



# AI for Imaging & Metrology

Virtual Special Issue

人工智能赋能计算成像与测量



SCILab



# “人工智能赋能计算成像与测量”虚拟专刊序言

光学成像的起源可以追溯到约公元前四百年前墨子的“小孔成像”，但直到最近半个多世纪才得以迅速发展。这得益于20世纪几项重要的发明：信号传输与通信、激光与光全息、光电信号数字化及数字信号处理技术，这些技术的融合催生了“光电成像”技术。2009年，我研究生考入南京理工大学，有幸加入导师陈钱教授领衔的“光电成像与信息处理”科研团队的一员。记得在那一年，诺贝尔物理学奖授予了对光的研究成果为现代数字时代奠定基础的科学家，其中包括贝尔实验室的威拉德·博伊尔（Willard Sterling Boyle）和乔治·史密斯（George Smith），以表彰他们发明了电荷耦合（Charge-coupled Device, CCD）图像传感芯片。这个小小的芯片，带领人类对光学影像的记录从金属感光板与胶片的“模拟时代”迈入了“数字时代”。

进入21世纪，“可调控”光电器件、高性能处理器/并行处理单元、新型数学/信号处理工具三方面的迅速发展与无缝结合推动了光电成像技术由传统的强度、彩色探测发展进入计算光学成像时代。这种转变体现在前端物理域的光学调控与后端数字域的信息处理的紧密结合上，为打破传统成像技术的限制提供了创新的解决方案，并预示着未来先进光学成像技术的发展趋势。在博士期间，我有幸见证了这一切的发生与发展，从红外探测器基于场景非均匀性校正，到高速结构光投影三维成像，再到非干涉相位恢复与定量相位显微成像，我不知不觉地成为了一名“计算成像”领域的“探索者”。

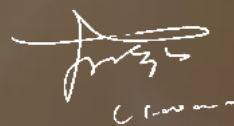
2014年底，我于南京理工大学博士毕业后留校工作，并创建了一个仅有不足十人的科研小组，命名为“智能计算成像实验室”（Smart Computational Imaging Laboratory, SCILab）。这或许是国内最早以“计算成像”来命名的实验室之一，而在“计算成像”之前，我加上了“智能”一词来加以修饰。因为我相信未来的成像技术，不单应该像通信与信息一样富含调制与解调的内涵，还应该像我们的眼睛与大脑一样强大而智能。近年来，我又一次有幸见证并参与了这一切的发生与发展——人工智能与深度学习技术的飞跃式发展为计算光学成像技术开启了一扇全新的大门：深度学习赋能了各类计算成像与测量技术，它不但解决了许多传统技术难以求解的非线性逆问题，还使成像系统的信息获取能力、功能、性能指标获得了质的提升。与国内外的很多同仁一起，我们不知不觉地成为了一群“智能计算成像”领域的“开拓者”。

时隔十年后的2024年，当地时间10月8日，瑞典皇家科学院宣布，将2024年诺贝尔物理学奖授予美国普林斯顿大学的约翰·霍普菲尔德（John J. Hopfield）和加拿大多伦多大学的杰弗里·辛顿（Geoffrey Hinton），以表彰他们为推动利用人工神经网络进行机器学习作出的基础性发现和发明。次日，诺贝尔化学奖则部分授予了谷歌旗下DeepMind公司AI科学家德米斯·哈萨比斯（Demis Hassabis）和约翰·江珀（John M. Jumper），以表彰他们研发出的“Alpha Fold2”模型在蛋白质结构预测方面的成就。当AI首次成为诺贝尔奖的主要元素，这不仅是对科学家的认可，也是对“AI赋能科学研究（AI for Science）”这一趋势的肯定。此前人们一直认为，诺贝尔奖主要授予在传统自然科学领域做出杰出贡献的个人或组织。此次诺奖标志着以人工智能驱动的科研方式已不再是“偏门”而是逐渐成为“主流”，并获得了传统自然科学领域的广泛认可。



回顾这十年的发展历程，我们从最初的不足十人的小团队起步，逐渐成长为一个结构多元、开放包容、充满活力的国际化研究团队。为了纪念这一历史性的时刻，促进学术交流，以及推动相关领域向纵深发展，我们南京理工大学智能计算成像实验室特别策划了本期“人工智能赋能计算成像与测量”虚拟专题，以集中展示实验室在该领域的最新研究进展。该专题共包含论文27篇，其中综述论文4篇，研究论文21篇，评述论文2篇，还包括Nature Portfolio Communities旗下“Behind the Paper”栏目邀请报道1篇。这些论文涵盖了当前计算光学成像领域的三大热点研究方向：①快速三维光学传感，涉及条纹分析与相位处理、结构光三维成像等；②生物医学显微成像，包括定量相位显微成像、光声成像、荧光超分辨显微成像等；③计算光电成像探测，覆盖光谱成像、合成孔径、红外探测等。这些论文反映了深度学习赋能计算光学成像的最新进展与发展趋势：深度学习摒弃了对传统“正向物理模型”和“逆向重构算法”的严格依赖，以“样本数据驱动”的方式给计算光学成像技术带来了颠覆性的变革，打破了传统技术的功能/性能疆界，从极少的原始图像数据中挖掘出更多场景的本质信息，显著提升了信息获取能力，为计算光学成像技术打开了一扇新的大门。未来，我们实验室将在“AI for Science”这一科技浪潮中逐浪前行，为人工智能与计算成像的无尽探索贡献自己的一份微薄之力。

最后，感谢实验室各位老师与论文作者对本专栏的付出与贡献；感谢各位专家同仁们长期以来对我们工作的帮助与指导；感谢各个期刊杂志社对我们团队成果的信任与纳荐；感谢每一位读者对我们实验室关注与支持。希望本“人工智能赋能计算成像与测量”专题能够为广大读者和相关从业人员提供有益的参考。



左超

—— 代表SCILab全体成员

2024年10月25日



## Preface to the Virtual Special Issue on “AI for Imaging & Metrology”

---

The origins of optical imaging can be traced back to Mozi’s “pinhole imaging” around 400 BC, but it has experienced rapid development only in the last half-century. This advancement is attributed to several key inventions of the 20th century, including signal transmission and communication, lasers and optical holography, photoelectric signal digitization, and digital signal processing. The integration of these technologies gave rise to “photoelectric imaging”. In 2009, I entered Nanjing University of Science and Technology as a graduate student and was fortunate to join the research team led by Professor Chen Qian, focusing on “Photoelectric Imaging and Information Processing”. I recall that year when the Nobel Prize in Physics was awarded to scientists from Bell Labs, Willard Boyle and George Smith, for their invention of the charge-coupled device (CCD) image sensor chip, which laid the foundation for the modern digital era. This small chip facilitated the transition from the “analog era” of photographic plates and film to the “digital era” of optical image recording.

As we entered the 21st century, the rapid development and seamless integration of “tunable” photoelectric devices, high-performance processors, and new mathematical and signal processing tools have propelled photoelectric imaging technology from traditional intensity and color detection into the era of computational optical imaging. This transformation is marked by the close coupling of optical modulation in the physical domain and information processing in the digital domain, providing innovative solutions to overcome the limitations of traditional imaging techniques and indicating future trends in advanced optical imaging technology. During my doctoral studies, I was fortunate to witness these developments, from scene-based non-uniformity correction in infrared detectors to high-speed structured light projection 3D imaging, and from non-interferometric phase retrieval to quantitative phase microscopy, unwittingly becoming an “explorer” in the field of “computational imaging”.

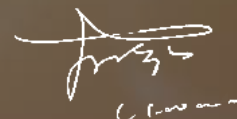
At the end of 2014, after graduating with a PhD from Nanjing University of Science and Technology, I stayed on to work and established a research group of fewer than ten members, named the “Smart Computational Imaging Laboratory” (SCILab). This may be one of the first laboratories in China specifically named for “computational imaging”. I prefixed it with “smart”, believing that future imaging technology should not only be rich in modulation and demodulation, like communication systems, but should also as powerful and intelligent as our eyes and brains. In recent years, I have once again witnessed and participated in the rapid advancements in artificial intelligence and deep learning, which have opened a new door for computational optical imaging technology. Deep learning has empowered various computational imaging and measurement technologies, solving many nonlinear inverse problems that traditional methods struggled to address, and qualitatively enhancing the capabilities, functions, and performance of imaging systems. Alongside many colleagues at home and abroad, we have unknowingly become “pioneers” in the field of “intelligent computational imaging”.

A decade later, on October 8, 2024, the Royal Swedish Academy of Sciences announced that the 2024 Nobel Prize in Physics was awarded to John J. Hopfield from Princeton University and Geoffrey Hinton from the University of Toronto for their foundational discoveries and inventions that advanced the use of artificial neural networks in machine learning. The following day, the Nobel Prize in Chemistry was awarded in part to AI scientists Demis Hassabis and John Jumper from DeepMind for their achievements with the “AlphaFold2” model in protein structure prediction. With AI becoming a key element of the Nobel Prizes for the first time, this not only recognizes the contributions of individual scientists but also affirms the trend of “AI for Science”. Previously, it was believed that the Nobel Prize primarily honored outstanding contributions in traditional natural sciences. This year’s Nobel recognition signifies that AI-driven research is no longer considered “niche” but is gradually becoming “mainstream” and gaining broad recognition in traditional natural science fields.



Reflecting on the past decade, we have grown from a small team of fewer than ten members into a vibrant, diverse, and internationally collaborative research group. To commemorate this historic moment, promote academic exchange, and advance related fields, the Smart Computational Imaging Laboratory at Nanjing University of Science and Technology has curated this virtual special issue on “AI for Imaging & Metrology” to showcase the latest research advancements in this area. This special issue features 27 papers, comprising 4 review articles, 21 research papers, and 2 commentary papers, along with an invited report in the “Behind the Paper” section of Nature Portfolio Communities. These papers explore three major research directions in current computational optical imaging: (1) rapid 3D optical sensing, including fringe analysis and phase processing, and structured light 3D imaging; (2) biomedical microscopic imaging, covering quantitative phase microscopy, photoacoustic imaging, and fluorescence super-resolution microscopy; (3) computational optoelectronic imaging detection, encompassing spectral imaging, synthetic aperture, and infrared detection. These papers reflect the latest progress and trends in deep learning-empowered computational optical imaging: deep learning has discarded strict reliance on traditional “forward physical models” and “inverse reconstruction algorithms”, bringing transformative changes through a “data-driven” approach, breaking the functional and performance boundaries of traditional technology, and extracting more essential scene information from minimal original image data, significantly enhancing information acquisition capabilities, thus opening new doors for computational optical imaging. In the future, our laboratory will continue to embrace the “AI for Science” trend, contributing our modest efforts to the endless exploration of artificial intelligence and computational imaging.

Finally, I would like to express my gratitude to all the teachers and authors in the laboratory for their contributions to this special issue; to the experts and colleagues for their long-term support and guidance; to the various journals for their trust in and recommendations of our team’s work; and to every reader for their attention to and support of our laboratory. I hope this special issue on “AI for Imaging & Metrology” will provide valuable references for a wide audience and professionals in the field.



Chao Zuo

— On behalf of all SCILab members

October 25, 2024



<b>01</b>	<b>研究综述与最新进展</b>	Research Overview and Recent Advances
	<b>Deep learning in optical metrology: a review</b>	1
	C. Zuo, J. Qian, S. Feng, W. Yin, Y. Li, P. Fan, K. Qian, and Q. Chen	
	<i>Light Sci. Appl.</i> 11(1), 1-54 (2022)	
	封面论文 ESI高被引论文 ESI热点论文	
	<b>深度学习下的计算成像: 现状, 挑战与未来</b>	55
	左超, 冯世杰, 张翔宇, 韩静, 陈钱	
	光学学报 40(1), 0111003 (2020)	
	入选学术精要高PSCI论文	
<b>02</b>	<b>条纹分析与相位处理</b>	Fringe Pattern Analysis and Phase Processing
	<b>Fringe pattern analysis using deep learning</b>	81
	S. Feng, Q. Chen, G. Gu, T. Tao, L. Zhang, Y. Hu, W. Yin, and C. Zuo	
	<i>Adv. Photonics</i> 1(2), 025001 (2019)	
	封面论文 ESI高被引论文 编辑精选	
	<b>Deep-learning-based fringe-pattern analysis with uncertainty estimation</b>	88
	S. Feng, C. Zuo, Y. Hu, Y. Li, and Q. Chen	
	<i>Optica</i> 8(12), 1507-1510 (2021)	
	封面论文	
	<b>Physics-informed deep learning for fringe pattern analysis</b>	93
	W. Yin, Y. Che, X. Li, M. Li, Y. Hu, S. Feng, EY. Lam, Q. Chen, and C. Zuo	
	<i>Opto-Electronic Adv.</i> 7(1), 230034-1-230034-12 (2024)	
	封面论文	
	<b>Generalized framework for non-sinusoidal fringe analysis using deep learning</b>	105
	S. Feng, C. Zuo, L. Zhang, W. Yin, and Q. Chen	
	<i>Photonics Res.</i> 9(6), 1084-1098 (2021)	
	ESI高被引论文 ESI热点论文	
	<b>Fringe-pattern analysis with ensemble deep learning</b>	120
	S. Feng, Y. Xiao, W. Yin, Y. Hu, Y. Li, C. Zuo, and Q. Chen	
	<i>Adv. Photonics Nexus</i> 2(3), 036010-036010 (2023)	
	<b>Temporal phase unwrapping using deep learning</b>	127
	W. Yin, Q. Chen, S. Feng, T. Tao, L. Huang, M. Trusiak, A. Asundi, and C. Zuo	
	<i>Sci. Rep.</i> 9, 20175 (2019)	



- 深度学习技术在条纹投影三维成像中的应用 139  
冯世杰, 左超, 尹维, 陈钱  
*红外与激光工程* 49(3), 0303018-0303018-17(2020)

入选中国精品科技期刊顶尖学术论文

- Deep-learning-enabled dual-frequency composite fringe projection  
profilometry for single-shot absolute 3D shape measurement 156  
Y. Li, J. Qian, S. Feng, Q. Chen, and C. Zuo  
*Opto-Electronic Adv.* 5, 210021 (2022)

封面论文

ESI高被引论文

- Deep-learning-enabled temporally super-resolved multiplexed fringe  
projection profilometry: high-speed kHz 3D imaging with low-speed camera 172  
W. Chen, S. Feng, W. Yin, Y. Li, J. Qian, Q. Chen, and C. Zuo  
*Photonix* 5(1), 25 (2024)

编辑精选

- Micro deep learning profilometry for high-speed 3D surface imaging 184  
S. Feng, C. Zuo, W. Yin, G. Gu, and Q. Chen  
*Opt. Lasers Eng.* 121, 416-427 (2019)

- Deep-learning-enabled geometric constraints and phase unwrapping  
for single-shot absolute 3D shape measurement 196  
J. Qian, S. Feng, T. Tao, Y. Hu, Y. Li, Q. Chen, and C. Zuo  
*APL Photonics* 5(4), 046105 (2020)

ESI高被引论文

ESI热点论文

- Single-shot absolute 3D shape measurement with deep-learning-based  
color fringe projection profilometry 207  
J. Qian, S. Feng, Y. Li, T. Tao, J. Han, Q. Chen, and C. Zuo  
*Opt. Lett.* 45(7), 1842-1845 (2020)

ESI高被引论文

ESI热点论文

- Single-shot 3D shape measurement using an end-to-end stereo matching  
network for speckle projection profilometry 211  
W. Yin, Y. Hu, S. Feng, L. Huang, K. Qian, Q. Chen, and C. Zuo  
*Opt. Express* 29(9), 13388-13407 (2021)

**Composite fringe projection deep learning profilometry for single-shot absolute 3D shape measurement** 231

Y. Li, J. Qian, S. Feng, Q. Chen, and C. Zuo

*Opt. Express* 30(3), 3424-3442 (2022)

封面论文

**Neural-field-assisted transport-of-intensity phase microscopy: partially coherent quantitative phase imaging under unknown defocus distance** 250

Y. Jin, L. Lu, S. Zhou, J. Zhou, Y. Fan, and C. Zuo

*Photonics Res.* 12(7), 1494-1501 (2024)

**4D spectral-spatial computational photoacoustic dermoscopy** 258

Y. Gao, T. Feng, H. Qiu, Y. Gu, Q. Chen, C. Zuo, and H. Ma

*Photoacoustics* 34, 100572 (2023)

**Deep learning-enabled pixel-super-resolved quantitative phase microscopy from single-shot aliased intensity measurement** 272

J. Zhou, Y. Jin, L. Lu, S. Zhou, H. Ullah, J. Sun, Q. Chen, R. Ye, J. Li, and C. Zuo

*Laser Photonics Rev.* 18(1), 230048 (2024)

封面论文

**Deep learning assisted variational Hilbert quantitative phase imaging** 288

Z. Li, J. Sun, Y. Fan, Y. Jin, Q. Shen, M. Trusiak, M. Cywińska, P. Gao, Q. Chen, and C. Zuo

*Opto-Electronic Sci.* 2(4), 220023-1-220023-11 (2023)

封面论文

**Deep Learning-powered biomedical photoacoustic imaging** 299

X. Wei, T. Feng, Q. Huang, Q. Chen, C. Zuo, and H. Ma

*Neurocomputing* 573(7), 127207 (2023)

**深度学习在超分辨显微成像中的研究进展** 322

鲁心怡, 黄昱, 张梓童, 吴天筱, 吴洪军, 刘永焘, 方中, 左超, 陈钱

*激光与光电子学进展* 61(16), 1611002 (2024)

封面论文



**Learning-based single-shot long-range synthetic aperture Fourier ptychographic imaging with a camera array** 340

B. Wang, S. Li, Q. Chen, and C. Zuo

*Opt. Lett.* 48(2), 263-266 (2023)

👍 编辑精选

**Model-based deep learning for fiber bundle infrared image restoration** 344

B. Wang, L. Li, W. Yang, J. Chen, Y. Li, Q. Chen, and C. Zuo

*Def. Technol.* 27, 38-45 (2022)

**Multimodal super-resolution reconstruction of infrared and visible images via deep learning** 352

B. Wang, Y. Zou, L. Zhang, Y. Li, Q. Chen, and C. Zuo

*Opt. Lasers Eng.* 156, 107078 (2022)

**Behind the paper: Deep learning in optical metrology: a review** 363

C. Zuo

*Nat. Portf. Eng. Community* (2022)

**Optical metrology embraces deep learning: keeping an open mind** 380

B. Pan

*Light Sci. Appl.* 11, 139 (2022)

**Exploiting optical degrees of freedom for information multiplexing in diffractive neural networks** 383

C. Zuo, and Q. Chen

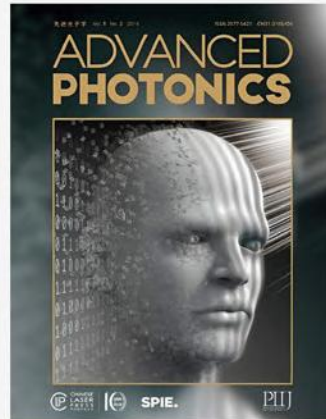
*Light Sci. Appl.* 208, 1-4 (2022)



Deep learning in optical metrology:  
a review

*Light Sci. Appl.* 11(1), 1-54 (2022)

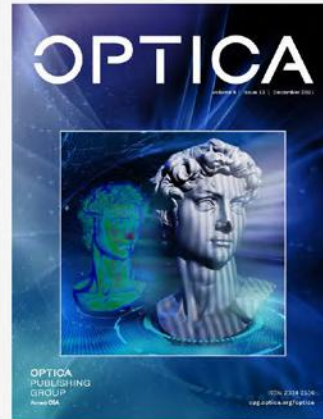
P-1



Fringe pattern analysis using  
deep learning

*Adv. Photonics* 1(2), 025001 (2019)

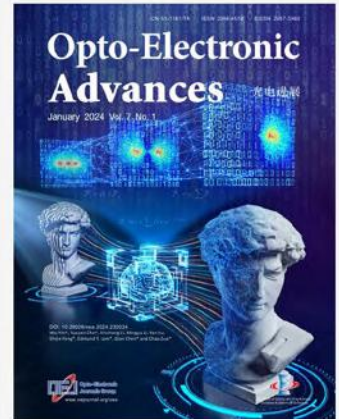
P-81



Deep-learning-based fringe-pattern  
analysis with uncertainty estimation

*Optica* 8(12), 1507-1510 (2021)

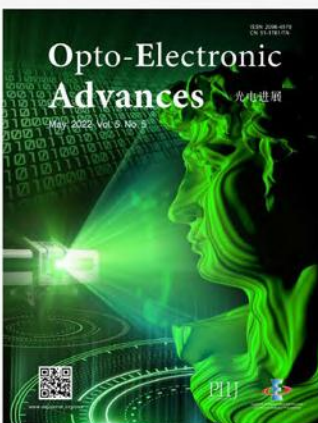
P-88



Physics-informed deep learning for  
fringe pattern analysis

*Opto-Electronic Adv.* 7(1),  
230034-1-230034-12 (2024)

P-93



Deep-learning-enabled dual-frequency  
composite fringe projection  
profilometry for single-shot absolute  
3D shape measurement

*Opto-Electronic Adv.* 5, 210021 (2022)

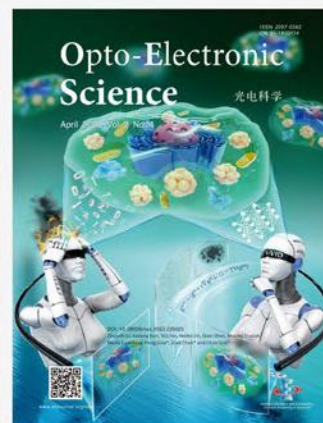
P-156



Deep learning-enabled pixel-super-  
resolved quantitative phase microscopy  
from single-shot aliased intensity  
measurement

*Laser Photonics Rev.* 18(1),  
230048 (2024)

P-272



Deep learning assisted variational  
Hilbert quantitative phase imaging

*Opto-Electronic Sci.* 2(4),  
220023-1-220023-11 (2023)

P-288



深度学习在超分辨显微成像中  
的研究进展

激光与光电子学进展  
61(16), 1611002 (2024)

P-322



REVIEW ARTICLE

Open Access

# Deep learning in optical metrology: a review

Chao Zuo<sup>1,2</sup>✉, Jiaming Qian<sup>1,2</sup>✉, Shijie Feng<sup>1,2</sup>, Wei Yin<sup>1,2</sup>✉, Yixuan Li<sup>1,2</sup>, Pengfei Fan<sup>1,2,3</sup>, Jing Han<sup>2</sup>, Kemao Qian<sup>4</sup>✉ and Qian Chen<sup>2</sup>✉

## Abstract

With the advances in scientific foundations and technological implementations, optical metrology has become versatile problem-solving backbones in manufacturing, fundamental research, and engineering applications, such as quality control, nondestructive testing, experimental mechanics, and biomedicine. In recent years, deep learning, a subfield of machine learning, is emerging as a powerful tool to address problems by learning from data, largely driven by the availability of massive datasets, enhanced computational power, fast data storage, and novel training algorithms for the deep neural network. It is currently promoting increased interests and gaining extensive attention for its utilization in the field of optical metrology. Unlike the traditional “physics-based” approach, deep-learning-enabled optical metrology is a kind of “data-driven” approach, which has already provided numerous alternative solutions to many challenging problems in this field with better performances. In this review, we present an overview of the current status and the latest progress of deep-learning technologies in the field of optical metrology. We first briefly introduce both traditional image-processing algorithms in optical metrology and the basic concepts of deep learning, followed by a comprehensive review of its applications in various optical metrology tasks, such as fringe denoising, phase retrieval, phase unwrapping, subset correlation, and error compensation. The open challenges faced by the current deep-learning approach in optical metrology are then discussed. Finally, the directions for future research are outlined.

## Introduction


Optical metrology is the science and technology of making measurements with the use of light as standards or information carriers<sup>1–3</sup>. Light is characterized by its fundamental properties, namely, amplitude, phase, wavelength, direction, frequency, speed, polarization, and coherence. In optical metrology, these fundamental properties of light are ingeniously utilized as information carriers of a measurand, enabling a wide range of optical metrology tools that allow the measurement of a wide range of subjects<sup>4–6</sup>. For example, optical

interferometry takes advantage of the wavelength of light as a precise dividing marker of length. The speed of light defines the international standard of length, the meter, as the length traveled in vacuum during a time interval of 1/299,792,458 of a second<sup>7</sup>. As a result, optical metrology is being increasingly adopted in many applications where reliable data about the distance, displacement, dimensions, shape, roughness, surface properties, strain, and stress state of the object under test are required<sup>8–10</sup>. Optical metrology is a broad and interdisciplinary field relating to diverse disciplines such as photomechanics, optical imaging, and computer vision. There is no strict boundary between those fields, and in fact, the term “optical metrology” is often interchangeably used with “optical measurement”, in which achieving higher precision, sensitivity, repeatability, and speed is always a priority<sup>11,12</sup>.

There are a few inventions that revolutionized optical metrology. The first is the invention of laser<sup>13,14</sup>. The

Correspondence: Chao Zuo (zuochoao@njjust.edu.cn) or Kemao Qian (mkmqian@ntu.edu.sg) or Qian Chen (chenqian@njjust.edu.cn)  
<sup>1</sup>Smart Computational Imaging (SCI) Laboratory, Nanjing University of Science and Technology, 210094 Nanjing, Jiangsu Province, China  
<sup>2</sup>Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing University of Science and Technology, 210094 Nanjing, Jiangsu Province, China  
Full list of author information is available at the end of the article  
These authors contributed equally: Chao Zuo, Jiaming Qian  
Smart Computational Imaging (SCI) Laboratory: <https://scilaboratory.com/>

© The Author(s) 2022, corrected publication 2022

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

advent of laser interferometry could be traced back to experiments conducted independently in 1962 by Denisyuk<sup>15</sup> and Leith and Upatnix<sup>16</sup> with the objective of marrying coherent light produced by lasers with Gabor's holography method<sup>17</sup>. The use of lasers as a light source in optical metrology marked the first time that such highly controlled light became available as a physical medium to measure the physical properties of samples, opening up new possibilities for optical metrology. The second revolution was initiated with the invention of charged coupled device (CCD) cameras in 1969, which replaced the earlier photographic emulsions by virtue of recording optical intensity signals from the measurand digitally<sup>8</sup>. The use of the CCD camera as a recording device in optical metrology represented another important milestone: the compatibility of light with electricity, i.e., "light" can be converted into "electrical quantity (current, voltage, etc.)". This means that the computational storage, access, analysis, and transmission of captured data are easily attainable, leading to the "digital transition" of optical metrology. Computer-based signal processing tools were introduced to automate the quantitative determination of optical metrology data, eliminating the inconvenience associated with the manual, labor-intensive, time-consuming evaluation of fringe patterns<sup>18–20</sup>. Methods such as digital interferometry<sup>21</sup>, digital holography<sup>22</sup>, and digital image correlation (DIC)<sup>23</sup> have become state of the art by now.

With the digital transition, image processing plays an essential role in optical metrology for the purpose of converting the observed measurements (generally displayed in the form of deformed fringe/speckle patterns) into the desired attributes (such as geometric coordinates, displacements, strain, refractive index, and others) of an object under study. Such information-recovery process is similar to those of computer vision and computational imaging, presenting as an inverse problem that is often ill-posed with respect to the existence, uniqueness, and stability of the solution<sup>24–27</sup>. Tremendous progress has been achieved in terms of accurate mathematical modeling (statistical models of noise and the observational data)<sup>28</sup>, regularization techniques<sup>29</sup>, numerical methods, and their efficient implementations<sup>30</sup>. For the field of optical metrology, however, the situation becomes quite different due to the fact that the optical measurements are frequently carried out in a highly controlled environment. Instead of explicitly interpreting optical metrology tasks from the perspective of solving inverse problems (based on a formal optimization framework), mainstream scientists in optical metrology prefer to bypass the ill-posedness and simplify the problem by means of active strategies, such as sample manipulation, system adjustment, and multiple acquisitions<sup>31</sup>. A typical example is the phase-shifting technique<sup>32</sup>, which sacrifices the time and effort

of capturing multiple fringe patterns to exchange for a deterministic and straightforward solution. Under such circumstances, the phase retrieval problem is well-posed or even over-determined (when the phase-shifting step is larger than  $\lambda$ ), and employing more evolved algorithms, such as compressed sensing<sup>33</sup> and nonconvex (low-rank) regularization<sup>34</sup> seem redundant and unnecessary, especially as they fail to demonstrate clear advantages over classical ones in terms of accuracy, adaptability, speed, and, more importantly, ease-of-use. This gives us the key question and motivation of this review paper: whether machine learning will be the driving force in optical metrology not only provides superior solutions to the growing new challenges but also tolerates imperfect measurement conditions with the least efforts, such as additive noise, phase-shifting error, intensity nonlinearity, motion, and vibration.

In the past few years, we have indeed witnessed the rapid progress on high-level artificial intelligence (AI), where deep representations based on convolutional and recurrent neural network models are learned directly from the captured data to solve many tasks in computer vision, computational imaging, and computer-aided diagnosis with unprecedented performance<sup>35–37</sup>. The early framework for deep learning was established on artificial neural networks (ANNs) in the 1980s<sup>38</sup>, yet only recently the real impact of deep learning became significant due to the advent of fast graphics processing units (GPUs) and the availability of large datasets<sup>39</sup>. In particular, deep learning has revolutionized the computer vision community, introducing non-traditional and effective solutions to numerous challenging problems such as object detection and recognition<sup>40</sup>, object segmentation<sup>41</sup>, pedestrian detection<sup>42</sup>, image super-resolution<sup>43</sup>, as well as medical image-related applications<sup>44</sup>. Similarly, in computational imaging, deep learning has led to rapid growth in algorithms and methods for solving a variety of ill-posed inverse computational imaging problems<sup>45</sup>, such as super-resolution microscopy<sup>46</sup>, lensless phase imaging<sup>47</sup>, computational ghost imaging<sup>48</sup>, and image through scattering media<sup>49</sup>. In this context, researchers in optical metrology have also made significant explorations in this regard with very promising results within just a few short years, as evidenced by the ever-increasing and the respectable number of publications<sup>50–55</sup>. Meanwhile, those research works are scattered rather than systematic, which gives us the second motivation to provide a comprehensive review to understand their principles, implementations, advantages, applications, and challenges. It should be noted that optical metrology covers a wide range of methods and applications today. It would be beyond the scope of this review to discuss all relevant technologies and trends. We, therefore, restrict our focus to



phase/correlation measurement techniques, such as interferometry, holography, fringe projection, and DIC. Although phase retrieval and wave-field sensing technologies, such as defocus variation (Gerchberg–Saxton–Fienup-type methods<sup>56,57</sup>), transport of intensity equation (TIE)<sup>58,59</sup>, aperture modulation<sup>60</sup>, ptychography<sup>61,62</sup>, and wavefront sensing (e.g., Shack–Hartmann<sup>63</sup>, Pyramid<sup>64</sup>, and computational shear interferometry<sup>65</sup>), has been recently introduced to optical metrology<sup>66–68</sup>, they may be more appropriately placed in the field of “computational imaging”. The reader is referred to the earlier review by Barbastathis et al.<sup>45</sup> for more detailed information on this topic. It is also worth mentioning that (passive) stereovision, which extracts depth information from stereo images, is an important branch of photogrammetry that has been extensively studied by the computer vision community. Although stereovision techniques do not strictly fall into the category of optical metrology, due to the fact that many ideas and algorithms in DIC and fringe projection were “borrowed” from stereovision, they are also included in this review.

The remainder of this review is organized as follows. We start by summarizing the relevant foundations and image formation models of different optical metrology approaches, which are generally required as a priori knowledge in conventional optical metrology methods. Next, we present a general hierarchy of the image-processing algorithms that are most commonly used in conventional optical metrology in the “Image processing in optical metrology” section. After a brief introduction to the history and basic concept of deep learning, we recapitulate the advantages of using deep learning in optical metrology tasks by interpreting the concept as an optimization problem. We then present a recollection of the deep learning methods that have been proposed in optical metrology, suggesting the pervasive penetration of deep learning in almost all aspects of the image-processing hierarchy. The “Challenges” section discusses both technical and implementation challenges faced by the current deep-learning approach in optical metrology. In the “Future directions” section, we give our outlook for the prospects for deep learning in optical metrology. Finally, conclusions and closing remarks are given in the “Conclusions” section.

### Image formation in optical metrology

Optical metrology methods often form images (e.g., fringe/speckle patterns) for processing. Thus image formation is essential to reconstruct various quantities. In most interferometric metrological methods, the image is formed by the coherent superposition of the object and reference beams. As a result, the raw intensity across the object is modulated by a harmonic function, resulting in

the bright and dark contrasts, known as fringe patterns. A typical fringe pattern can be written as<sup>18,19</sup>

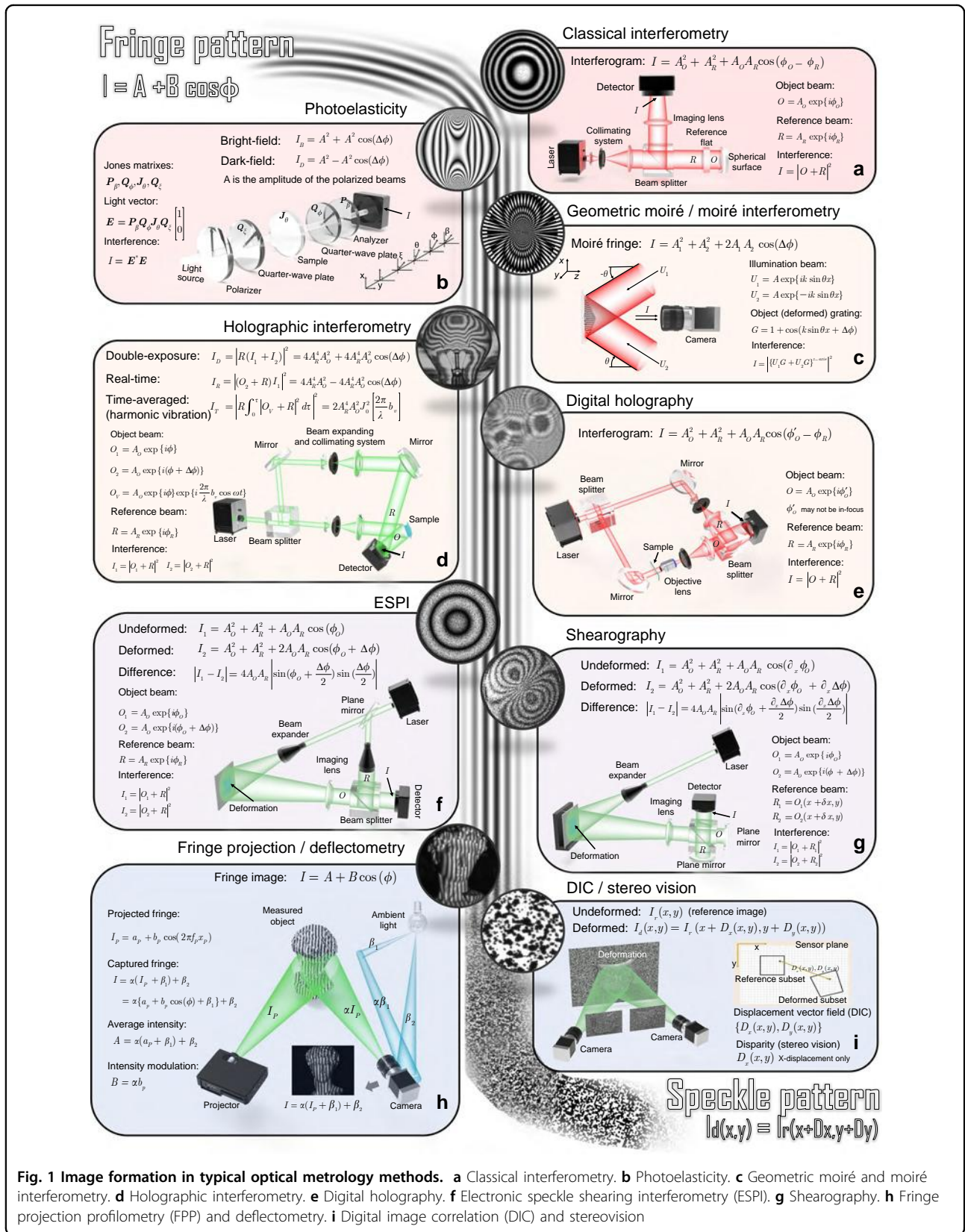
$$I(x, y) = A(x, y) + B(x, y) \cos[\phi(x, y)] \quad (1)$$

where  $(x, y)$  refers to the spatial coordinates along the horizontal and vertical directions,  $A(x, y)$  is the background intensity,  $B(x, y)$  is the fringe amplitude,  $\phi(x, y)$  is the phase distribution. In most cases, phase is the primary quantity of the fringe pattern to be retrieved as it is related to the final object quantities of interest, such as surface shape, mechanical displacement, 3D coordinates, and their derivations. The related techniques include classical interferometry, photoelasticity, holographical interferometry, digital holography, etc. On a different note, the fringe patterns can also be created noninterferometrically by overlapping of two periodic gratings as in geometric moiré, or incoherent projection of structured patterns onto the object surface as in fringe projection profilometry (FPP)/deflectometry. As summarized in Fig. 1, though the final fringe patterns obtained in all forms of fringe-based techniques discussed herein are similar in form, the physics behind the image formation process and the meanings of the fringe parameters are different. In DIC, the measured intensity images are speckle patterns of the specimen surface before and after deformation,

$$I_d(x, y) = I_r(x + D_x(x, y), y + D_y(x, y)) \quad (2)$$

where  $(D_x(x, y), D_y(x, y))$  refers to the displacement vector-field mapping from the undeformed/reference pattern  $I_r(x, y)$  to the deformed one  $I_d(x, y)$ . It directly provides full-field displacements and strain distributions of the sample surface. The DIC technique can also be combined with binocular stereovision or stereophotogrammetry to recover depth and out-of-plane deformation of the surface from the displacement field (so-called disparity) by exploiting the unique textures present in two or more images of the object taken from different viewpoints. The image formation processes for typical optical metrology methods are briefly described as follows.

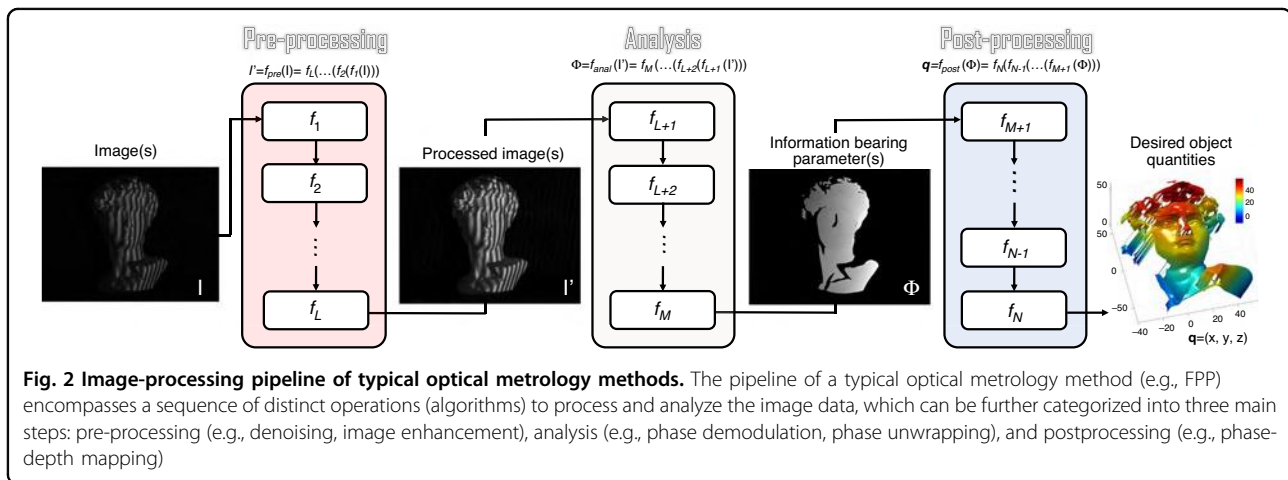
- (1) **Classical interferometry:** In classical interferometry, the fringe pattern is formed by superposition of two smooth coherent wavefronts, one of which is typically a flat or spherical reference wavefront and the other a distorted wavefront formed and directed by optical components<sup>69,70</sup> (Fig. 1a). The phase of the fringe pattern reflects the difference between the ideal reference wavefront and object wavefront. Typical examples of classical interferometry include the use of configurations such as the Michelson, Fizeau,



**Fig. 1** Image formation in typical optical metrology methods. **a** Classical interferometry. **b** Photoelasticity. **c** Geometric moiré and moiré interferometry. **d** Holographic interferometry. **e** Digital holography. **f** Electronic speckle shearing interferometry (ESPI). **g** Shearography. **h** Fringe projection profilometry (FPP) and deflectometry. **i** Digital image correlation (DIC) and stereovision



- Twyman Green, and Mach-Zehnder interferometers to characterize the surface, aberration, or roughness of optical components with high accuracy, of the order of a fraction of the wavelength.
- (2) **Photoelasticity:** Photoelasticity is a nondestructive, full-field, optical metrology technique for measuring the stress developed in transparent objects under loading<sup>71,72</sup>. Photoelasticity is based on an optomechanical property, so-called “double refraction” or “birefringence” observed in many transparent polymers. Combined with two circular polarizers (linear polarizer coupled with quarter waveplate) and illuminated with a conventional light source, a loaded photoelastic sample (or photoelastic coating applied to an ordinary sample) can produce fringe patterns whose phases are associated with the difference between the principal stresses in a plane perpendicular to the light propagation direction<sup>73</sup> (Fig. 1b).
  - (3) **Geometric moiré/Moiré interferometry:** In optical metrology, the moiré technique is defined as the utilization of the moiré phenomenon to measure shape, deformation, or displacements of surfaces<sup>74,75</sup>. A moiré pattern is formed by the superposition of two periodic or quasi-periodic gratings. One of these gratings is called reference grating, and the other one is object grating mounted or engraved on the surface to be studied, which is subjected to distortions induced by surface changes. For in-plane displacement and strain measurements, moiré technology has evolved from low-sensitivity geometric moiré<sup>75–77</sup> to high-sensitivity moiré interferometry<sup>75,78</sup>. In moiré interferometry, two collimated coherent beams interfere to produce a virtual reference grating with high frequencies, which interacts with the object grating to create the moiré pattern with fringes representing subwavelength in-plane displacements per contour (Fig. 1c).
  - (4) **Holographic interferometry:** Holography, invented by Gabor<sup>17</sup> in the 1940 s, is a technique that records an interference pattern and uses diffraction to reproduce a wavefront, resulting in a 3D image that still has the depth, parallax, and other properties of the original scene. The principle of holography can also be utilized as an optical metrology tool. In holographic interferometry, a wavefront is first stored in the hologram and later interferometrically compared with another, producing fringe patterns that yield quantitative information about the object surface deriving these two wavefronts<sup>79,80</sup>. This comparison can be made in three different ways that constitute the basic approaches of holographic interferometry:
    - real-time<sup>81</sup>, double-exposure<sup>82</sup>, and time-average holographic interferometry<sup>83,84</sup> (Fig. 1d), allowing for both qualitative visualization and quantitative measurement of real-time deformation and perturbation, changes of the state between two specific time points, and vibration mode and amplitude, respectively.
    - (5) **Digital holography:** Digital holography utilizes a digital camera (CMOS or CCD) to record the hologram produced by the interference between a reference wave and an object wave emanating from the sample<sup>85,86</sup> (Fig. 1e). Unlike classical interferometry, the sample may not be precisely in-focus and can even be recorded without using any imaging lenses. The numerical propagation using Fresnel transform or angular spectrum algorithm enables digital refocusing at any depths of the sample without physically moving it. In addition, digital holography also provides an alternative and much simpler way to realize double-exposure<sup>87</sup> and time-averaged holographic interferometry<sup>88,89</sup>, without additional benefits of quantitative evaluation of holographic interferograms and flexible phase-aberration compensation<sup>86,90</sup>.
    - (6) **Electronic speckle pattern interferometry (ESPI):** In ESPI, the tested object generally has an optically rough surface. When illuminated by a coherent laser beam, it will create a speckle pattern with random phase, amplitude, and intensity<sup>91,92</sup>. If the object is displaced or deformed, the object-to-image distance will change, and the phase of the speckle pattern will change accordingly. In ESPI, two speckle patterns are acquired one each for the undeformed and deformed states, by double exposure, and the absolute difference between these two deformed patterns results in the form of fringes superimposed on the speckle pattern where each fringe contour normally represents a displacement of half a wavelength (Fig. 1f).
    - (7) **Electronic speckle shearing interferometry (shearography):** Electronic speckle shearing interferometry, commonly known as shearography, is an optical measurement technique similar to ESPI. However, instead of using a separate known reference beam, shearography uses the test object itself as the reference; and the interference pattern is created by two sheared speckle fields originated from the light scattered by the surface of the object under test<sup>93,94</sup>. In shearography, the phase encoded in the fringe pattern depicts the derivatives of the surface displacements, i.e., to the strain developed on the object surface (Fig. 1g). Consequently, the anomalies or defects on the surface of the object



can be revealed more prominently, rendering shearography one of the most powerful tools for nondestructive testing applications.

(8) **Fringe projection profilometry/deflectometry:**

Fringe projection is a widely used noninterferometric optical metrology technique for measuring the topography of an object at a certain angle between the observation and the projection point<sup>95,96</sup>. The sinusoidal pattern in fringe projection techniques is generally incoherently formed by a digital video projector and directly projected onto the object surface. The corresponding distorted fringe pattern is recorded by a digital camera. The average intensity and intensity modulation of the captured fringe pattern are associated with the surface reflectivity and ambient illuminations, and the phase is associated with the surface height<sup>32</sup> (Fig. 1h). Deflectometry is another structured light technique similar to FPP, but instead of being produced by a projector, similar types of fringe patterns are displayed on a planar screen and distorted by the reflective (mirror-like) test surface<sup>97,98</sup>. The phase measured in deflectometry is directly sensitive to the surface slope (similar to shearography), so it is more effective for detecting shape defects<sup>99,100</sup>.

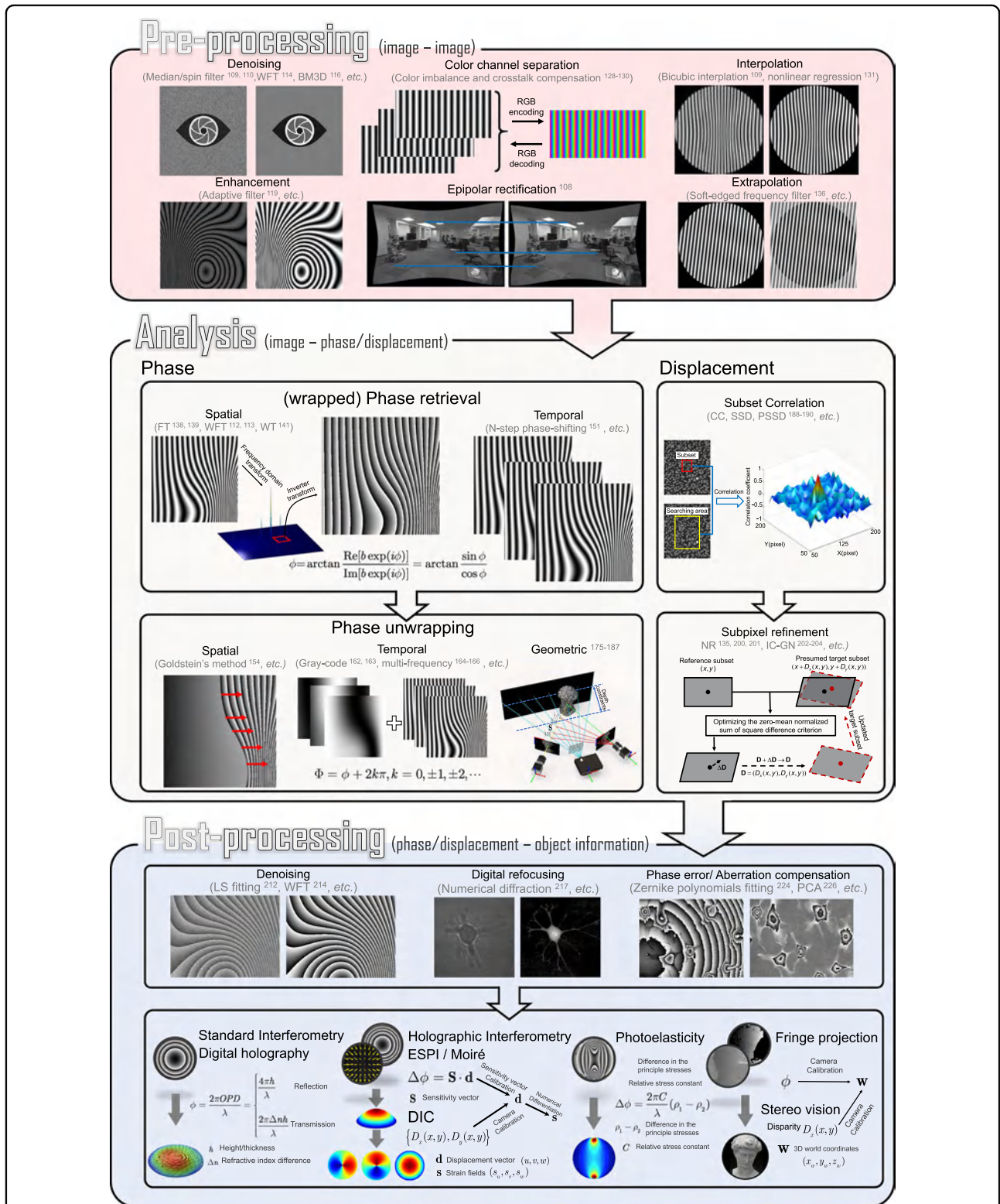
(9) **Digital image correction (DIC)/stereovision:**

DIC is another important noninterferometric optical metrology method that employs image correlation techniques for measuring full-field shape, displacement, and strains of an object surface<sup>23,101,102</sup>. Generally, the object surface should have a random intensity distribution (i.e., a random speckle pattern), which distorts together with the sample surface as a carrier of deformation information. Images of the object at different loadings are captured with one (2D-DIC)<sup>23</sup>, or two

synchronized cameras (3D-DIC)<sup>103</sup>, and then these images are analyzed with correlation-based matching (tracking or registration) to extract full-field displacement and strain distributions (Fig. 1i). Unlike 2D-DIC that is limited to in-plane deformation measurement of nominal planar objects, 3D-DIC, also known as stereo-DIC, allows for the measurement of 3D displacements (both in-plane and out-of-plane) for both planar and curved surfaces<sup>104,105</sup>. 3D-DIC is inspired by binocular stereovision or stereophotogrammetry in the computer vision community, which recovers the 3D coordinates by finding pixel correspondence (i.e., disparity) of unique features that exist in two or more images of the object taken from different points of view<sup>106,107</sup>. Nevertheless, unlike DIC, in which the displacement vector can be along both  $x$  and  $y$  directions, in stereophotogrammetry, after epipolar rectification, disparities between the images are along the  $x$  direction only<sup>108</sup>.

### Image processing in optical metrology

The elementary task of digital image processing in optical metrology can be defined as the conversion of the captured raw intensity image(s) into the desired object quantities taking into account the physical model of the intensity distribution describing the image formation process. In most cases, image processing in optical metrology is not a one-step procedure, and a logical hierarchy of image processing steps should be accomplished. As illustrated in Fig. 2, the image-processing hierarchy typically encompasses three main steps, pre-processing, analysis, and postprocessing, each of which includes a series of mapping functions that are cascaded to form a pipeline structure. For each operation, the corresponding  $f$  is an operator that transforms



**Fig. 3 Hierarchy and typical algorithms of image processing in optical metrology.** Image processing in optical metrology is not a one-step procedure. Depending on the purpose of the evaluation, a logical hierarchy of processing steps should be implemented before the desired information can be extracted from the image. In general, the image processing architecture in optical metrology consists of three main steps: pre-processing, analysis, and post-processing.



the image-like input into an output of corresponding (possibly resampled) spatial dimensions. Figure 3 shows the big picture of the image-processing hierarchy with various types of algorithms distributed in different layers. Next, we will zoom in one level deeper on each of the hierarchical steps.

### Pre-processing

The purpose of pre-processing is to assess the quality of the image data and improve the data quality by suppressing or minimizing unwanted disturbances (noise, aliasing, geometric distortions, etc.) before being fed to the following image analysis stage. It takes place at the lowest level (so-called iconic level) of image processing—the input and output of the corresponding mapping function (s) are both intensity images, i.e.,  $f_{anal} : I \rightarrow I'$ . Representative image pre-processing algorithms in optical metrology includes but not limited to:

- **Denoising:** In optical metrology, noise in captured raw intensity data has several sources that are related to the electronic noise of photodetectors and the coherent noise (so-called speckle). Typical numerical approaches to noise reduction include median filter<sup>109</sup>, spin filter<sup>110</sup>, anisotropic diffusion<sup>111</sup>, coherence diffusion<sup>112</sup>, Wavelet<sup>113</sup>, windowed Fourier transform (WFT)<sup>114,115</sup>, block matching 3D (BM3D)<sup>116</sup>, etc. For more detailed information and comparisons of these algorithms, the reader may refer to the reviews by Kulkarnia and Rastogi<sup>117</sup> and Bianco et al.<sup>118</sup>.
- **Enhancement:** Image enhancement is a crucial pre-processing step in intensity-based fringe analysis approaches, such as fringe tracking or skeletonizing. Referring to the intensity model, the fringe pattern may still be disturbed by locally varying background and intensity modulation after denoising. Several algorithms have been developed for fringe pattern enhancement, e.g., adaptive filter<sup>119</sup>, bidimensional empirical mode decomposition<sup>120,121</sup>, and dual-tree complex wavelet transform<sup>122</sup>.
- **Color channel separation:** Because a Bayer color sensor-camera captures three monochromatic (red, green, and blue) images at once, color multiplexing techniques are often employed in optical metrology to speed up the image acquisition process<sup>123–127</sup>. However, the separation of three color channels is not so straightforward due to the coupling and imbalance among the three color channels. Many cross-talk-matrix-based color channel calibration and leakage correction algorithms have been proposed to minimize such side effects<sup>128–130</sup>.
- **Image registration and rectification:** Image registration and rectification are aimed at aligning two or more images of the same object to a reference or correcting image distortion due to lens aberration.

In stereophotogrammetry, epipolar (stereo) rectification determines a reprojection of each image plane so that pairs of conjugate epipolar lines in both images become collinear and parallel to one of the image axes<sup>108</sup>.

- **Interpolation:** Image interpolation algorithms, such as the nearest neighbor, bilinear, bicubic<sup>109</sup>, and nonlinear regression<sup>131</sup> are necessary when the measured intensity image is sampled at an insufficient dense grid. In DIC, to reconstruct displacements with subpixel accuracy, the correlation criterion must be evaluated at non-integer-pixel locations<sup>132–134</sup>. Therefore, image interpolation is also a key algorithm for DIC to infer subpixel gray values and gray-value gradients in many subpixel displacement registration algorithms, e.g., the Newton–Raphson method<sup>133–135</sup>.
- **Extrapolation:** Image extrapolation, especially fringe extrapolation is often employed in Fourier transform (FT) fringe analysis methods to minimize the boundary artifacts induced by spectrum leakage. Schemes for the extrapolation of the fringe pattern beyond the borders have been reported, such as soft-edged frequency filter<sup>136</sup> and iterative FT<sup>137</sup>.

### Analysis

Image analysis is the core component of the image-processing architecture to extract the key information-bearing parameter(s) reflecting the desired physical quantity being measured from the input images. In phase measurement techniques, image analysis refers to the reconstruction of phase information from the fringe-like modulated intensity distribution(s), i.e.,  $f_{anal} : I \rightarrow \phi$ .

- **Phase demodulation:** The aim of phase demodulation, or more specifically, fringe analysis, is to obtain the wrapped phase map from the quasi-periodic fringe patterns. Various techniques for fringe analysis have been developed to meet different requirements in diverse applications, which can be broadly classified into two categories:

**Spatial phase demodulation:** Spatial phase-demodulation methods are capable of estimating the phase distribution through a single-fringe pattern. FT<sup>138,139</sup>, WFT<sup>114,115,140</sup>, and wavelet transform (WT)<sup>141</sup> are classical methods for the spatial carrier fringe analysis. For closed-fringe patterns without the carrier, alternative methods, such as Hilbert spiral transform<sup>142,143</sup>, regularized phase tracking (RPT)<sup>144,145</sup> and frequency-guided sequential demodulation<sup>146,147</sup>, can be applied provided that the sinusoidal component of the fringe pattern can be extracted by pre-processing algorithms of denoising, background removal, and fringe normalization. The interested reader may refer to the book by Servin et al.<sup>148</sup> for further details.

**Temporal phase demodulation:** Temporal phase-demodulation techniques detect the phase distribution from the temporal variation of fringe signals, as typified by heterodyne interferometry<sup>149</sup> and phase-shifting techniques<sup>150</sup>. Many phase-shifting algorithms have originally been proposed for optical interferometry/holography and later been adapted and extended to fringe projection, for example, standard N-step phase-shifting algorithm<sup>151</sup>, Hariharan 5-step algorithm<sup>21</sup>, 2 + 1 algorithm<sup>152</sup> etc. The interested reader may refer to the chapter “Phase shifting interferometry”<sup>153</sup> of the book edited by Malacara<sup>4</sup> and the review article by Zuo et al.<sup>32</sup> for more details about phase-shifting techniques in the contexts of optical interferometry and FPP, respectively.

- **Phase unwrapping:** No matter which phase-demodulation technique is used, the retrieved phase distribution is mathematically wrapped to the principal value of the arctangent function ranging between  $-\pi$  and  $\pi$ . The result is what is known as a wrapped phase image, and phase unwrapping has to be performed to remove any  $2\pi$ -phase discontinuities. Phase unwrapping algorithms can be broadly classified into three categories:

**Spatial phase unwrapping:** Spatial phase unwrapping methods use only a single wrapped phase map to retrieve the corresponding unwrapped phase distribution, and the unwrapped phase of a given pixel is derived based on the adjacent phase values. Representative methods include Goldstein’s method<sup>154</sup>, reliability-guided method<sup>155</sup>, Flynn’s method<sup>156</sup>, minimal Lp-norm method<sup>157</sup>, and phase unwrapping max-flow/min-cut (PUMA) method<sup>158</sup>. The interested reader may refer to the book by Ghiglia et al. for more technical details. There are also many reviews on the performance comparisons of different unwrapping algorithms for specific applications<sup>159–161</sup>. Limited by the assumption of phase continuity, spatial phase unwrapping methods cannot fundamentally address the inherent fringe order ambiguity problem when the phase difference between neighboring pixels is greater than  $\pi$ .

**Temporal phase unwrapping:** To remove the phase ambiguity, temporal phase unwrapping methods generally generate different or synthetic wavelengths by adjusting flexible system parameters (wavelength, angular separation of light sources, spatial frequency, orientation of the projected fringe patterns) step by step, so that the object can be covered by fringes with different periods. Representative temporal phase unwrapping algorithms include gray-code methods<sup>162,163</sup>, multi-frequency (hierarchical) methods<sup>164–166</sup>, multi-wavelength (heterodyne) methods<sup>167–169</sup>, and number-theoretical methods<sup>170–173</sup>. For more detailed information about these methods, the reader can refer to the

comparative review by Zuo et al.<sup>174</sup> The advantage of temporal phase unwrapping lies in that the unwrapping is neighborhood-independent and proceeds along the time axis on the pixel itself, enabling an absolute evaluation of the mod- $2\pi$  phase distribution.

**Geometric phase unwrapping:** Geometric phase unwrapping approaches can solve the phase ambiguity problem by exploiting the epipolar geometry of projector–camera systems. If the measurement volume can be predefined, depth constraints can be incorporated to preclude some phase ambiguities corresponding to the candidates falling out of the measurement range<sup>175–185</sup>. Alternatively, an adaptive depth-constraint strategy can provide pixel-wise depth constraint ranges according to the shape of the measured object<sup>186</sup>. By introducing more cameras, tighter geometry constraints can be enforced so as to guarantee the unique correspondence and improve the unwrapping reliability<sup>185,187</sup>.

In stereomatching techniques, image analysis refers to determining (tracking or matching) the displacement vector of each pixel point between a pair of acquired images, i.e.,  $f_{anal} : (I_r, I_d) \rightarrow (D_x, D_y)$ . In the routine implementation for DIC and stereophotogrammetry, a region of interest (ROI) or subset in the image is specified at first. The subset is further divided into an evenly spaced virtual grid. The similarity is evaluated at each point of the virtual grid in the reference image to obtain the displacement between two subsets. A full-field displacement map can be obtained by sliding the subset in the searching area of the reference image and obtaining the displacement at each location.

- **Subset correlation:** In DIC, to quantitatively evaluate the similarity or difference between the selected reference subset and the target subset, several correlation criteria have been proposed, such as cross-correlation (CC), the sum of absolute difference (SAD), the sum of squared difference (SSD), zero-mean normalized cross-correlation criterion (ZNCC), zero-mean normalized sum of squared difference (ZNSSD), and the parametric sum of squared difference (PSSD)<sup>188–190</sup>. The subsequent matching procedure is realized by identifying the peak (or valley) position of the correlation coefficient distribution based on certain optimization algorithms. In stereophotogrammetry, nonparametric costs rely on the local ordering (i.e., Rank<sup>191</sup>, Census<sup>192</sup>, and Ordinal measures<sup>193</sup>) of intensity values, which are more frequently used due to their robustness against radiometric changes and outliers, especially near object boundaries<sup>192–194</sup>.

- **Subpixel refinement:** The subset correlation methods mentioned above can only provide integer-pixel displacements. To further improve the measurement resolution and accuracy, many

subpixel refinement methods were developed, including intensity interpolation (i.e., the coarse–fine search method)<sup>195,196</sup>, correlation coefficient curve-fitting<sup>133,197</sup>, gradient-based method<sup>198,199</sup>, Newton–Raphson (NR) algorithm<sup>135,200,201</sup>, and inverse compositional Gauss–Newton (IC-GN) algorithm<sup>202–204</sup>. Among these algorithms, NR and IC-GN are most commonly used for their high registration accuracy and effectiveness in handling high-order surface transformations. However, they suffer from expensive computation cost stemming from their iterative nonlinear optimization and repeated subpixel interpolation. Therefore, accurate initial guesses obtained by integer-pixel subset correlation methods are critical to ensure the rapid convergence<sup>205</sup> and reduce the computational cost<sup>206</sup>. In stereovision, the matching algorithms can be classified as local<sup>207–209</sup>, semi-global<sup>210</sup>, and global methods<sup>211</sup>. Local matching methods utilize the intensity information of a local subset centered at the pixel to be matched. Global matching methods take the result obtained by local matching methods as the initial value and then optimize the disparity by minimizing a predefined global energy function. Semi-global matching methods reduce the 2D global energy minimization problem into a 1D one, enabling faster and more efficient implementations of stereomatching.

### Postprocessing

In optical metrology, the main task of postprocessing is to further refine the measured phase or retrieved displacement field, and finally transform them into the desired physical quantity of the measured object, i.e., the corresponding operator  $f_{post} : \phi / (D_x, D_y) \rightarrow q$ , where  $q$  is the desired sample quantity.

- **Denoising:** Instead of applying to raw fringe patterns, image denoising can also be used as a postprocessing algorithm to remove noise directly from the retrieved phase distribution. Various phase denoising algorithms have been proposed, such as least-square (LS) fitting<sup>212</sup>, anisotropic average filter<sup>213</sup>, WFT<sup>214</sup>, total variation<sup>215</sup>, and nonlocal means filter<sup>216</sup>.
- **Digital refocusing:** The numerical reconstruction of propagating wavefronts by diffraction is a unique feature of digital holography. Since the hologram of the object may not be recorded in the in-focus plane. Numerical diffraction or backpropagation algorithms (e.g., Fresnel diffraction and angular spectrum methods) should be used to obtain a focused image by performing a plane-by-plane refocusing after the image acquisition<sup>217–219</sup>.

- **Error compensation:** There are various types of phase errors associated with optical metrology systems, such as phase-shifting error, intensity nonlinearity, and motion-induced error, which can be compensated with different types of postprocessing algorithms<sup>60,220,221</sup>. In digital holographic microscopy, the microscope objective induces additional phase curvature on the measured wavefront, which needs to be compensated in order to recover the phase information induced by the sample. Typical numerical phase-aberration compensation methods include double exposure<sup>222</sup>, 2D spherical fitting<sup>223</sup> Zernike polynomials fitting<sup>224</sup>, Fourier spectrum filtering<sup>225</sup>, and principal component analysis (PCA)<sup>226</sup>.
- **Quantity transformation:** The final step of postprocessing and also the whole measurement chain is to convert the phase or displacement field into the desired sample quantity, such as height, thickness, displacement, stress, strains, and 3D coordinates, based on sample parameters (e.g., refractive index, relative stress constant) or calibrated system parameters (e.g., sensitivity vector and camera (intrinsic, extrinsic) parameters). The optical setup should be carefully designed to optimize the sensitivity with respect to the measuring quantity in order to achieve a successful and efficient measurement<sup>227,228</sup>.

Finally, it should be mentioned that since optical metrology is a rapidly expanding field in both its scientific foundations and technological developments, the image-processing hierarchy used here cannot provide full coverage of all relevant methods and technologies. For example, phase retrieval and wave-field sensing technologies have shown great promise for inexpensive, vibration-tolerant, non-interferometric, optical metrology of optical surfaces and systems<sup>66,67</sup>. These methods constitute an important aspect of computational imaging as they often involve solving ill-posed inverse problems. There are also some optical metrology methods based on solving constrained optimization problems with added penalties and relaxations (e.g., RPT phase demodulation<sup>144,145</sup> and minimal Lp-norm phase unwrapping methods<sup>157</sup>), which may make pre- and postprocessing unnecessary. For a detailed discussion on this topic, please refer to the subsection “Solving inverse optical metrology problems: issues and challenges”.

### Brief introduction to deep learning

Deep learning is a subset of machine learning, which is defined as the use of specific algorithms that enable machines to automatically learn patterns from large amounts of historical data, and then utilize the uncovered patterns to make predictions about the future or enable



decision making under uncertain intelligently<sup>229,230</sup>. The key specific algorithm used in machine learning is the ANN, which exploits input data  $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^n$  to predict an unknown output  $\mathbf{y} \in \mathcal{Y}$ . The tasks accomplished by the ANN can be broadly divided as classification tasks or regression tasks, depending on whether  $\mathbf{y}$  is a discrete label or a continuous value. The objective of machine learning is then to find a mapping function  $f: \mathbf{x} \rightarrow \mathbf{y}$ . The choice of such functions is given by the neural network models with additional parameters  $\boldsymbol{\theta} \in \Theta$ : i.e.,  $\hat{\mathbf{y}} = f(\mathbf{x}, \boldsymbol{\theta}) \approx \mathbf{y}$ . The goal of this section is to provide a brief introduction to deep learning, as a preparation for the introduction of its applications in optical metrology later on.

### Artificial neural network (ANN)

Inspired by the biological neural network (Fig. 4a), ANNs are composed of interconnected computational units called artificial neurons. As illustrated in Fig. 4b, the simplest neural network following the above concept is the perceptron, which consists of only one single artificial neuron<sup>231</sup>. An artificial neuron takes a bias  $b$  and weight vector  $\mathbf{w} = (w_1, w_2, \dots, w_n)^T$  as parameters  $\boldsymbol{\theta} = (b, w_1, w_2, \dots, w_n)^T$  to map the input  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$  to the output  $f_p(\mathbf{x})$  through a nonlinear activation function  $\sigma$  as

$$f_p(\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x} + b) \quad (3)$$

Typical choices for such activation functions are the sign function  $\sigma(x) = \text{sgn}(x)$ , sigmoid function  $\sigma(x) = \frac{1}{1+e^{-x}}$ , hyperbolic tangent function  $\sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$  and rectified linear unit (ReLU)  $\sigma(x) = \max(0, x)$ <sup>232</sup>. A single perceptron can only model a linear function, but because of the activation functions and in combination with other neurons, the modeling capabilities will increase dramatically. Arranged in a single layer, it has already been shown that neural networks can approximate any continuous function  $f(\mathbf{x})$  on a compact subset of  $\mathbb{R}^n$ . A single-layer network, also called single-layered perceptron (SLP), is represented as a linear combination of  $M$  individual neurons:

$$f_{1NN}(\mathbf{x}) = \sum_{i=1}^M v_i \sigma(\mathbf{w}_i^T \mathbf{x} + b_i) \quad (4)$$

where  $v_i$  is the combination weight of the  $i$ th neuron. We can further extend the mathematical specification of SLP by stacking several single-layer networks into a multi-layered perceptron (MLP)<sup>233</sup>. As the network goes deeper (number of layers increase), the number of free parameters increases, as well as the capability of the network

to represent highly nonlinear functions<sup>234</sup>. We can formalize this mathematically by stacking several single-layer networks into a deep neural network (DNN) with  $N$  layers, i.e.

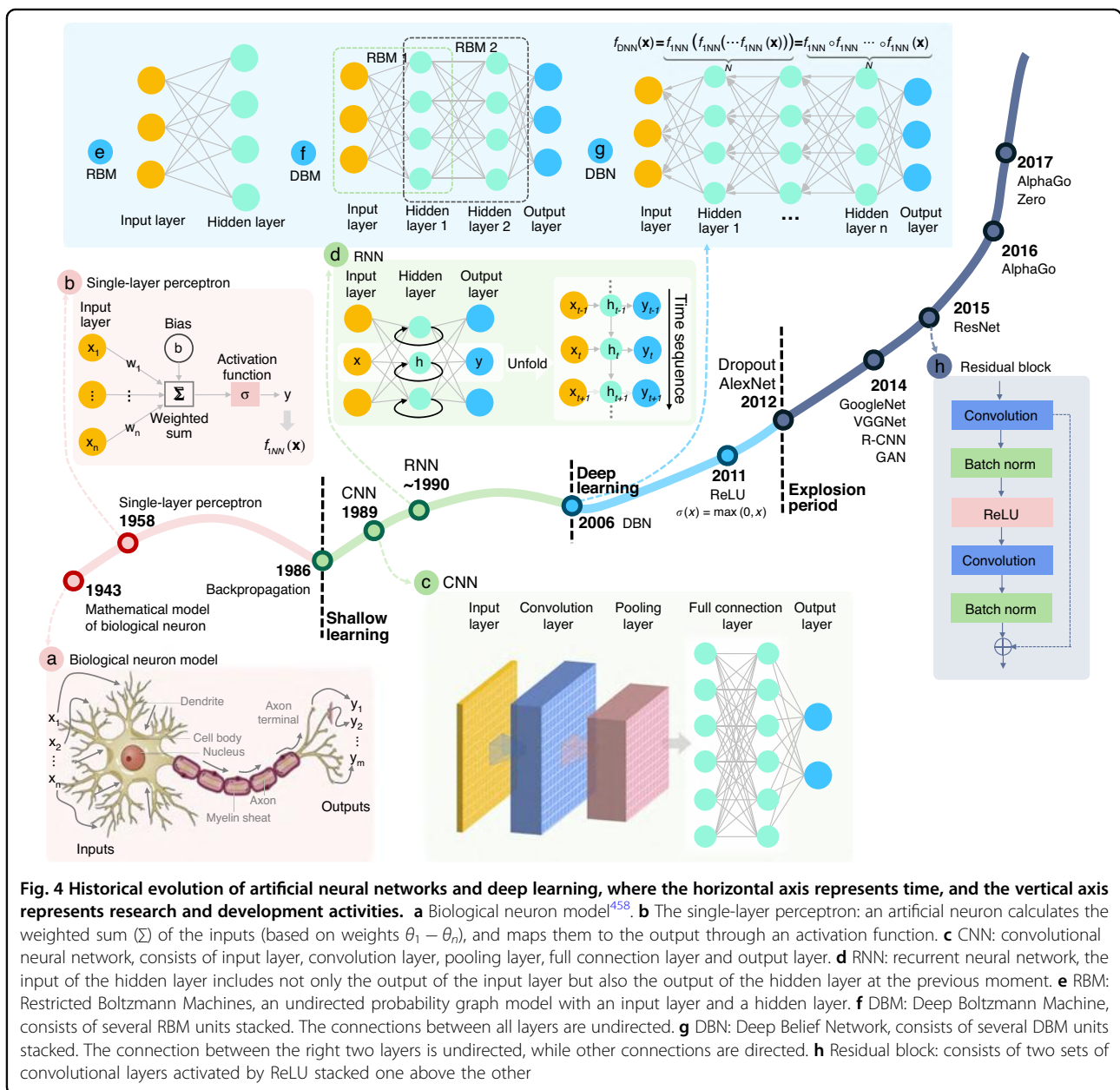
$$f_{DNN}(\mathbf{x}) = \underbrace{f_{1NN}(f_{1NN}(\dots f_{1NN}(\mathbf{x})))}_N = \underbrace{f_{1NN} \circ f_{1NN} \dots \circ f_{1NN}(\mathbf{x})}_N \quad (5)$$

where the circle  $\circ$  is the symbol for the composition of functions. The first layer is referred to as the input layer, the last as the output layer, and the layers in between the input and output are termed as hidden layers. We refer to these using the term “deep”, when a neural network contains many hidden layers, hence the term “deep learning”.

### Neural network training

Having gained basic insights into neural networks and their basic topology, we still need to discuss how to train the neural network, i.e., how its parameters  $\boldsymbol{\theta}$  are actually determined. In this regard, we need to select the appropriate model topology for the problem to be solved and specify the various parameters associated with the model (known as “hyper-parameters”). In addition, we need to define a function that assesses the quality of the network parameter set  $\boldsymbol{\theta}$ , the so-called loss function  $L$ , which quantifies the error between the predicted value  $\hat{\mathbf{y}} = f_{\boldsymbol{\theta}}(\mathbf{x})$  and the true observation  $\mathbf{y}$  (label)<sup>235</sup>.

Depending on the type of task accomplished by the network, the loss function can be divided into classification loss and regression loss. Commonly used classification loss functions include hinge loss ( $L_{\text{Hinge}} = \sum_{i=1}^n \max[0, 1 - \text{sgn}(y_i)\hat{y}_i]$ ) and cross-entropy loss  $L_{\text{CE}} = -\sum_{i=1}^n [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$ <sup>236</sup>. Since the optical metrology tasks involved in this review mainly belong to regression tasks, here we focus on the regression loss functions. The mean absolute error (MAE) loss ( $L_{\text{MAE}} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$ ) and the mean squared error (MSE) loss ( $L_{\text{MSE}} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ ) are the two most commonly used loss functions, which are also known as  $L1$  loss and  $L2$  loss, respectively. In image-processing tasks, MSE is usually converted into a peak signal-to-noise ratio (PSNR) metric:  $L_{\text{PSNR}} = 10 \log_{10} \frac{\text{MAX}^2}{L_{\text{MSE}}}$ , where  $\text{MAX}$  is the maximum pixel intensity value within the dynamic range of the raw image<sup>237</sup>. Other variants of  $L1$  and  $L2$  loss include RMSE, Euclidean loss, smooth  $L1$ , etc.<sup>238</sup>. For natural images, the structural similarity (SSIM) index is a representative image fidelity measurement, which judges the structural similarity of two images based on three metrics (luminance, contrast, and structure):  $L_{\text{SSIM}} = l(\mathbf{y}, \hat{\mathbf{y}})c(\mathbf{y}, \hat{\mathbf{y}})s(\mathbf{y}, \hat{\mathbf{y}})$ <sup>239</sup>, where  $l(\mathbf{y}, \hat{\mathbf{y}})$ ,  $c(\mathbf{y}, \hat{\mathbf{y}})$ , and  $s(\mathbf{y}, \hat{\mathbf{y}})$  are the similarities of the local



patch luminances, contrasts, and structures, respectively. For more details about these loss functions, readers may refer to the article by Wang and Bovik<sup>240</sup>. With the defined loss function, the objective behind the training process of ANNs can be formalized as an optimization problem<sup>241</sup>

$$\hat{\theta} = \arg \min_{\theta \in \Theta} L(f_{\theta}(\mathbf{x}, \mathbf{y})) \quad (6)$$

The learning schemes can be broadly classified into three categories, supervised learning, semi-supervised learning, and unsupervised learning<sup>36,242–244</sup>. Supervised learning dominates the majority of practical applications,

in which a neural network model is optimized based on a large amount dataset of labeled data pairs  $(\mathbf{x}, \mathbf{y})$ , and the training process amounts to find the model parameters  $\hat{\theta}$  that best predict the data based on the loss function  $L(\hat{\mathbf{y}}, \mathbf{y})$ . In unsupervised learning, training algorithms process input data  $\mathbf{x}$  without corresponding labels  $\mathbf{y}$ , and the underlying structure or distribution in the data has to be modeled based on the input itself. Semi-supervised learning sits in between both supervised and unsupervised learning, where a large amount of input data  $\mathbf{x}$  is available and only some of the data is labeled. More detailed discussions about semi-supervised and unsupervised learning can be found in the “Future directions” section.

### From perceptron to deep learning

As summarized in Fig. 4, despite the overall upward trend, a broader look at the history of deep learning reveals three major waves of development. Concepts of machine learning and deep learning commenced with the research into the artificial neural network, which was originated from the simplified mathematical model of biological neurons established by McCulloch and Pitts in 1943<sup>245</sup>. In 1958, Rosenblatt<sup>231</sup> proposed the idea of perceptron, which was the first ANN that allows neurons to learn. The emergence of perceptron marked the first peak of neural network development. However, a single-layer perceptron model can only solve linear classification problems and cannot solve simple XOR and XNOR problems<sup>246</sup>. These limitations caused a major dip in their popularity and stagnated the development of neural networks for nearly two decades.

In 1986, Rumelhart et al.<sup>247</sup> proposed the idea of a backpropagation algorithm (BP) for MLP, which constantly updates the network parameters to minimize the network loss based on a chain rule method. It effectively solves the problems of nonlinear classification and learning, leading neural networks into a second development phase of “shallow learning” and promoting a boom of shallow learning. Inspired by the mammalian visual cortex (stimulated in the restricted visual field)<sup>248</sup>, LeCun et al.<sup>249</sup> proposed the biologically inspired CNN model based on the BP algorithm in 1989, establishing the foundation of deep learning for modern computer vision. During this wave of development, various models like long short-term memory (LSTM) recurrent neural network (RNN), distributed representation, and processing were developed and continue to remain key components of various advanced applications of deep learning to this date. Adding more hidden layers to the network allows a deep architecture to be built, which can accomplish more complex mappings. However, training such a deep network is not trivial because once the errors are back-propagated to the first few layers, they become negligible (so-called gradient vanishing), making the learning process very slow or even fails<sup>250</sup>. Moreover, the limited computational capacity of the available hardware at that time could not support training large-scale neural networks. As a result, deep learning suffered a second major roadblock.

In 2006, Hinton et al.<sup>251,252</sup> proposed a Deep Belief Network (DBN) (the composition of simple, unsupervised networks such as Deep Boltzmann Machines (DBMs)<sup>253</sup> (Fig. 4f) or Restricted Boltzmann Machines (RBMs)<sup>254</sup> (Fig. 4e)) training approach based on the brain graphical models, trying to overcome the gradient-vanishing problem. They gave the new name “deep learning” to multilayer neural network-related learning methods<sup>251,252</sup>. This milestone revolutionized the approaching prospects

in machine learning, leading neural networks into the third upsurge along with the development of computer hardware performance, the development of GPU acceleration technology, and the availability of massive labeled datasets.

In 2012, Krizhevsky et al.<sup>255</sup> proposed a deep CNN architecture — AlexNet, which won the 2012 ImageNet competition, making CNN<sup>249,256</sup> become the dominant framework for deep learning after more than 20 years of silence. Meanwhile, several new deep-learning network architectures and training approaches (e.g., ReLU<sup>232</sup> given by  $\sigma(x) = \max(0, x)$ , and Dropout<sup>257</sup> that discards a small but random portion of the neurons during each iteration of training to prevent neurons from co-adapting to the same features) were developed to further combat the gradient vanishing and ensure faster convergence. These factors have led to the explosive growth of deep learning and its applications in image analysis and computer vision-related problems. Different from CNN, RNN is another popular type of DNN inspired by the brain’s recurrent feedback system. It provides the network with additional “memory” capabilities for previous data, where the inputs of the hidden layer consist of not only the current input but also the output from the previous step, making it a framework specialized in processing sequential data<sup>258–260</sup> (Fig. 4d). CNNs and RNNs usually operate on Euclidean data like images, videos, texts, etc. With the diversification of data, some non-Euclidean graph-structured data, such as 3D-point clouds and biological networks, are also considered to be processed by deep learning. Graph neural networks (GNNs), where each node aggregates feature vectors of its neighbors to compute its new feature vector (a recursive neighborhood aggregation scheme), are effective graph representation learning frameworks specifically for non-Euclidean data<sup>261,262</sup>.

With the focus of more attention and efforts from both academia and industry, different types of deep neural networks have been continuously proposed in recent years with exponential growth, such as VGGNet<sup>263</sup> (VGG means “Visual Geometry Group”), GoogLeNet<sup>264</sup> (using “GoogLe” instead of “Google” is a tribute to LeNet, one of the earliest CNNs developed by LeCun<sup>256</sup>), R-CNN (regions with CNN features)<sup>265</sup>, generative adversarial network (GAN)<sup>266</sup>, etc. In 2015, the emergence of the residual block (Fig. 4h), containing two convolutional layers activated by ReLU that allow the information (from the input or those learned in earlier layers) to penetrate more into the deeper layers, significantly reduces the vanishing gradient problem as the network gets deeper, making it possible to train large-scale CNNs efficiently<sup>267</sup>. In 2016, the Google-owned AI company DeepMind shocked the world by beating Lee Se-dol with its AlphaGo AI system, alerting the world to deep learning, a new



breed of machine learning that promised to be smarter and more creative than before<sup>268</sup>. For a more detailed description of the history and development of deep learning, readers can refer to the chronological review article by Schmidhuber<sup>39</sup>.

### Convolutional neural network (CNN)

In the subsection “Artificial neural network”, we talked about the simplest DNN, so-called MLPs, which basically consist of multiple layers of neurons, each fully connected to those in the adjacent layers. Each neuron receives some inputs, which are multiplied by their weights, with non-linearity applied via activation functions. In this subsection, we will talk about CNNs, which are considered an evolution of the MLP architecture that is developed to process data in single or multiple arrays, and thus are more appropriate to handle image-like input. Given the prevalence of CNNs in image processing and analysis tasks, here we briefly review some basic ideas and concepts widely used in CNNs. For a comprehensive introduction to CNN, we refer readers to the excellent book by Goodfellow et al.<sup>36</sup>.

CNN follows the same pattern as MLP: artificial neurons are stacked in hidden layers on top of each other; parameters are learned during network training with nonlinearity applied via activation functions; the loss  $L(\hat{\mathbf{y}}, \mathbf{y})$  is calculated and back-propagated to update the network parameters. The major difference between them is that instead of regular fully connected layers, CNN uses specialized convolution layers to model locality and abstraction (Fig. 5b). At each layer, the input image  $\mathbf{x}$  (lexicographically ordered) is convolved with a set of convolutional filters  $\mathbf{W}$  (note here  $\mathbf{W}$  represents block-Toeplitz convolution matrix) and added biases  $\mathbf{b}$  to generate a new image, which is subjected to an elementwise nonlinear activation function  $\sigma$  (normally use ReLU function  $\sigma(x) = \max(0, x)$ ), and the same structure is repeated for each convolution layer  $k$ :

$$\mathbf{x}^k = \sigma(\mathbf{W}^{k-1}\mathbf{x}^{k-1} + \mathbf{b}^{k-1}) \quad (7)$$

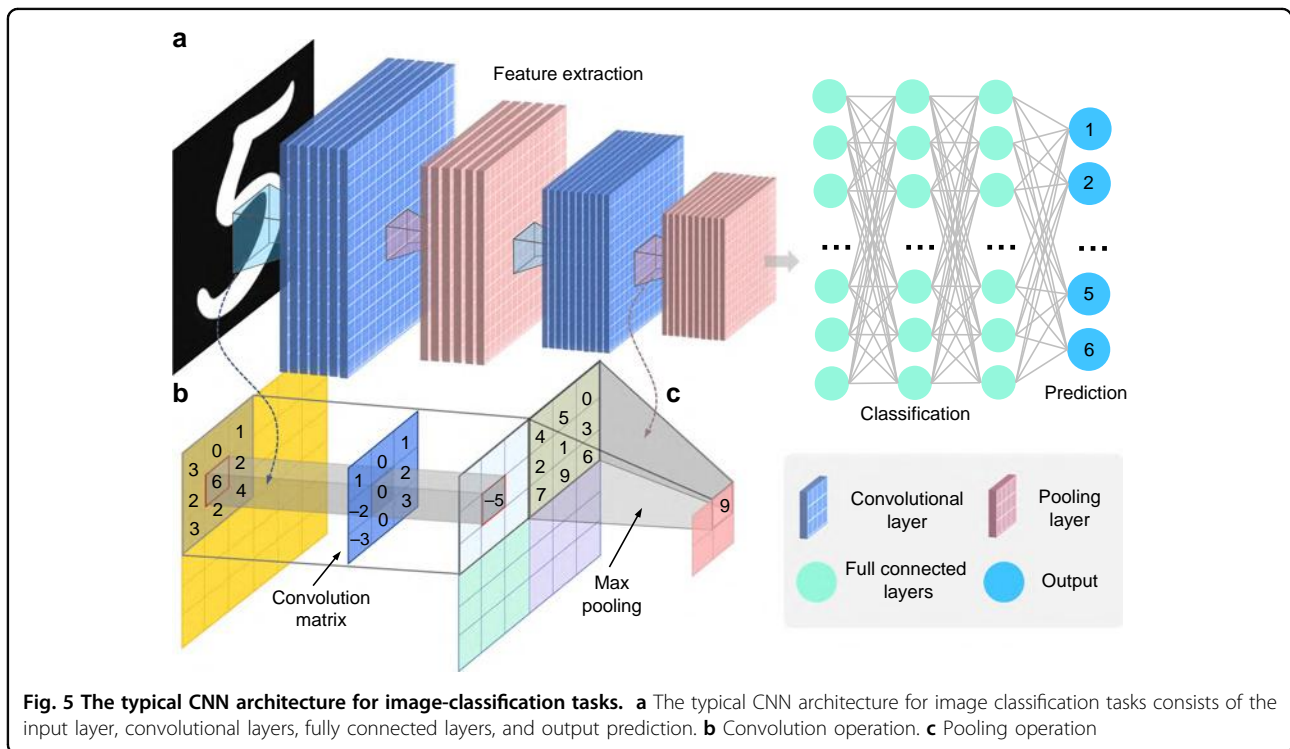
The second key difference between CNNs and MLPs is the typical incorporation of pooling layers in CNNs, where pixel values of neighborhoods are aggregated by applying a permutation invariant function, such as the max or mean operation, to reduce the dimensionality of the convolutional layers and allows significant features to propagate downstream without being affected by neighboring pixels (Fig. 5c). The major advantage of such an architecture is that CNNs exploit spatial dependencies in the image and only consider a local neighborhood for each neuron, i.e., the network parameters are shared in such a way that the network performs convolution

operations on images. In other words, the idea of a CNN is to take advantage of a pyramid structure to first identify features at the lowest level before passing these features to the next layer, which, in turn, create features of a higher level. Since the local statistics of images are invariant to location, the model does not need to learn weights for the same feature occurring at different positions in an image, making the network equivariant with respect to translations of the input. It makes CNNs especially suitable for processing images captured in optical metrology, e.g., a fringe pattern consisting of sinusoidal signal repeated over different image locations. In addition, it also drastically reduces the number of parameters (i.e., the number of weights no longer depends on the size of the input image) that need to be learned.

Figure 5a shows a CNN architecture for the image-classification task. Every layer of a CNN transforms the input volume to an output volume of neuron activation, eventually leading to the final fully connected layers, resulting in a mapping of the input data to a 1D feature vector. A typical CNN configuration consists of a sequence of convolution and pooling layers. After passing through a few pairs of convolutional and pooling layers, all the features of the image have been extracted and arranged into a long tube. At the end of the convolutional stream of the network, several fully connected layers (i.e., regular neural network architecture, MLP, that discussed in the previous subsection) are usually added to fatten the features into a vector, with which tasks, such as classifications, can be performed. Starting with LeNet<sup>256</sup>, developed in 1998 for recognizing handwritten characters with two convolutional layers, CNN architectures have evolved since then to deeper CNNs like AlexNet<sup>264</sup> (5 convolutional Layers) and VGGNet<sup>263</sup> (19 convolutional Layers) and beyond to more advanced and super-deep networks like GoogLeNet<sup>264</sup> and ResNet<sup>267</sup>, respectively. These CNNs have been extremely successful in computer vision applications, such as object detection<sup>269</sup>, action recognition<sup>270</sup>, motion tracking<sup>271</sup>, and pose estimation<sup>272</sup>.

### Fully convolutional network architectures for image processing

Conventionally, CNNs have been used for solving classification problems. Due to the presence of a parameter-rich fully connected layer at the end of the network, typical CNNs throw away spatial information and produce non-spatial outputs. However, for most image-processing tasks that we encountered earlier in the Section “Image processing in optical metrology”, the network must have a whole-resolution output with the same or even larger size compared with the input, which is commonly referred to as dense prediction (contrary to the single target category per image)<sup>273</sup>. Specifically, fully convolutional network architectures without fully connected layers should be used for this purpose, which accepts input of any size, is trained with



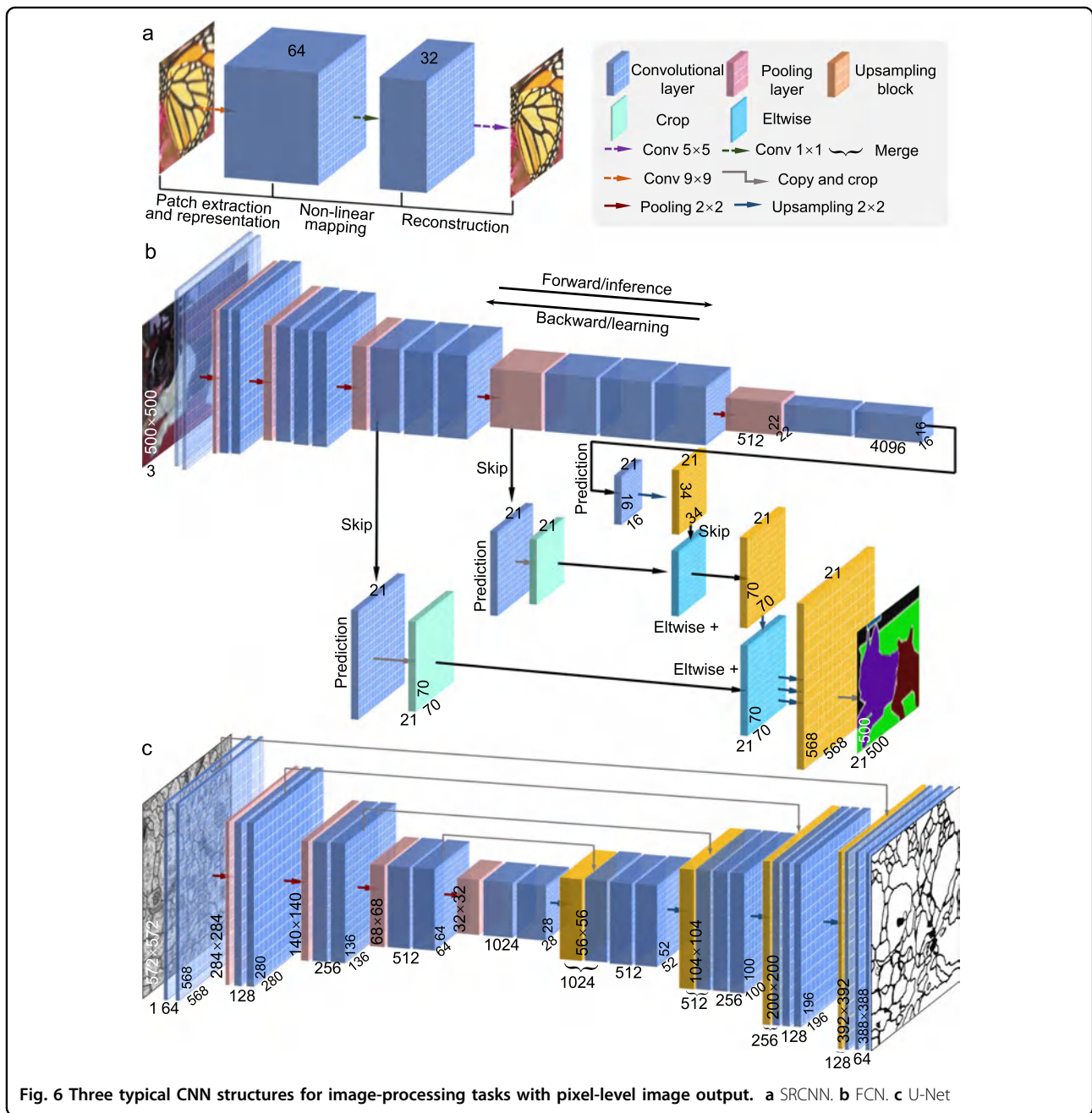
**Fig. 5** The typical CNN architecture for image-classification tasks. **a** The typical CNN architecture for image classification tasks consists of the input layer, convolutional layers, fully connected layers, and output prediction. **b** Convolution operation. **c** Pooling operation

a regression loss, and produces an output of the corresponding dimensions<sup>273,274</sup>. Here, we briefly review three representative network architectures with such features.

- SRCNN:** In conventional CNN, the downsampling effect of pooling layers results in an output with a far lower resolution than the input. Thus, a relatively naive and straightforward solution is simply stacking several convolutions layers while skipping pooling layers to preserve the input dimensions. Dong et al.<sup>275</sup> firstly adopt this idea and propose SRCNN for the image super-resolution task. SRCNN utilizes traditional upsampling algorithms to obtain low-resolution images and then refine them by learning an end-to-end mapping from interpolated coarse images to high-resolution images of the same dimension but with more details, as illustrated in Fig. 6a. Due to its simple ideal and implementation, SRCNN has gradually become one of the most popular frameworks in image super-resolution<sup>276</sup> and been extended to many other tasks such as radar image enhancing<sup>277</sup>, underwater image high definition display<sup>278</sup>, and computed tomography<sup>279</sup>. One major disadvantage of SRCNN is the cost of time and space to keep the whole resolution through the whole network, limiting SRCNN only practical for relatively shallow network structures.
- FCN:** The fully convolutional network (FCN) proposed by Long et al.<sup>273</sup> is a popular strategy and baseline for semantic-segmentation tasks. FCN is

inspired by the fact that the fully connected layers in classification CNN (Fig. 5) can also be viewed as convolutions with kernels that cover their entire input regions. As illustrated in Fig. 6b, FCN uses the existing classification CNN as the encoder module of the network and replace these fully connected layers into  $1 \times 1$  convolution layers (also termed as deconvolution layers) as the decoding module, enabling the CNN to upsample the input feature maps and get pixel-wise output. In FCN, skip connections combining (simply adding) information in fine layers and coarse layers enhances the localization capability of the network, allowing for the reconstruction of accurate fine details that respect global structure. FCN and its variants have achieved great success in the application of dense pixel prediction as required in many advanced computer vision understanding tasks<sup>280</sup>.

- U-Net:** Ronneberger et al.<sup>281</sup> took the idea of FCN one step further and proposed the U-Net architecture, which replaces the one-step upsampling part with a bunch of complimentary upsampling convolutions layers, resulting in a quasi-symmetrical encoder-decoder model architecture. As illustrated in Fig. 6c, the basic structure of U-Net consists of a contractive branch and an expansive branch, which enables multiresolution analysis and general multiscale image-to-image transforms. The contractive branch (encoder) downsamples the image using conventional strided convolution, producing a compressed feature

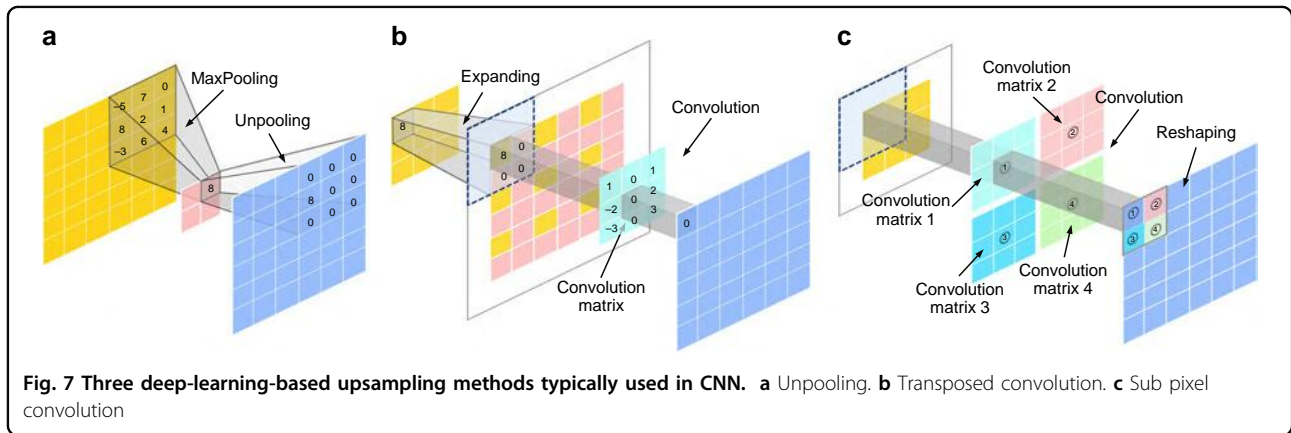


**Fig. 6** Three typical CNN structures for image-processing tasks with pixel-level image output. **a** SRCNN. **b** FCN. **c** U-Net

representation of the input image. The expansive branch (decoder), complimentary to the contractive branch, uses upsampling methods like transpose convolution to provide the processed result with the same size as the input. In addition, U-Net features skip connections that concatenate the matching resolution levels of the contractive branch and the expansive branch. Ronneberger’s U-Net is a breakthrough toward automatic image segmentation and has been successfully applied in many tasks that require image-to-image transforms<sup>282</sup>.

Since the feature extraction is only performed in low-dimensional space, the computation and spatial complexity of the above encoder-decoder structured networks (FCN and U-Net) can be much reduced. Therefore, the encoder-decoder CNN structure has become the mainstream for image segmentation and reconstruction<sup>283</sup>. The encoder is usually a classic CNN (Alexnet, VGG, Resnet, etc.) in which downsampling (pooling layers) is adopted to reduce the input dimension so as to generate low-resolution feature maps. The decoder tries to mirror the encoder to upsample these feature representations





and restore the original size of the image. Thus, how to perform upsampling is of great importance. Although traditional upsampling methods, e.g., nearest neighbor, bilinear, and bicubic interpolations, are easy to implement, deep-learning-based upsampling methods, e.g., unpooling<sup>284</sup>, transpose convolution<sup>273</sup>, subpixel convolution<sup>285</sup>, has gradually become a trend. All these approaches can be combined with the model mentioned above to prevent the decrease in resolution and obtain a full-resolution image output.

- **Unpooling upsampling:** Unpooling upsampling reverts maxpooling by remembering the location of the maxima in the maxpooling layers and in the unpooling layers copy the value to exactly this location, as shown in Fig. 7a.
- **Transposed convolution:** The opposite of the convolutional layers are the transposed convolution layers (also misinterpreted as deconvolution layers<sup>280</sup>), i.e., predicting the possible input based on feature maps sized like convolution output. Specifically, it increases the image resolution by expanding the image by inserting zeros and performing convolution, as shown in Fig. 7b.
- **Sub pixel convolution:** The subpixel layer performs upsampling by generating a plurality of channels by convolution and then reshaping them, as Fig. 7c shows. Within this layer, a convolution is firstly applied for producing outputs with  $M$  times channels, where  $M$  is the scaling factor. After that, the reshaping operation (*a.k.a.* shuffle) is performed to produce outputs with size  $M$  times larger than the original.

As discussed in the Section “Image processing in optical metrology”, despite their diversity, the image-processing algorithms used in optical metrology share a common characteristic—they can be regarded as a mapping operator that transforms the content of arbitrary-sized inputs into pixel-level outputs, which

fits exactly with DNNs with a fully convolutional architecture. In principle, any fully convolutional network architectures presented here can be used for a similar purpose. By applying different types of training datasets, they can be trained for accomplishing different types of image-processing tasks that we encountered in optical metrology. This provides an alternative approach to process images such that the produced results resemble or even outperform conventional image-processing operators or their combinations. There are also many other potential desirable factors for such a substitution, e.g., accuracy, speed, generality, and simplicity. All these factors were crucial to enable the fast rise of deep learning in the field of optical metrology.

### Invoking deep learning in optical metrology: principles and advantages

Let us return to optical metrology. It is essential that the image formation is properly understood in order to reconstruct the required geometrical or mechanical quantities of the sample, as we discussed in Section “Image formation in optical metrology”. In general, the relation between the observed images  $\mathbf{I} \in \mathbb{R}^m$  (frame-stacked lexicographically ordered with  $m \times 1$  in dimension) and the desired sample parameter (or information-bearing parameter that clearly reflects the desired sample quantity, e.g., phase or displacement field)  $\mathbf{p} \in \mathbb{R}^m$  (or  $\mathbb{C}^n$ ) can be described as

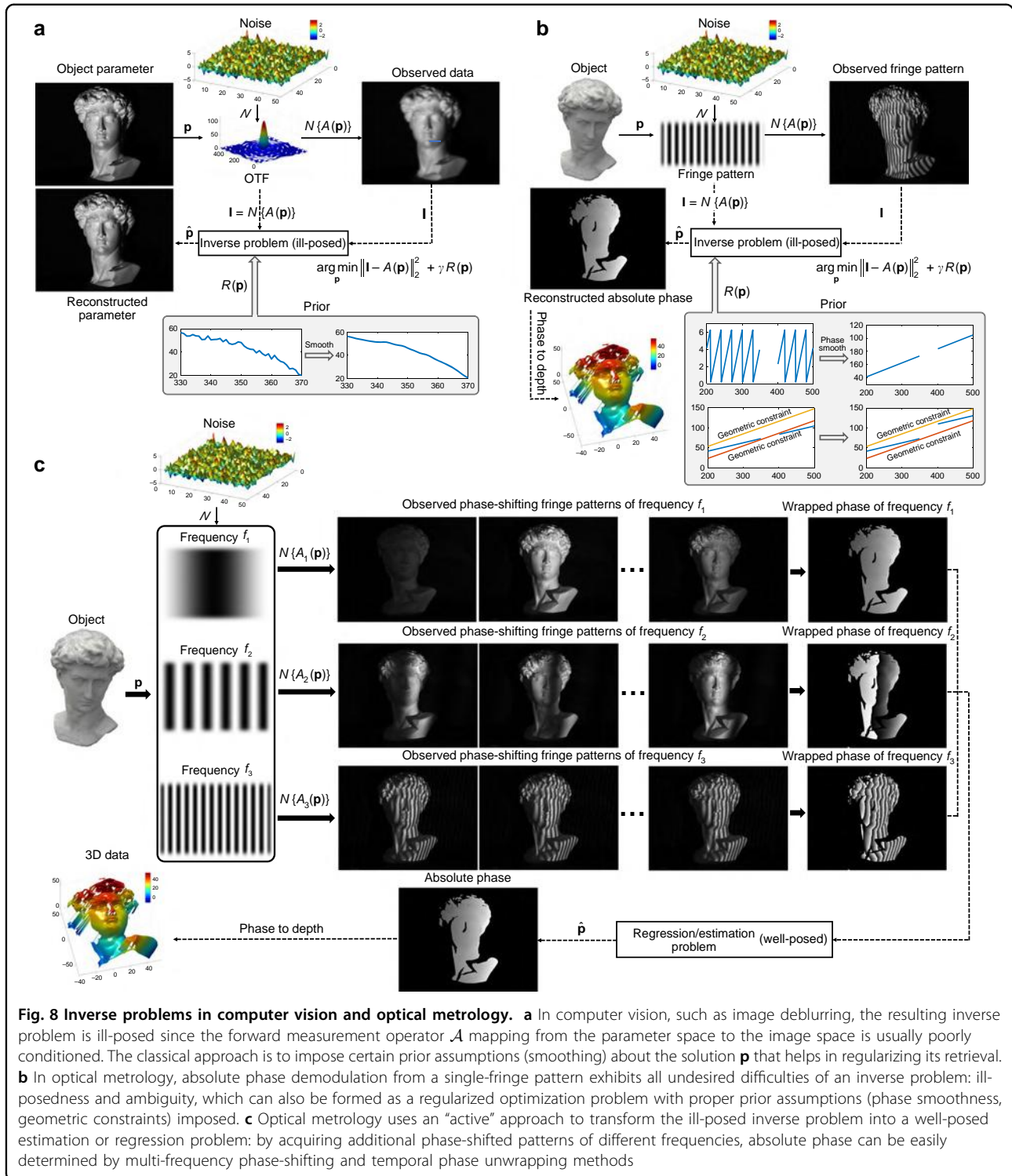
$$\mathbf{I} = \mathcal{N}\{\mathcal{A}(\mathbf{p})\} \tag{8}$$

where  $\mathcal{A}$  is the (possibly nonlinear) forward measurement operator mapping from the parameter space to the image space, which is given by the physics laws governing the formation of data;  $\mathcal{N}$  represents the effect of noise (not necessarily additive). This model seems general enough to cover almost all image formation processes in optical metrology. However, this does not mean that  $\mathbf{p}$  can be

directly obtained from  $\mathbf{I}$ . More specifically, we have to conclude in general from the effect (i.e., the intensity at the pixel) to its cause (i.e., shape, displacement, deformation, or stress of the surface), suggesting that an inverse problem has to be solved.

**Solving inverse optical metrology problems: issues and challenges**

Given the forward model represented by Eq. (8), our task is to find the parameters by an approximate inverse of  $\mathcal{A}$  (denoted as  $\tilde{\mathcal{A}}^{-1}$ ) such that  $\hat{\mathbf{p}} = \tilde{\mathcal{R}}(\mathbf{I}) = \tilde{\mathcal{A}}^{-1}(\mathbf{I}) \approx \mathbf{p}$ .



**Fig. 8 Inverse problems in computer vision and optical metrology.** **a** In computer vision, such as image deblurring, the resulting inverse problem is ill-posed since the forward measurement operator  $\mathcal{A}$  mapping from the parameter space to the image space is usually poorly conditioned. The classical approach is to impose certain prior assumptions (smoothing) about the solution  $\mathbf{p}$  that helps in regularizing its retrieval. **b** In optical metrology, absolute phase demodulation from a single-fringe pattern exhibits all undesired difficulties of an inverse problem: ill-posedness and ambiguity, which can also be formed as a regularized optimization problem with proper prior assumptions (phase smoothness, geometric constraints) imposed. **c** Optical metrology uses an “active” approach to transform the ill-posed inverse problem into a well-posed estimation or regression problem: by acquiring additional phase-shifted patterns of different frequencies, absolute phase can be easily determined by multi-frequency phase-shifting and temporal phase unwrapping methods

However, in real practice, there are many problems involved in this process:

- **Unknown or mismatched forward model.** The success of conventional optical metrology approaches relies heavily on the precise pre-knowledge about the forward model  $\mathcal{A}$ , so they are often regarded as model-driven or knowledge-driven approaches. In practical applications, the forward model  $\mathcal{A}$  used is always an approximate description of reality, and extending it might be challenging due to a limited understanding of experimental perturbations (noise, aberrations, vibration, motion, nonlinearity, saturation, and temperature variations) and non-cooperative surfaces (shiny, translucent, coated, shielded, highly absorbent, and strong scattering). These problems are either difficult to model or result in a too complicated (even intractable) model with a large number of parameters.
- **Error accumulation and suboptimal solution.** As described in the section “Image processing in optical metrology”, “divide-and-conquer” is a common practice for solving complex problems with a sequence of cascaded image-processing algorithms to obtain the desired object parameter. For example, in FPP, the entire image-processing pipeline is generally divided into several sub-steps, i.e., image pre-processing, phase demodulation, phase unwrapping, and phase-to-height conversion. Although each sub-problem or sub-step becomes simpler and easier to handle, the disadvantages are also apparent: error accumulation and suboptimal solution, i.e., the aggregation of optimum solutions to subproblems may not be equivalent to the global optimum solution.
- **Ill-posedness of the inverse problem.** In many computer vision and computational imaging tasks, such as image deblurring<sup>24</sup>, sparse computed tomography<sup>25</sup>, and imaging through scattering media<sup>27</sup>, the difficulty in retrieving the desired information  $\mathbf{p}$  from the observation  $\mathbf{I}$  arises from the fact that the operator  $\mathcal{A}$  is usually poorly conditioned, and the resulting inverse problem is ill-posed, as illustrated in Fig. 8a. Due to the similar indirect measurement principle, there are also many important inverse problems in optical metrology that are ill-posed, among which the phase demodulation from a single-fringe pattern and phase unwrapping from single wrapped phase distributions are the best known for specialists in optical metrology (Fig. 8b). The simplified model for the intensity distribution of fringe patterns (Eq. (1)) suggests that the observed intensity  $\mathbf{I}$  results from the integration of several unknown components: the average intensity  $A(x, y)$ ,

the intensity modulation  $B(x, y)$ , and the desired phase function  $\phi(x, y)$ . Simply put, we do not have enough information to solve the corresponding inverse problem uniquely and stably.

In the fields of computer vision and computational imaging, the classical approach in solving an ill-posed inverse problem is to reformulate the ill-posed original problem into a well-posed optimization problem by imposing certain prior assumptions about the solution  $\mathbf{p}$  that helps in regularizing its retrieval:

$$\hat{\mathbf{p}} = \arg \min_{\mathbf{p}} \|\mathbf{I} - \mathcal{A}(\mathbf{p})\|_2^2 + \gamma R(\mathbf{p}) \quad (9)$$

where  $\|\cdot\|_2$  indicates the Euclidean norm,  $R(\mathbf{p})$  is a regularization penalty function that incorporates the prior information about  $\mathbf{p}$ , such as smoothness<sup>286</sup>, sparsity in some basis<sup>287</sup> or dictionary<sup>288</sup>.  $\gamma$  is a real positive parameter (regularization parameter) that governs the weight given to the regularization against the need to fit the measurement and should be selected carefully to make an admissible compromise between the prior knowledge and data fidelity. Such an optimization problem can be solved efficiently with a variety of algorithms<sup>289,290</sup> and provide theoretical guarantees on the recoverability and stability of the approximate solution to an inverse problem<sup>291</sup>.

Instead of regularizing the numerical solution, in optical metrology, we prefer to reformulate the original ill-posed problem into a well-posed and adequately stable one by actively controlling the image acquisition process so as to add systematically more knowledge about the object to be investigated into the evaluation process<sup>31</sup>. Due to the fact that the optical measurements are frequently carried out in a highly controlled environment, such a solution is often more practical and effective. As illustrated by Fig. 8c, by acquiring additional multi-frequency phase-shifted patterns, absolute phase retrieval becomes a well-posed estimation or regression problem, and the simple standard (unconstrained, regularization-free) least-square methods in regression analysis provides a stable, precise, and efficient solution<sup>292,293</sup>:

$$\hat{\mathbf{p}} = \arg \min_{\mathbf{p}} \|\mathbf{I} - \mathcal{A}(\mathbf{p})\|_2^2 \quad (10)$$

The situation may become very different when we step out of the laboratory and into the complicated environment of the real world<sup>294</sup>. The active strategies mentioned above often impose stringent requirements on the measurement conditions and the object under test. For instance, high-sensitivity interferometric measurement in general needs a laboratory environment where the thermal-mechanical settings are carefully controlled to preserve beam path conditions and minimize external



disturbances. Absolute 3D shape profilometry usually requires multiple fringe pattern projections, which requires that the measurement conditions remain invariant while sequential measurements are performed. However, harsh operating environments where the object or the metrology system cannot be maintained in a steady-state may make such active strategies a luxurious or even unreasonable request. Under such conditions, conventional optical metrology approaches will suffer from severe physical and technical limitations, such as a limited amount of data and uncertainties in the forward model.

To address these challenges, researchers have made great efforts to improve state-of-the-art methods from different aspects over the past few decades. For example, phase-shifting techniques were optimized from the perspective of signal processing to achieve high-precision robust phase measurement and meanwhile minimize the impact of experimental perturbations<sup>32,153</sup>. Single-shot spatial phase-demodulation methods have been explicitly formulated as a constrained optimization problem similar to Eq. (9) with an extra regularization term enforcing a priori knowledge about the recovered phase (spatially smooth, limited spectral extension, piecewise constant, etc.)<sup>140,148</sup>. Multi-frequency temporal phase unwrapping techniques have been optimized by utilizing the inherent information redundancy in the average intensity and the intensity modulation of the fringe images, allowing for absolute phase retrieval with the reduced number of patterns<sup>32,295</sup>. Geometric constraints were introduced in FPP to solve the phase ambiguity problem without additional image acquisition<sup>175,183</sup>. Despite these extensive research efforts for decades, how to extract the absolute (unambiguous) phase information, with the highest possible accuracy, from the minimum number (preferably single shot) of fringe patterns remains one of the most challenging open problems in optical metrology. Consequently, we are looking forward to innovations and breakthroughs in the principles and methods of optical metrology, which are of significant importance for its future development.

### Solving inverse optical metrology problems via deep learning

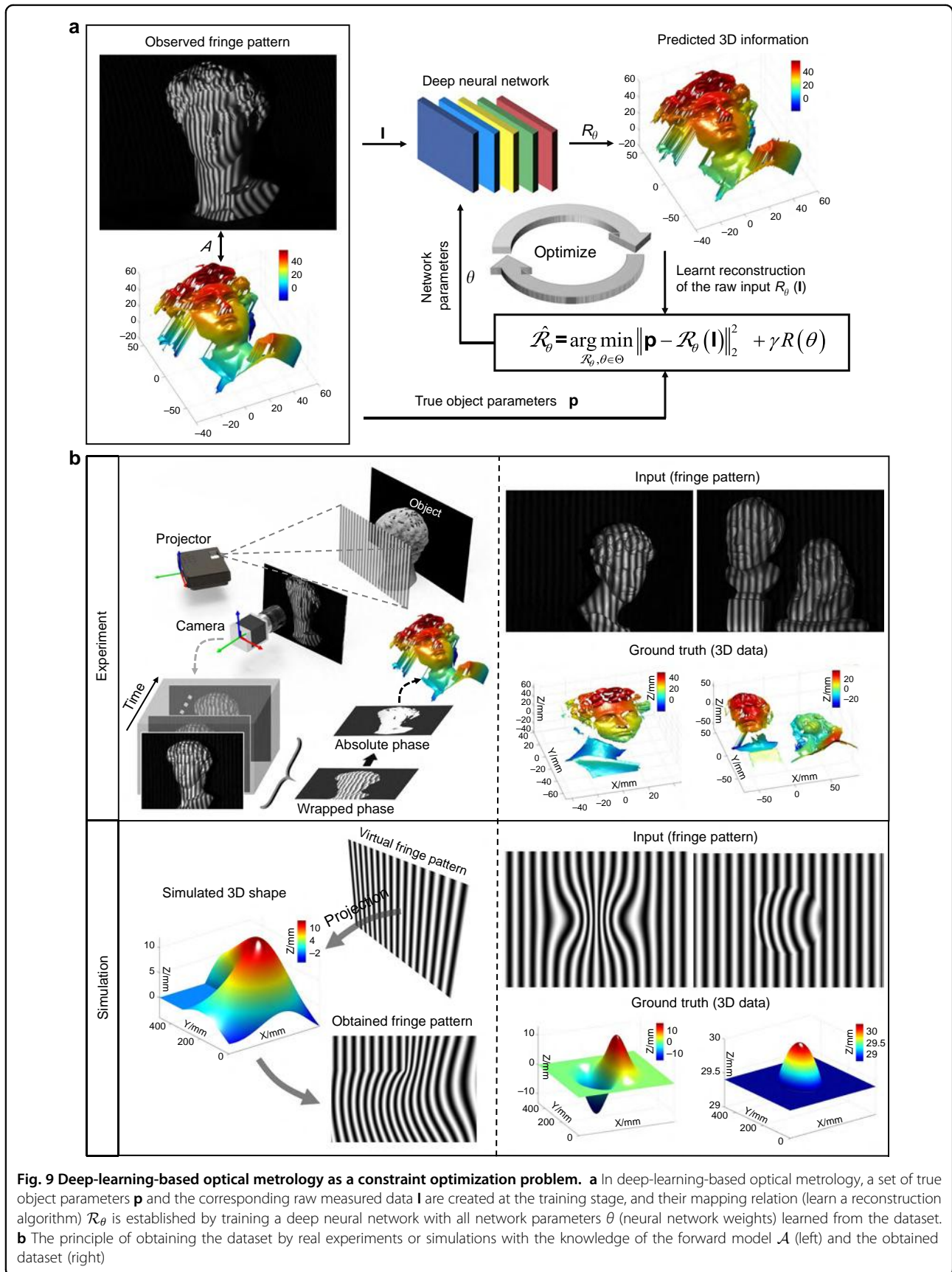
As a “data-driven” technology that has emerged in recent years, deep learning has received increasing attention in the field of optical metrology and made fruitful achievements in very recent years. Different from the conventional physical model and knowledge-driven approaches that the objective function (Eqs. (9) and (10)) is built based on the image formation model  $\mathcal{A}$ , in deep-learning approaches, we create a set of true object parameters  $\mathbf{p}$  and the corresponding raw measured data  $\mathbf{I}$ , and establish their mapping relation  $\mathcal{R}_\theta$  based on a deep

neural network with all network parameters  $\theta$  learned from the dataset by solving the following optimization problem (Fig. 9):

$$\widehat{\mathcal{R}}_\theta = \underset{\mathcal{R}_\theta, \theta \in \Theta}{\operatorname{argmin}} \|\mathbf{p} - \mathcal{R}_\theta(\mathbf{I})\|_2^2 + R(\theta) \quad (11)$$

with  $\|\cdot\|_2^2$  being the  $L_2$ -norm error (loss) function once again (different types of loss functions discussed in the subsection “Neural network training” can be specified depending on the type of training data) and  $R$  is a regularizer of the parameters to avoid overfitting. A key element in deep-learning approaches is to parameterize  $\widehat{\mathcal{R}}_\theta$  by parameters  $\theta \in \Theta$ . The “learning” process refers to finding an “optimal” set of network parameters from the given training data by minimizing Eq. (11) over all possible network parameters  $\theta \in \Theta$ . And the “optimality” is quantified through the loss function that measures the quality of the learned  $\mathcal{R}_\theta$ . Different deep-learning approaches can be thought of as different ways to parameterize the reconstruction network  $\mathcal{R}_\theta$ . Different from conventional approaches that solving the optimization problem directly gives the final solution  $\widehat{\mathcal{R}}_\theta$  to the inverse problem corresponding to a current given input, in deep-learning-based approaches, the optimization problem is phrased as to find a “reconstruction algorithm”  $\widehat{\mathcal{R}}_\theta$  satisfying the pseudo-inverse property  $\widehat{\mathbf{p}} = \widehat{\mathcal{R}}_\theta(\mathbf{I}) = \widehat{\mathcal{A}}^{-1}(\mathbf{I}) \approx \mathbf{p}$  from the prepared (previous) dataset, which is then used for the reconstruction of the future input.

Most of the deep-learning techniques currently used in optical metrology belong to supervised learning, i.e., a matched dataset of ground-truth parameters  $\mathbf{p}$  and corresponding measurements  $\mathbf{I}$  should be created to train the network. Ideally, the dataset should be collected by physical experiments based on the same metrology system to account for all experimental conditions (which are usually difficult to be fully described by the forward image formation model). The ground truth can be obtained by measuring various samples that one is likely to encounter by employing active strategies mentioned above, without considering the ill-posedness of the real problem. To be more precise, in deep-learning-based optical metrology approaches, active strategies frequently used in conventional optical metrology approaches are shifted from the actual measurement stage to the preparation (network training) stage. Although the situation faced during the preparation stage may be different from that in the actual measurement stage, the information obtained in the former can be transferred to the latter in many cases. What we should do during the training stage is to reproduce the sample (using representative test objects), the system (using the same measurement system), and the error sources (noise, vibration, background illumination) during the measurement stage to ensure that the captured input data is as close as possible to those in the real

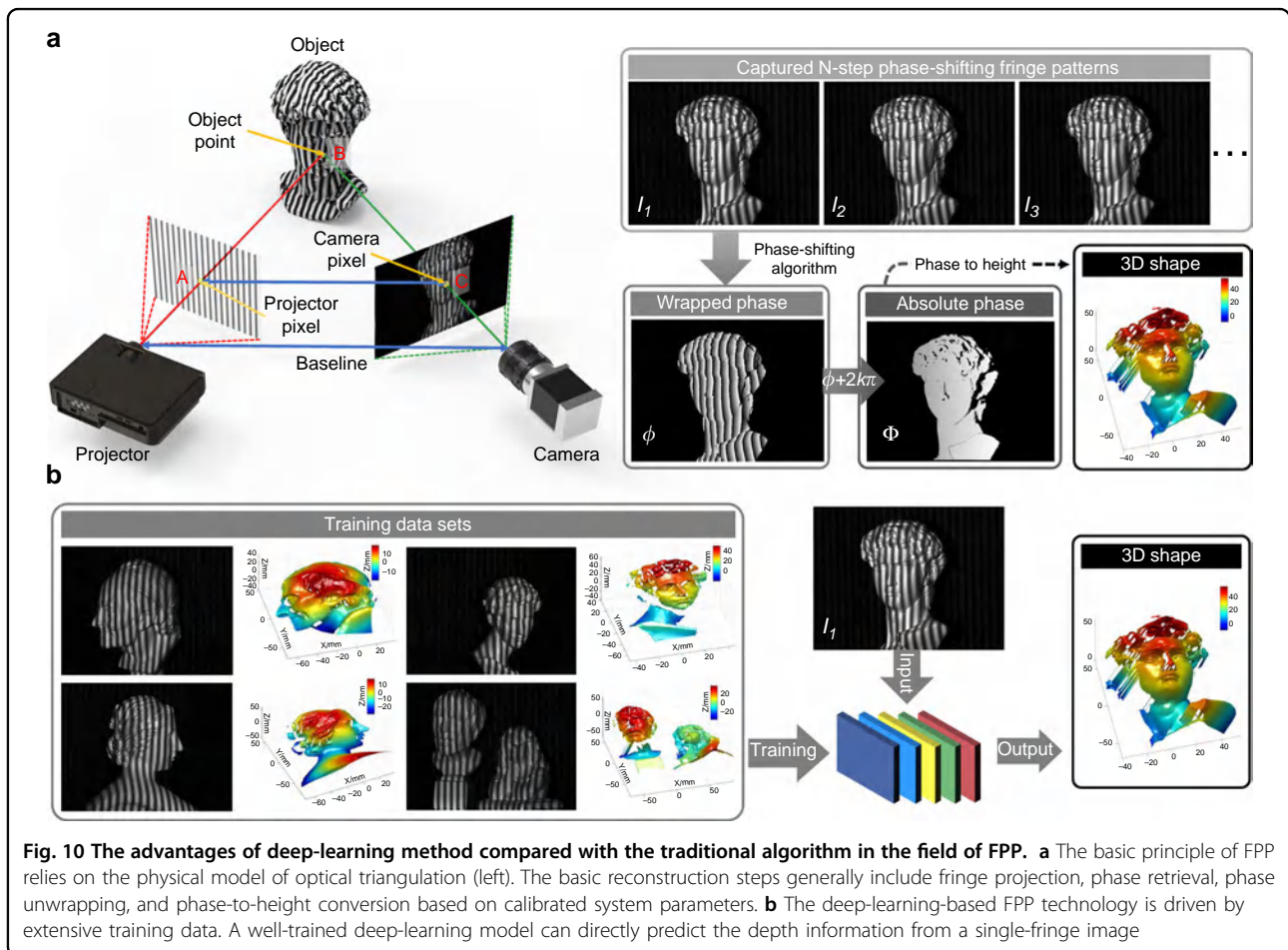


measurement. On the other hand, we should make the remaining environmental variables as controllable as possible so that more active strategies (sample manipulation, illumination changing, multiple acquisitions) can be involved in the training stage to derive the ground truth corresponding to these captured data. Once the network is trained, we can then strip out these ideal environment variables and make the network run in a realistic experimental condition.

For example, for an interferometric system working in a harsh environment or a FPP system designed for measuring dynamic objects, phase demodulation from a single-fringe pattern is the most desirable choice. The inherent ill-posedness of the problem makes it a very good example for deep learning in this regard. In the training stage, we reproduce all the experimental conditions except that we employ the multi-frame phase-shifting technique with large phase-shifting steps to obtain the ground truth for the training samples. Once the network is established, it can map from only one single-fringe pattern to the desired phase distribution, and thus can be used in harsh environments where the single-shot phase-demodulation technique should be applied. Note

that in this example, all the training data is fully generated by experiments, so the reconstruction algorithm (inverse mapping)  $\widehat{\mathcal{R}}_{\theta}$  can be established without the knowledge of the forward model  $\mathcal{A}$  in principle. Even though, since we have sufficient real-world training observations of the form  $(\mathbf{p}, \mathbf{I})$ , it can be expected that those experimental data can reflect the true  $\mathcal{A}$  in a complete and realistic way.

It should be noted that there are also many cases that the ground truth corresponding to the experimental data is inaccessible. In such cases, the matched dataset can be obtained by a “learning from simulation” scheme — simulating the forward operator (with the knowledge of the forward image formation model  $\mathcal{A}$ ) on ideal sample parameters. However, due to the complexity of real experimental conditions, we typically only know an approximation of  $\mathcal{A}$ . Subsequently, the inconsistency or uncertainty in the forward operator  $\mathcal{A}$  may lead to a compromised performance in real experiments (see the “Challenges” section for detailed discussions). On the other hand, partial knowledge of the forward model  $\mathcal{A}$  can be leveraged and incorporated in the deep neural network design to alleviate the “black box” nature of conventional neural network architectures, which may reduce the



**Fig. 10** The advantages of deep-learning method compared with the traditional algorithm in the field of FPP. **a** The basic principle of FPP relies on the physical model of optical triangulation (left). The basic reconstruction steps generally include fringe projection, phase retrieval, phase unwrapping, and phase-to-height conversion based on calibrated system parameters. **b** The deep-learning-based FPP technology is driven by extensive training data. A well-trained deep-learning model can directly predict the depth information from a single-fringe image



amount of required training data and provide more accurate and reliable network reconstruction (see the “Future directions” section for more details).

### Advantages of invoking deep learning in optical metrology

In light of the above discussions, we summarize the potential advantages that can be gained by using a deep-learning approach in optical metrology. Figure 10 shows the advantages of deep-learning techniques compared to traditional optical metrology algorithms by taking FPP as an example. One may have noticed that FPP has appeared a few times, and in fact, it will appear more times. The reason is that FPP is currently one of the most promising and well-researched areas at the intersection of deep learning and optical metrology, offering a representative and convincing example of the use of deep learning in optical metrology.

**(1) From “physics-model-driven” to “data-driven”** Deep learning subverts the conventional “physics-model-driven” paradigm and opens up the “data-driven” learning-based representation paradigm. The reconstruction algorithm (inverse mapping)  $\widehat{\mathcal{R}}_\theta$  can be learned from the experimental data without resorting to the pre-knowledge of the forward model  $\mathcal{A}$ . If the training data is collected under an environment that reproduces the real experimental conditions (including metrology system, sample types, measurement environment, etc.), and the amount (diversity) of data are sufficient, the trained model  $\widehat{\mathcal{R}}_\theta$  should reflect the true  $\mathcal{A}$  more precisely and comprehensively and is expected to produce better reconstruction results than conventional physics-model-driven or knowledge-driven approaches. The “data-driven” learning-based paradigm eliminates the need to design different processing flows for specific image-processing algorithm based on experience and pre-knowledge. By applying different types of training datasets, one specific class of neural network can be trained to perform various types of transformation for different tasks, significantly improving the universality and reducing the complexity of solving new problems.

**(2) From “divide-and-conquer” to “end-to-end learning”** In contrast to the traditional optical metrology approach that solves the sequence of tasks independently, deep learning allows for an “end-to-end” learning structure, where the neural network can learn the direct mapping relation between the raw image data and the desired sample parameters in one step, i.e.,  $\widehat{\mathbf{p}} = \widehat{\mathcal{R}}_\theta(\mathbf{I})$ , as illustrated in Fig. 10b. Compared with the “divide-and-conquer” scheme, the “end-to-end” learning allows to jointly solve multiple tasks, with great potential to alleviate the total computational burden. Such an approach has the advantage of synergy: it enables sharing information (features) between parts of the network that perform different tasks, which is more likely to get better

overall performance compared to solving each task independently.

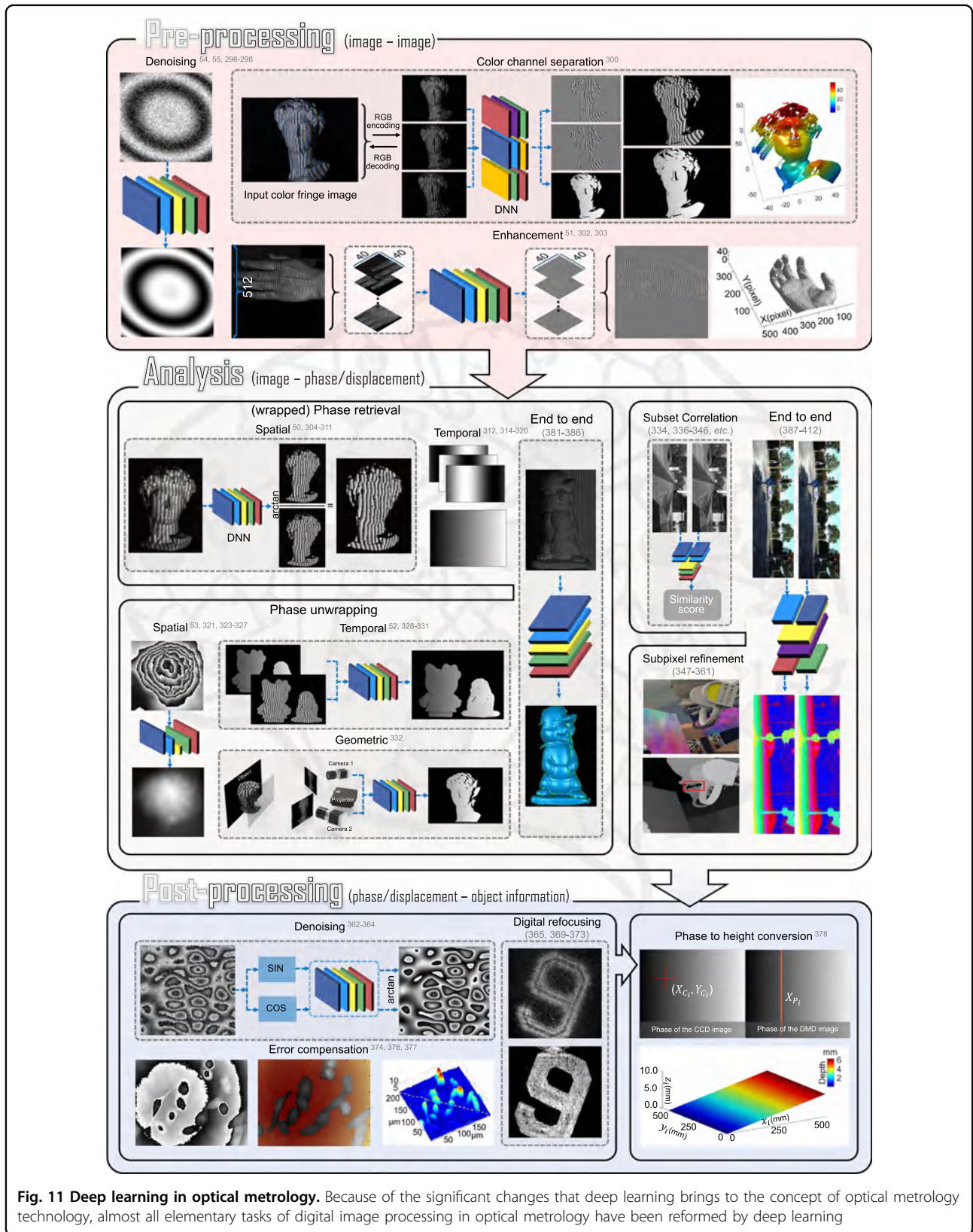
**(3) From “solving ill-posed inverse problems” to “learning pseudo-inverse mapping”** Deep learning utilizes complex neural network structures and nonlinear activation functions to extract high-dimensional features of the sample data, remove irrelevant information, and finally establish a nonlinear pseudo-inverse mapping model that is sufficient to describe the entire measurement process. The major reason for the success of deep learning is the abundance of training data and the explicit agnosticism from a priori knowledge of how such data are generated. Instead of hand-crafting a regularization function or specifying prior, deep learning can automatically learn it from the example data. Consequently, the learned prior  $R(\theta)$  is tailored to the statistics of real experimental data and, in principle, provides stronger and more reasonable regularization to the inverse problem pertaining to a specific metrology system. Consequently, the obstacle of “solving nonlinear ill-posed inverse problems” can be bypassed, and the pseudo-inverse mapping relation between the input and the desired output can be established directly.

### The use of deep learning in optical metrology

#### Deep-learning-enabled image processing in optical metrology

Owing to the above-mentioned advantages, deep learning has been gaining increasing attention in optical metrology, demonstrating promising performance in various optical metrology tasks and in many cases exceeding that of classic techniques. In this section, we review these existing researches leveraging deep learning in optical metrology according to an architecture similar to that introduced in the section “Image processing in optical metrology”, as summarized in Fig. 11. The basic network types, loss functions, and data acquisition methods of some representative examples are listed in Table 1.

- (1) **Pre-processing:** Many early works applying deep learning to optical metrology focused on image pre-processing tasks, such as denoising and enhancement. This is mainly due to the fact that the successful use cases of deep learning to such pre-processing tasks can be easily found in the computer vision community. Many image pre-processing algorithms in optical metrology could receive a performance upgrade by simply reengineering these existing neural network architectures for a similar kind of problem.
  - **Denoising:** Yan et al.<sup>55</sup> constructed a CNN composed of 20 convolutional layers for fringe denoising (Fig. 12a). Simulated fringe patterns with artificial Gaussian noise were generated as the

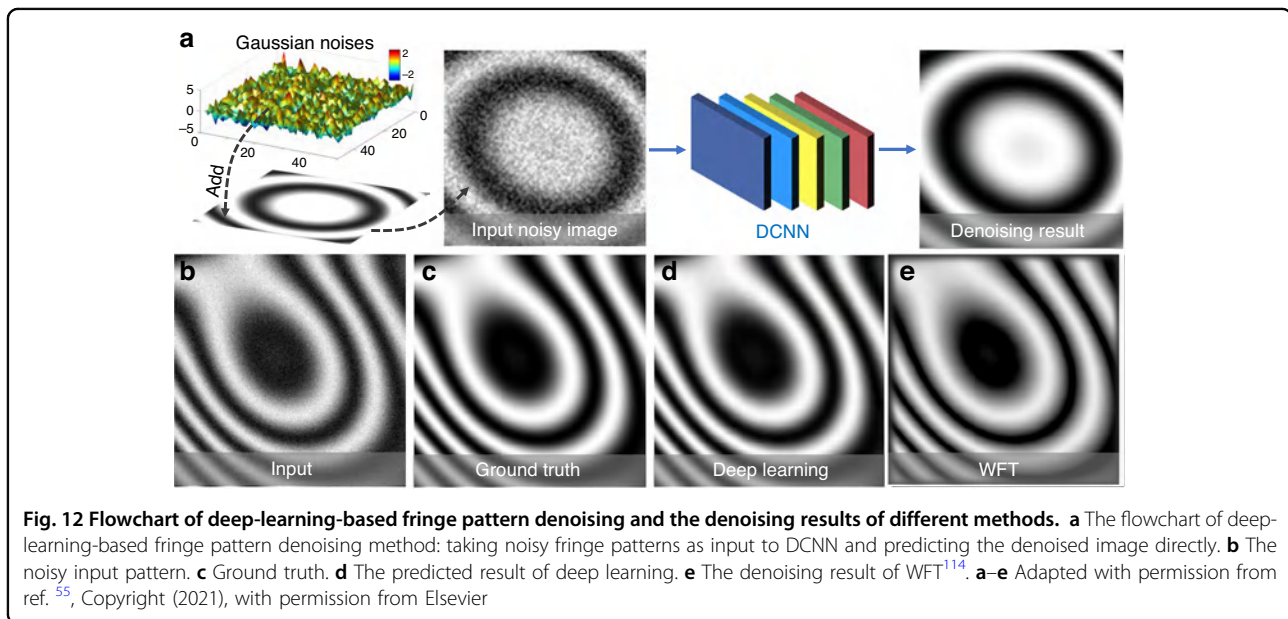


**Fig. 11 Deep learning in optical metrology.** Because of the significant changes that deep learning brings to the concept of optical metrology technology, almost all elementary tasks of digital image processing in optical metrology have been reformed by deep learning

**Table 1 Basic network structures, loss functions, and data acquisition methods for deep-learning methods applied to optical metrology tasks**

	Task	Reference	Network structure	Training database	Loss function
Pre-processing	Denoising	Yan et al. <sup>55</sup>	SRCNN	Simulation	MAE
		Jeon et al. <sup>296</sup>	U-Net + ResNet	Simulation	MAE
		Hao et al. <sup>54</sup>	SRCNN + ResNet	Simulation	Euclidean loss
		Lin et al. <sup>297</sup>	SRCNN + ResNet	Simulation	Euclidean loss
	Color channel separation	Qian et al. <sup>300</sup>	FCN + ResNet	Experiment	MSE
	Enhancement	Shi et al. <sup>51</sup>	SRCNN	Experiment	MSE
		Yu et al. <sup>303</sup>	FCN	Experiment	MSE
Analysis	Phase demodulation	Feng et al. <sup>50,304</sup>	FCN + ResNet	Experiment	MSE
		Yang et al. <sup>307</sup>	GAN	Experiment	GAN loss
		Li et al. <sup>311</sup>	U-Net	Experiment	MSE
		Zhang et al. <sup>315</sup>	GAN	Simulation	GAN loss
	Phase unwrapping	Wang et al. <sup>321</sup>	U-Net	Simulation	SSIM
		Spoorthi et al. <sup>323</sup>	FCN	Simulation	MSE
		Zhang et al. <sup>325</sup>	FCN + U-Net	Simulation	Cross-entropy
		Kando et al. <sup>326</sup>	U-Net	Simulation	RMSE
		Yin et al. <sup>52</sup>	FCN + ResNet	Experiment	MSE
	Subset correlation	Žbontar and LeCun <sup>334</sup>	CNN	KITTI <sup>459</sup> , Middlebury <sup>460</sup>	Hinge loss
		Luo et al. <sup>336</sup>	CNN	KITTI <sup>459</sup>	Cross-entropy
		Guo et al. <sup>344</sup>	FCN	Scene Flow <sup>388</sup> , KITTI <sup>459</sup>	Smooth L1
	Subpixel refinement	Pang et al. <sup>347</sup>	FCN	FlyingThings3D <sup>387</sup> , Middlebury <sup>460</sup> , KITTI <sup>459</sup>	MAE
		Hartmann et al. <sup>338</sup>	CNN + ResNet	Scene Flow <sup>388</sup> , KITTI <sup>459</sup>	Smooth L1
	Denoising	Montresor et al. <sup>362</sup>	SRCNN + ResNet	Simulation	MSE
		Yan et al. <sup>363</sup>	SRCNN + ResNet	Simulation	MSE
	Digital refocusing	Ren et al. <sup>365</sup>	SRCNN + ResNet	Experiment	MSE
		Wang et al. <sup>309</sup>	U-Net	Experiment	MSE
		Lee et al. <sup>370</sup>	CNN	Simulation	MSE
Shinmobaba et al. <sup>371</sup>		CNN	Experiment	MSE	
Nguyen et al. <sup>374</sup>		U-Net	Experiment	Cross-entropy	
Error compensation	Aguénounon et al. <sup>377</sup>	U-Net	Experiment	Mse	
	Li et al. <sup>378</sup>	BP neural network	Experiment	–	
Postprocessing	Phase to height conversion	Nguyen et al. <sup>381</sup>	FCN, U-Net	Experiment	MSE
		Van et al. <sup>382</sup>	SRCNN	Simulation	RMSE
		Machineni et al. <sup>384</sup>	FCN + ResNet	Simulation	Smooth L1
		Zheng et al. <sup>385</sup>	U-Net	Simulation	RMSE
	From stereo images to disparity	Kendall et al. <sup>389</sup>	SRCNN + ResNet	Scene Flow, KITTI	MSE
		Chang et al. <sup>390</sup>	FCN + ResNet	Scene Flow, KITTI	Smooth L1



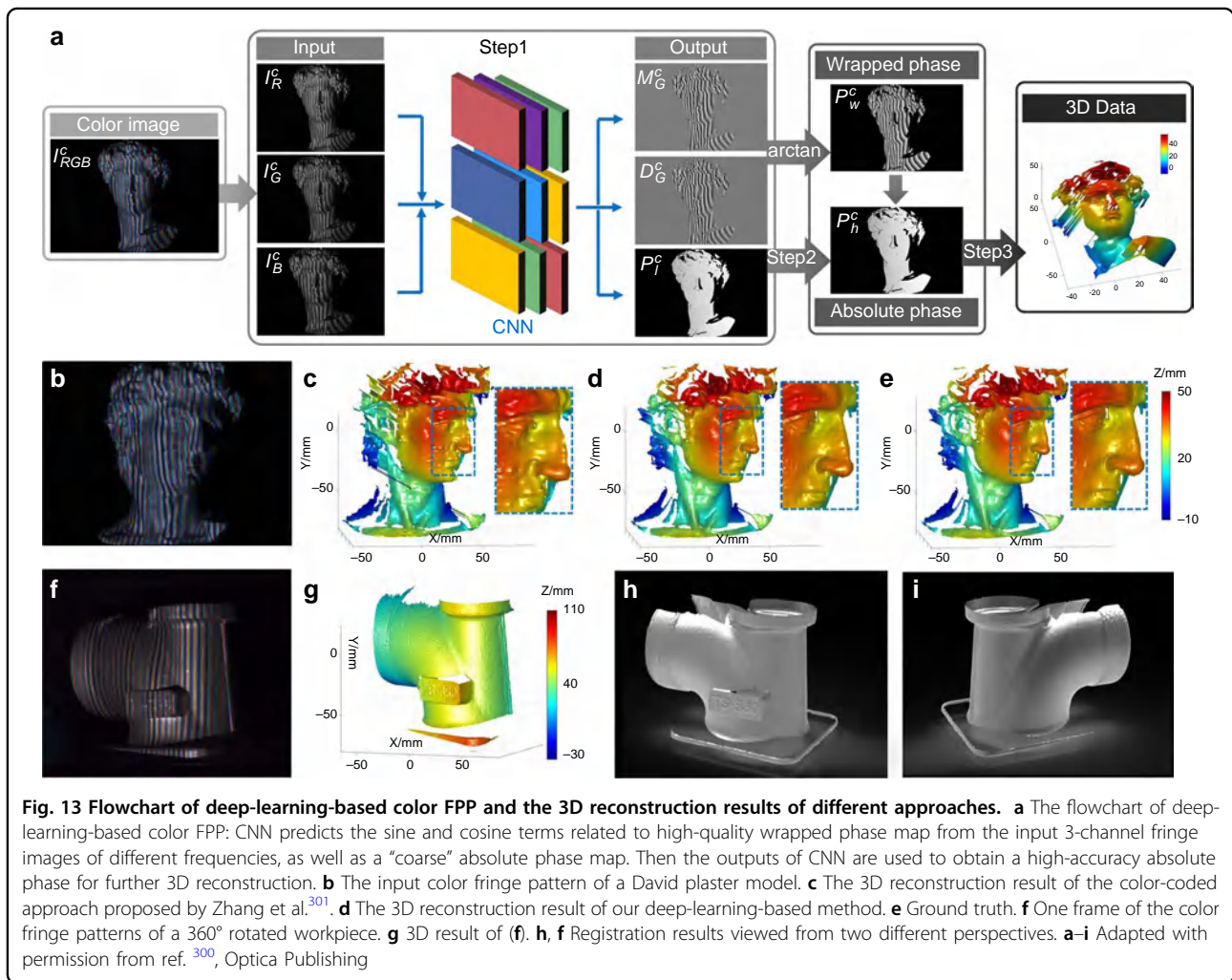


training dataset, and corresponding noise-free versions were used as ground truth. Figure 12d, e shows the denoising results of WFT<sup>114</sup> and the deep-learning-based method, showing that their method was free of the boundary artifacts in WFT and achieved comparable denoising performance in the central region. Jeon et al.<sup>296</sup> proposed a fast speckle-noise reduction method based on U-Net, which showed robust and excellent denoising performance for digital holographic images. Hao et al.<sup>54</sup> constructed a fast and flexible denoising convolutional neural network (FFDNet) for batch denoising of ESPI fringe images. Lin et al.<sup>297</sup> developed a denoising CNN (DnCNN) for speckle-noise suppression of fringe patterns. Reyes-Figueroa and Rivera<sup>298</sup> proposed a fringe pattern filtering and normalization technique based on autoencoder<sup>299</sup>. The autoencoder was able to fine-tune the U-Net network parameters and reduce residual errors, thereby improving the stability and repeatability of the neural network. Since it is difficult to access noise-free ground-truth images in real experimental conditions, the training datasets of these deep-learning-based denoising methods are all generated based on simulations.

- **Color channel separation:** Our group reported a single-shot 3D shape measurement approach with deep-learning-based color fringe projection profilometry that can automatically eliminate color cross-talk and channel imbalance<sup>300</sup>. As shown in Fig. 13a, the network predicted the sine and cosine terms related to high-quality cross-talk-free phase information from the input 3-channel

fringe images of different wavelengths. In order to get rid of color cross-talk and chromatic aberration, the green monochromatic fringe patterns were projected and only the green channel of the captured patterns was used to generate labels. Figure 13b–d shows 3D reconstruction results of a David plaster model measured by the traditional color-coded method<sup>301</sup> and our method, showing that the deep-learning-based method yielded more accurate surface details. The quality of the 3D reconstruction was comparable to the ground truth (Fig. 13e) obtained by the non-composite (monochromatic) multi-frequency phase-shifting method<sup>174</sup>. The deep-learning-based method was applied for dynamic 360° 3D digital modeling, demonstrating its potential in rapid reverse engineering and related industrial applications (Fig. 13f–i).

- **Enhancement:** Shi et al.<sup>51</sup> proposed a fringe-enhancement method based on deep learning, and the flowchart of which is given in Fig. 14a. The captured fringe image and the corresponding enhanced one obtained by the subtraction of two fringe patterns with  $\pi$  relative phase shift were used to establish the mapping between the raw fringe and the desired enhanced versions. Figure 14b–d shows the 3D reconstruction results of a moving hand using the traditional FT method<sup>138</sup> and the deep-learning method, suggesting that the deep-learning method outperformed FT in terms of detail preservation and SNR. Goy et al.<sup>302</sup> proved that DNN could recover an image with decent quality under low-photon conditions, and successfully



**Fig. 13** Flowchart of deep-learning-based color FPP and the 3D reconstruction results of different approaches. **a** The flowchart of deep-learning-based color FPP: CNN predicts the sine and cosine terms related to high-quality wrapped phase map from the input 3-channel fringe images of different frequencies, as well as a “coarse” absolute phase map. Then the outputs of CNN are used to obtain a high-accuracy absolute phase for further 3D reconstruction. **b** The input color fringe pattern of a David plaster model. **c** The 3D reconstruction result of the color-coded approach proposed by Zhang et al.<sup>301</sup>. **d** The 3D reconstruction result of our deep-learning-based method. **e** Ground truth. **f** One frame of the color fringe patterns of a 360° rotated workpiece. **g** 3D result of (f). **h, i** Registration results viewed from two different perspectives. **a–i** Adapted with permission from ref.<sup>300</sup>, Optica Publishing

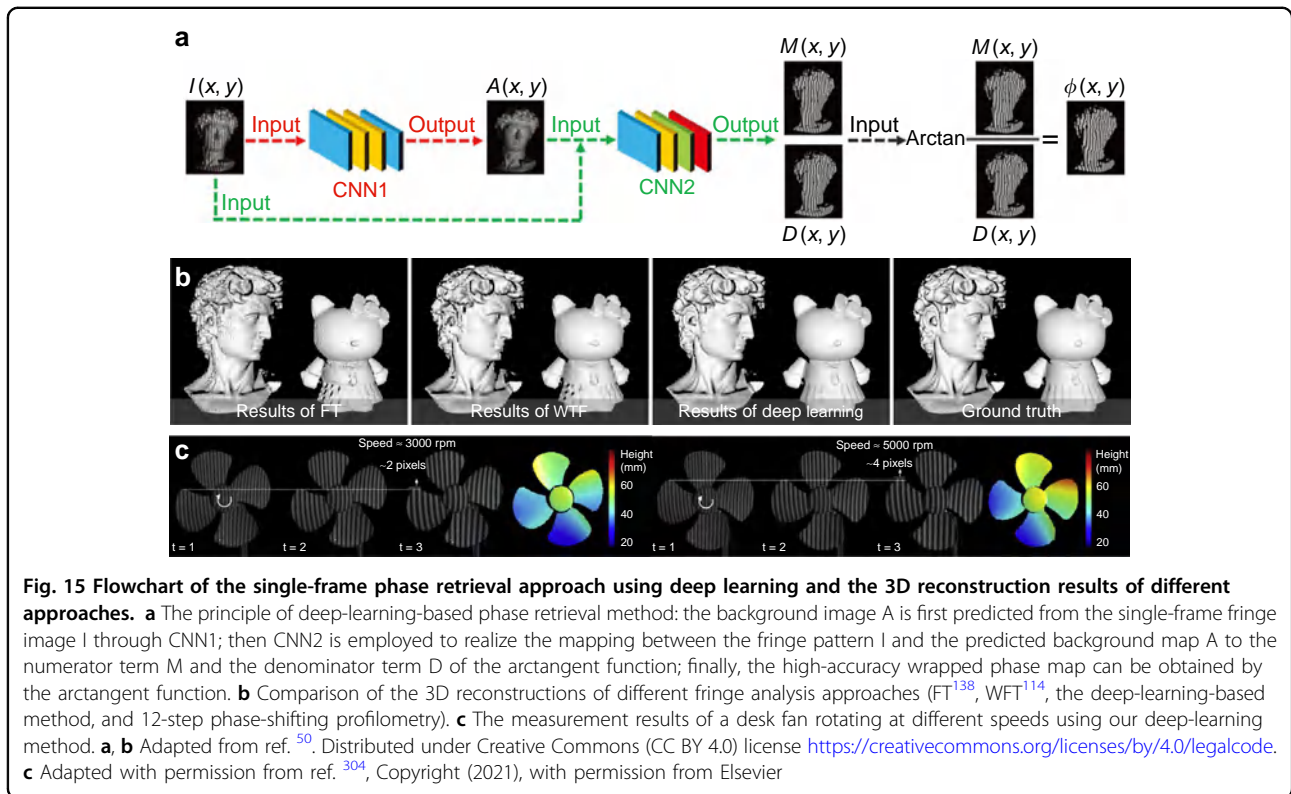
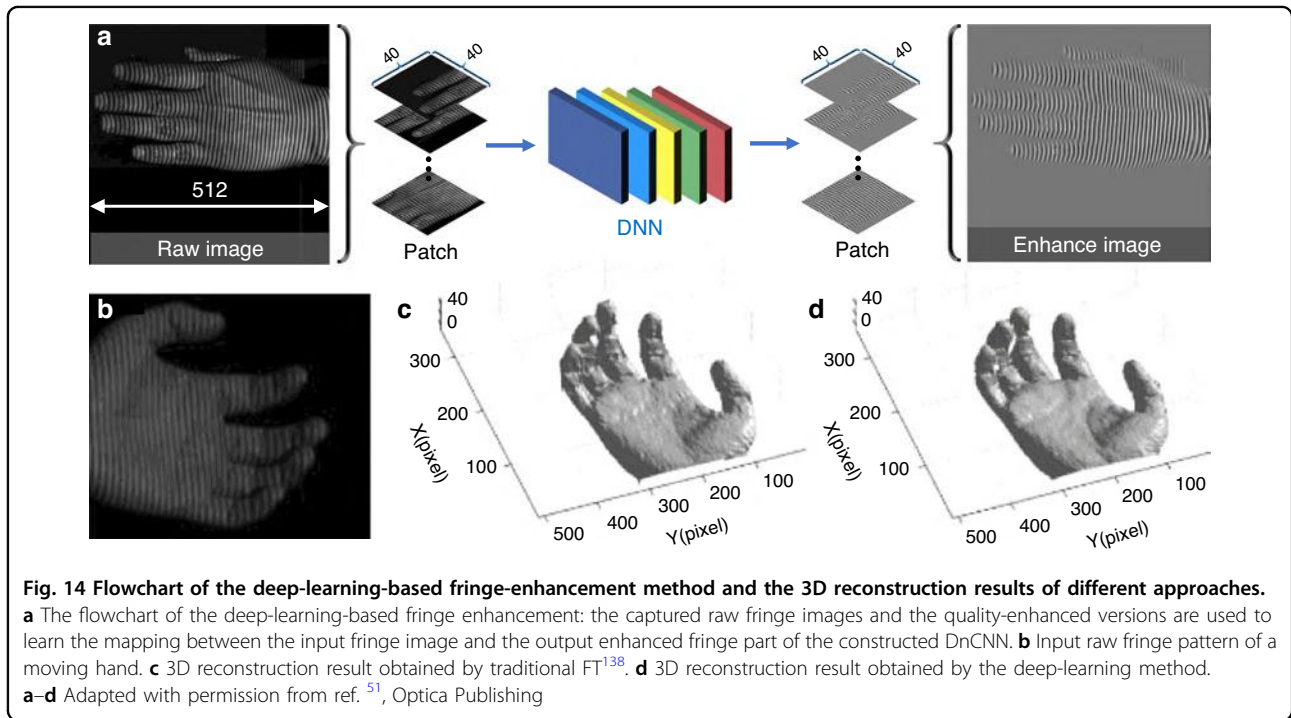
applied their method to phase retrieval. Yu et al.<sup>303</sup> proposed a fringe-enhancement method in which the fringe modulation was improved by deep learning, facilitating high-dynamic 3D shape measurement without resorting to conventional multi-exposure schemes.

- (2) **Analysis:** Image analysis is the most critical step in the image-processing architecture of optical metrology. Consequently, most deep-learning techniques applied to optical metrology are proposed to accomplish the tasks associated with image analysis. For phase measurement techniques, deep learning is extensively explored for (both spatial and temporal) phase demodulation and (spatial, temporal, and geometric) phase unwrapping.

• **Phase demodulation:**

**Spatial phase retrieval:** To address the contradiction between the measurement efficiency and accuracy of traditional phase retrieval methods, our group, for the first time, introduced deep learning to fringe pattern

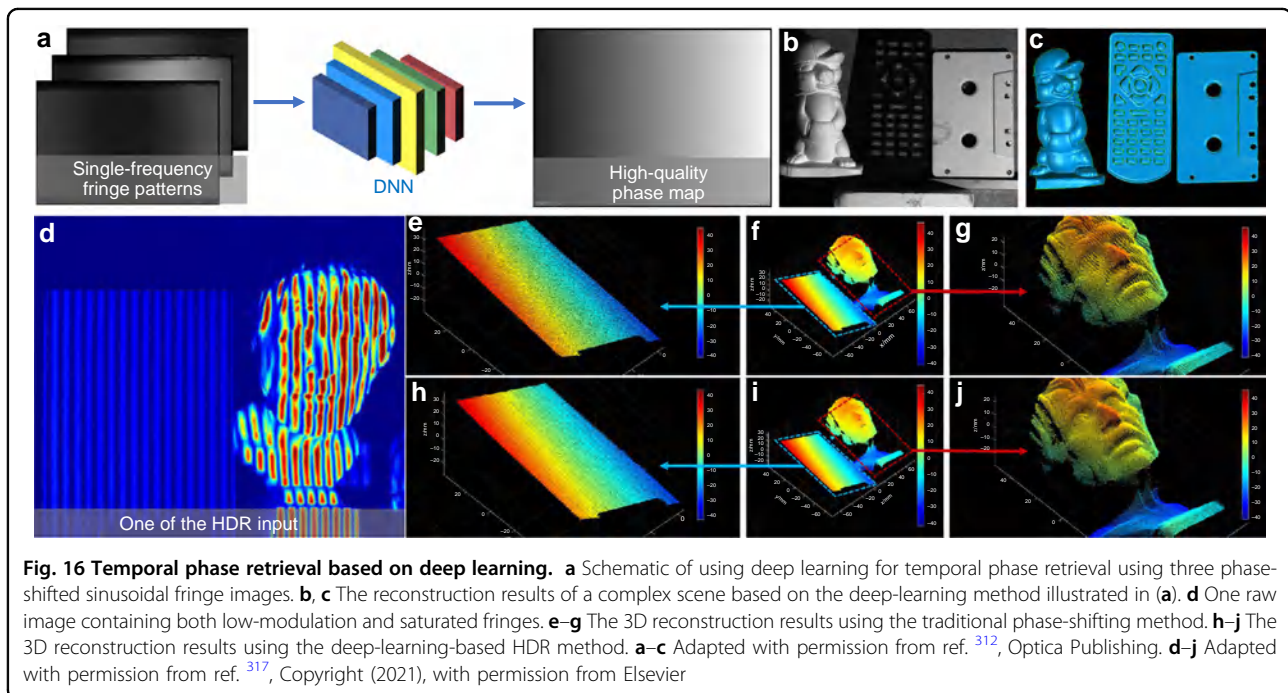
analysis, substantially enhancing the phase-demodulation accuracy from a single-fringe pattern<sup>50</sup>. As illustrated in Fig. 15a, the background image A was first predicted from the acquired fringe image I through CNN1. Then CNN2 was employed to realize the mapping from I and A to the numerator (sine) term M and denominator (cosine) term D. Finally, the wrapped phase information can be acquired by computing the arctangent of M/D. Figure 15b compares the phases retrieved by two representative traditional single-frame phase retrieval methods (FT<sup>138</sup>, WFT<sup>114</sup>) and the deep-learning method, revealing that our deep-learning-based single-frame phase retrieval method achieved the highest reconstruction quality, which almost visually reproduced the ground-truth information obtained by the 12-step phase-shifting method. We have incorporated the deep-learning-based phase retrieval technique into the micro-Fourier transform profilometry ( $\mu$ FTP) technique to eliminate the need for additional uniform patterns, doubling the measurement speed and achieving an



unprecedented 3D imaging frame rate up to 20,000 Hz<sup>304</sup>. Figure 15c shows the 3D measurement results of a rotating fan at different speeds (3000 and 5000 revolutions per minute (RPM)), suggesting that the

3D shape of fan blades can be intactly reconstructed without any motion-induced artifacts visible. Qiao et al.<sup>305</sup> applied this deep-learning-based phase extraction technique for phase measuring deflectometry, and





achieved single-shot high-accuracy 3D shape measurement of specular surfaces. Some other network structures, such as structured light CNN (SL-CNN)<sup>306</sup> and deep convolutional GAN<sup>307</sup> were also adopted for single-frame phase retrieval. In addition, deep learning can also be applied to Fourier transform profilometry for automatic spectrum extraction by identifying the carrier frequency components bearing the object information in the Fourier domain, facilitating automatic spectrum extraction, and achieving higher phase retrieval accuracy without human intervention<sup>308</sup>. Wang et al.<sup>309</sup> proposed an automatic holographic reconstruction framework (Y-Net) consisting of two symmetrical U-Nets, allowing for simultaneous recovery of phase and intensity information from a single off-axis digital hologram. They also doubled the capability of Y-Net, extending it to the reconstruction of dual-wavelength complex amplitudes, while overcoming the spectral overlapping issue in common-path dual-wavelength digital holography<sup>310</sup>. Recently, our group used U-Net to realize aliasing-free phase retrieval from a dual-frequency composite fringe pattern<sup>311</sup>. Compared with the traditional Fourier transform profilometry, the deep-learning-enabled approach avoids the complexities associated with dual-frequency spectra separation and extraction, allowing for higher-quality single-shot absolute 3D shape reconstruction.

**Temporal phase retrieval:** Wang et al.<sup>312</sup> introduced a deep-learning scheme to the phase-shifting technique in FPP. As shown in Fig. 16a, by introducing a fully connected DNN, the link between three low- and unit-frequency phase-shifting fringe patterns and high-quality

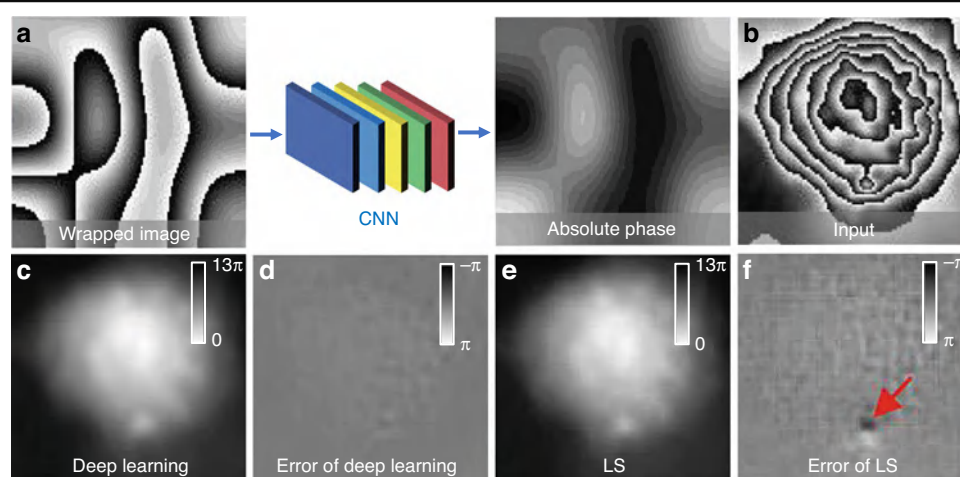
absolute phases calculated from high-frequency fringe images were established, and thus, the 3D measurement accuracy could be significantly enhanced. The three unit-frequency phase-shifting patterns were encoded in three monochrome channels of a color image and projected by a 3LCD projector. The individual fringe patterns were then decoded and projected by the projector sequentially and rapidly<sup>313,314</sup>. Consequently, the hardware system allowed for real-time 3D surface imaging of multiple objects at a speed of 25.6 fps. Zhang et al.<sup>315</sup> developed a deep-phase-shift network (DPS-Net) based on GAN, with which multi-step phase-shifting interferograms with accurate arbitrary phase shifts for calculating high-quality phase information were predicted from a single interferogram. Besides random intensity noise, conventional phase-shifting algorithms are also sensitive to other experimental imperfections, such as phase-shifting error, illumination fluctuations, intensity nonlinearity, lens defocusing, motion-induced artifacts, and detector saturation. Deep learning also provides a potential solution to eliminate or at least partially alleviate the impact of these error sources on phase measurement. For example, Li et al.<sup>316</sup> proposed a deep-learning-based phase-shifting interferometric phase recovery approach. The constructed U-Net was capable of predicting the accurate wrapped phase map from two interferogram inputs with unknown phase shifts. Zhang et al.<sup>317</sup> applied CNN to extract a high-accuracy wrapped phase map from conventional 3-step phase-shifting fringe patterns. In the training stage, low-modulation or saturated fringe patterns were used as the raw dataset,

and the relation between these imperfect raw fringe and high-quality error-free unwrapped phase (obtained by 12-step phase-shifting algorithms) were established based on CNN. Consequently, the deep-learning-based approach could accommodate both dark and reflective surfaces, and the related phase errors (noise and saturation) in the conventional three-step phase-shifting method were significantly suppressed, making it a promising approach for high-dynamic-range (HDR) 3D measurement of surfaces with large reflectivity variations (Fig. 16d–g). Wu et al.<sup>318</sup> proposed a deep-learning-based phase-shifting approach to overcome the phase errors associated with intensity nonlinearity. Through a well-trained FCN, the distortion-free high-quality phase map could be reconstructed conveniently and efficiently from the raw phase-shifting fringe patterns with a strong gamma effect. Yang et al.<sup>319</sup> constructed a three-to-three deep-learning framework (Tree-Net) based on U-Net to compensate for the nonlinear effect in the phase-shifting images, which effectively and robustly reduced the phase errors by about 90%. Recently, our group demonstrated that the nonsinusoidal errors (e.g., residual high-order harmonics in binary defocusing projection, intensity saturation, gamma effect of projectors and cameras, and their coupling) in phase-shifting profilometry could be handled by an integrated deep-learning framework. A well-trained U-Net could effectively suppress the phase errors caused by different types of nonsinusoidal fringe with only a minimum of three fringe patterns as input<sup>320</sup>.

- **Phase unwrapping:**

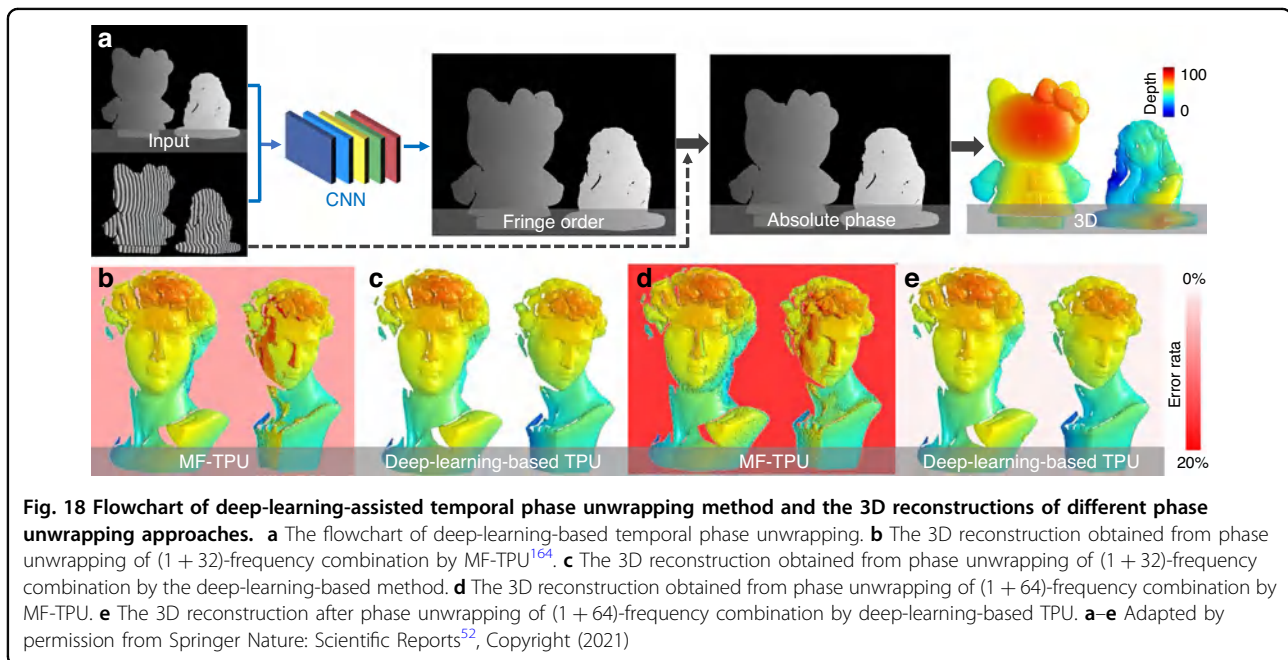
**Spatial phase unwrapping:** Wang et al.<sup>321</sup> proposed a one-step phase unwrapping approach based on deep

learning. Various ideal (noise-free) continuous phase distributions and the corresponding wrapped phase maps with different types of noises (Gaussian, salt and pepper, or multiplicative noises) were simulated and used as the training dataset for a CNN based on U-Net. Upon completion of the training, the absolute phases can be predicted directly from a noisy wrapped phase map, as illustrated in Fig. 17a. Figure 17b–f shows the comparisons of phase unwrapping results obtained by the traditional least-square (LS) method<sup>322</sup> and deep-learning-based method, demonstrating that deep learning can directly fulfill the complicated nonlinear phase unwrapping task in one step with improved anti-noise and anti-aliasing ability. Spoorthi et al.<sup>323</sup> developed a CNN-based phase unwrapping framework—PhaseNet. The fringe order ( $2\pi$  integer phase jumps) used for phase unwrapping can be obtained pixel by pixel through a semantic segmentation-based deep-learning framework of the encoder-decoder structure. Recently, they developed an enhanced phase unwrapping framework—PhaseNet 2.0, which could directly map a noisy wrapped phase to a denoised absolute one<sup>324</sup>. Zhang et al.<sup>325</sup> transferred the task of phase unwrapping to a multi-class classification problem and generated fringe orders by feeding the wrapped phase into a convolutional segmentation network. Zhang et al.<sup>53</sup> proposed a deep-learning-based approach for rapid 2D phase unwrapping, which demonstrated good denoising and unwrapping performance and outperformed the conventional path-dependent and path-independent methods. Kando et al.<sup>326</sup> applied U-Net to achieve absolute phase prediction from a single interferogram, and the quality



**Fig. 17** Flowchart of the one-step deep-learning-based phase unwrapping approach and the unwrapping results of different methods.

**a** The flowchart of the one-step deep-learning-based phase unwrapping approach: the absolute phase can be predicted directly from a noisy wrapped phase based on the trained CNN. **b** The wrapped phase map of living mouse osteoblast. **c** Unwrapped phase of **(b)** obtained by deep learning. **d** Phase errors of **(c)**. **e** Unwrapped phase of **(b)** obtained by the conventional LS method<sup>322</sup>. **f** Phase errors of **(e)**. **a–f** Adapted with permission from ref. 321, Optica publishing



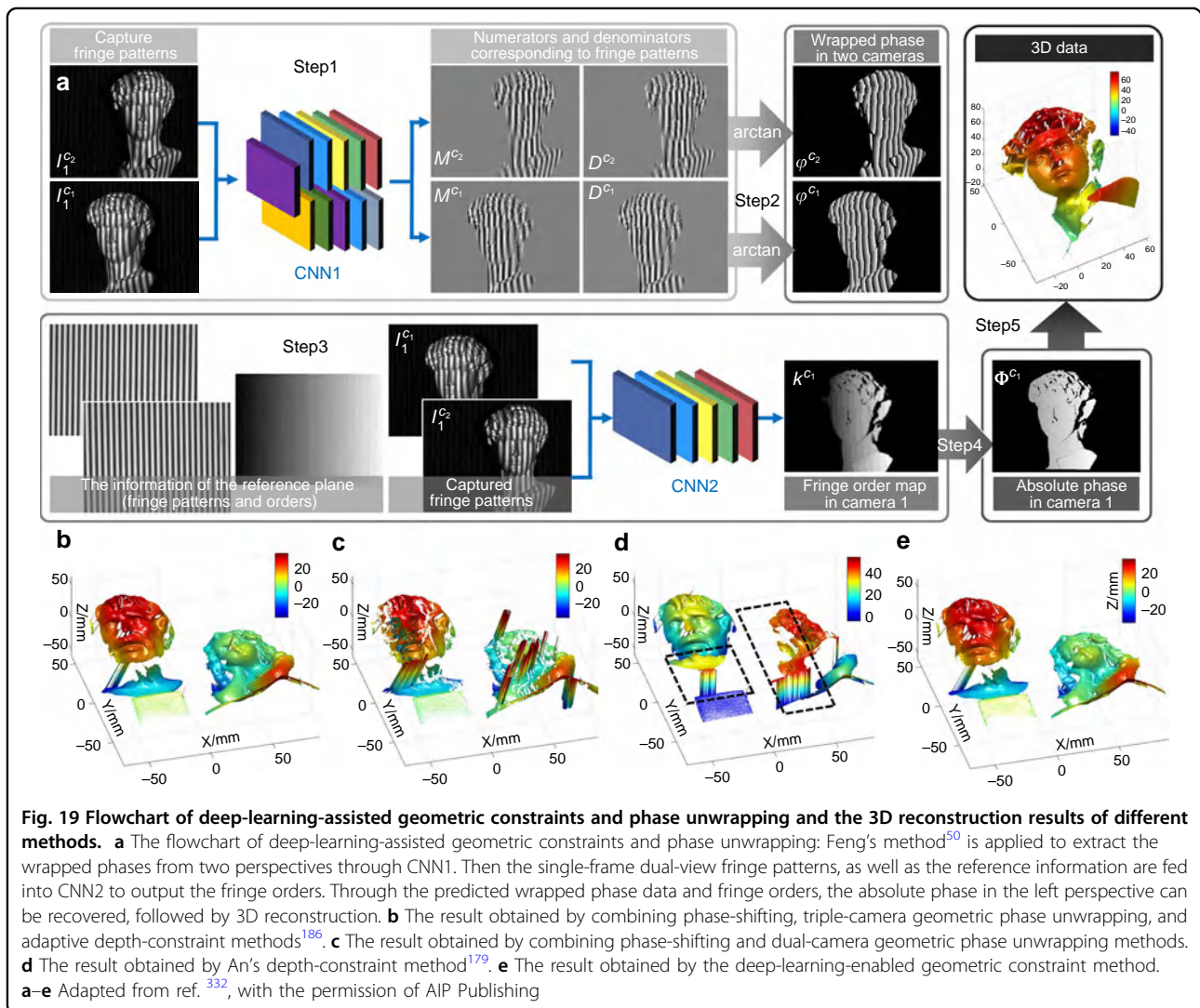
of the recovered phase was superior to that obtained by the conventional FT method, especially for closed-fringe patterns. Li et al.<sup>327</sup> proposed a deep-learning-based phase unwrapping strategy for closed fringe patterns. They compared four different network structures for phase unwrapping and found that the improved FCN architecture performed the best in terms of accuracy and speed. However, it should be mentioned that, similar to the case of fringe denoising, true absolute phase maps corresponding to the real experimentally obtained wrapped phase maps are generally quite hard to obtain in many interferometric techniques (which requires sophisticated multi-wavelength illuminations and heterodyne operations). Therefore, the training datasets used in the above-mentioned deep-learning-based spatial phase unwrapping methods are generated based on numerical simulation instead of real experiments. Moreover, since only one single wrapped phase map is used as input, the above-mentioned deep-learning-based spatial phase unwrapping methods still suffers from the  $2\pi$  ambiguity problem inherent in traditional phase measurement techniques.

**Temporal phase unwrapping:** Our group developed a deep-learning-based temporal phase unwrapping framework, as illustrated in Fig. 18a<sup>52</sup>. The inputs of the network are a single-period (wrap-free) phase map and a high-frequency wrapped phase map, from which the constructed CNN could directly predict the fringe orders corresponding to the high-frequency phase to be unwrapped. Figure 18b–e gives the comparison between the traditional multi-frequency temporal phase unwrapping (MF-TPU) method<sup>174</sup> and the deep-learning-based

approach for the 3D reconstructions obtained by unwrapping the wrapped phase maps using the (1–32) and (1–64) frequency combination of fringe patterns, respectively. In comparison with MF-TPU, the deep-learning-assisted method produced phase unwrapping results with higher accuracy and robustness even in the case of different types of error sources (low SNR, intensity nonlinearity, and object motion). Liu et al.<sup>328</sup> further improved this approach by using a lightweight classification CNN to extract the fringe orders from a pair of low-high-frequency phase maps, which saved a large amount of training time and made it possible to deploy the network on mobile devices. Li et al.<sup>329</sup> proposed a deep-learning-based dual-wavelength phase unwrapping approach in which only a single-wavelength interferogram was used to predict another interferogram recorded at a different wavelength with a conditional GAN (CGAN). Though their approach still suffered from the phase ambiguity problem when measuring discontinuous surface or isolated objects, it provided an effective and potential solution to phase unwrapping and extended the measurement range of single-wavelength interferometry and holography techniques. Yao et al. designed FCNs by incorporating residual layers to predict the fringe orders of wrapped phases from only two<sup>330</sup> or even single<sup>331</sup> Gray-code image(s), significantly reducing the required images compared with the conventional Gray-code technique.

**Geometric phase unwrapping:** Our group proposed a deep-learning-assisted geometric phase unwrapping approach for single-shot 3D surface measurement<sup>332</sup>. The flowchart of this approach is shown in Fig. 19a.

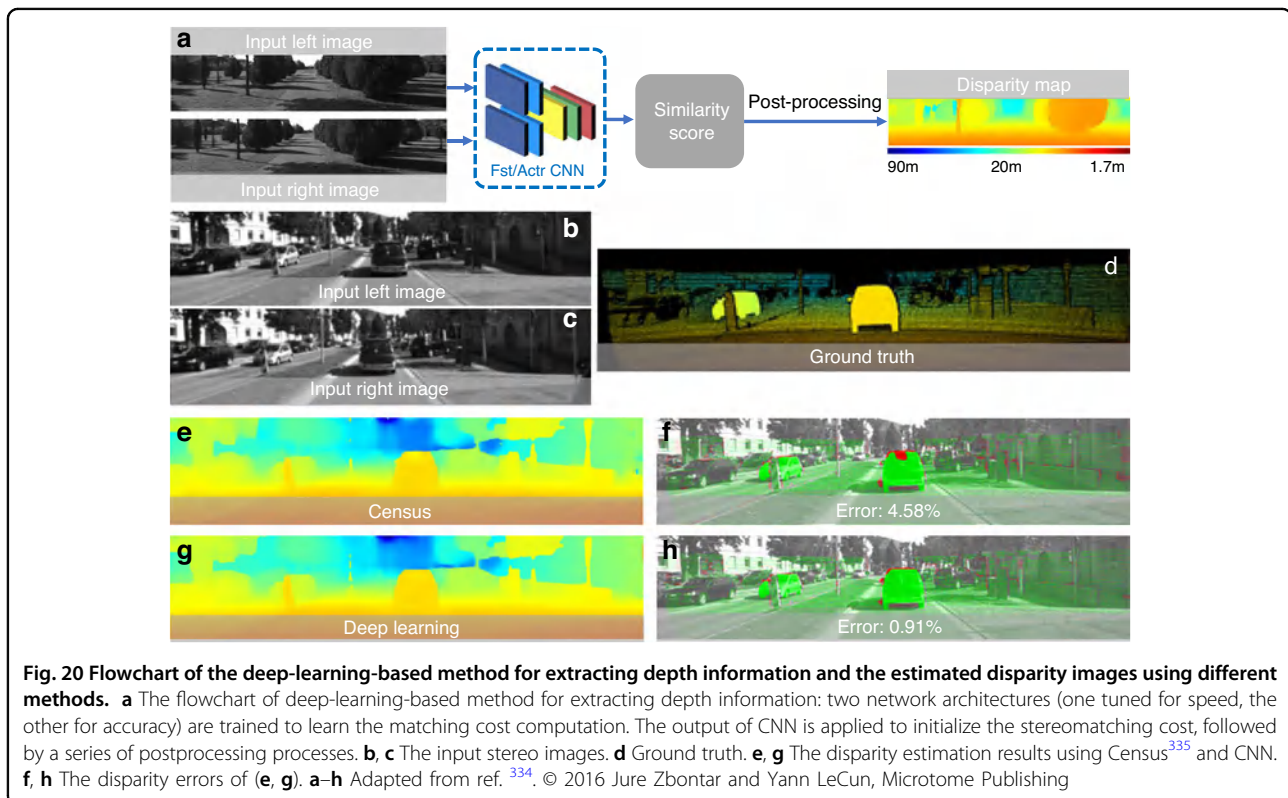




Two CNNs (CNN1 and CNN2) were constructed for phase retrieval and phase unwrapping, respectively. Based on a stereo camera system, dual-view single-shot fringe patterns, as well as the reference plane images, were fed into CNN2 to determine the fringe orders. With the predicted wrapped phases and fringe orders, the absolute phase map can be recovered. Figure 19b–e is the comparison of 3D reconstructions obtained through different conventional geometric phase unwrapping methods<sup>175,179,186</sup> and the deep-learning-based method, demonstrating that the deep-learning-based method can robustly unwrap the wrapped phases of dense fringe patterns within a larger measurement volume under the premise of single-frame projection. It should be mentioned that it is indeed a straightforward idea to establish the relationship between the fringe pattern to the corresponding absolute phase directly. However, since the rationality of the deep-learning-based approach is largely dependent on the input data,

when the input fringe itself is ambiguous, the network can never always produce reliable phase unwrapped results. For example, in Yu's work<sup>333</sup>, when there exist large depth discontinuities and isolated objects, even with the assistance of deep learning, one fringe image is insufficient to eliminate the  $2\pi$  phase ambiguity. In DIC and stereophotogrammetry, image analysis aims to determine the displacement vector of each pixel point between a pair of acquired images. Recently, deep learning has also been extensively applied to stereomatching in order to achieve improved performance compared with traditional subset correlation and subpixel refinement methods.

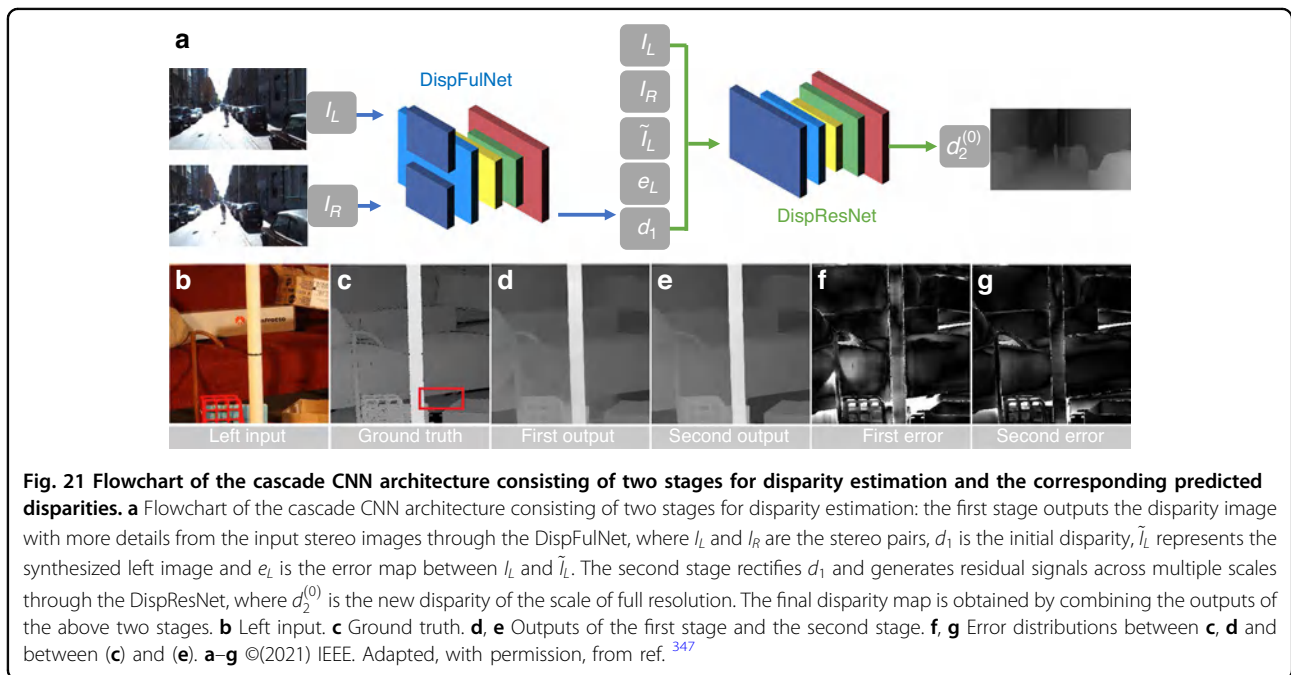
**Subset Correlation:** Zbontar and LeCun<sup>334</sup> presented a deep-learning-based approach for estimating the disparity map from a rectified stereo image pair. A siamese-structured CNN was reconstructed to address the matching cost computation problem through learning the similarity measure from small image patches.



The output of CNN was utilized for initializing the stereomatching cost, followed by some postprocessing processes, as shown in Fig. 20a. Figure 20d–h is the disparity images obtained from the traditional Census transform method<sup>335</sup> and the deep-learning-based method, from which we can see that the deep-learning-based approach achieved a lower error rate and better prediction result. Luo et al.<sup>336</sup> exploited a siamese CNN connected by point product layer to speed up the calculation of matching score and obtained improved matching performance. Recently, our group improved Luo’s network by introducing additional residual blocks and convolutional layers to the head of the neural network and replacing the original inner product with the fully connected layers with shared weights<sup>337</sup>. The improved network can extract a more accurate initial absolute disparity map from speckle image blocks after epipolar correction, and showed better matching capability than Luo’s network. Hartmann et al.<sup>338</sup> constructed a CNN with five siamese branches to learn a matching function, which could directly predict a scalar similarity score from multiple image patches. It should be noted that the siamese CNN is one of the most widely used network structures in stereovision applications, which has been frequently employed and continuously improved for subset correlation tasks<sup>339–343</sup>. On a different note, Guo et al.<sup>344</sup>

improved the 3D-stacked hourglass network to obtain the cost volume by group-wise correlation and then realized stereomatching. Besides conventional supervised learning approaches, unsupervised learning was also introduced to subset correlation. Zhou et al.<sup>345</sup> proposed an unsupervised deep-learning framework for learning the stereomatching cost, using a left-right consistency check to guide the training process to converge to a stable state. Kim et al.<sup>346</sup> constructed a semi-supervised network to estimate stereo confidence. First, the matching probability was calculated according to the matching cost with residual networks. Then, the confidence measure was estimated based on a unified deep network. Finally, the confidence feature of the disparity map is extracted by synthesizing the results obtained by the two networks.

**Subpixel refinement:** Pang et al.<sup>347</sup> proposed a cascade (two-stage) CNN architecture for subpixel stereomatching. Figure 21a shows the flowchart of their method. In the first stage, the disparity image with more details was obtained from the input stereo images through DispFulNet (“Ful” means full resolution) equipped with extra upsampling modules. Then the initialized disparity was rectified and the residual signals across multiple scales were generated through the hourglass structure DispResNet (“Res” means residual) in the second stage. According to the combination of the outputs from the



two stages, the final disparity with subpixel accuracy can be obtained. Figure 21d–g shows the predicted disparity images and error distributions of the input stereo image pairs (Fig. 21b) obtained by DispFulNet and DispResNet. It can be seen from the experimental results that after the second stage of optimization, the quality of the disparity was significantly improved. Based on different considerations, a large variety of network structures were proposed for subpixel refinement, e.g., StereoNet<sup>348</sup>, LGC-Net<sup>349</sup>, DeepMVS<sup>350,351</sup>, StereoDRNet<sup>352</sup>, DeepPruner<sup>353</sup>, LAF-Net<sup>354</sup>, 3D CNN<sup>355</sup>, MADNet<sup>356</sup>, Unos<sup>357</sup>, left-right comparative recurrent model<sup>358</sup>, CNN-based disparity map optimization<sup>359</sup>, deep-learning-based fringe-image-assisted stereomatching method<sup>360</sup>, and UltraStereo<sup>361</sup>.

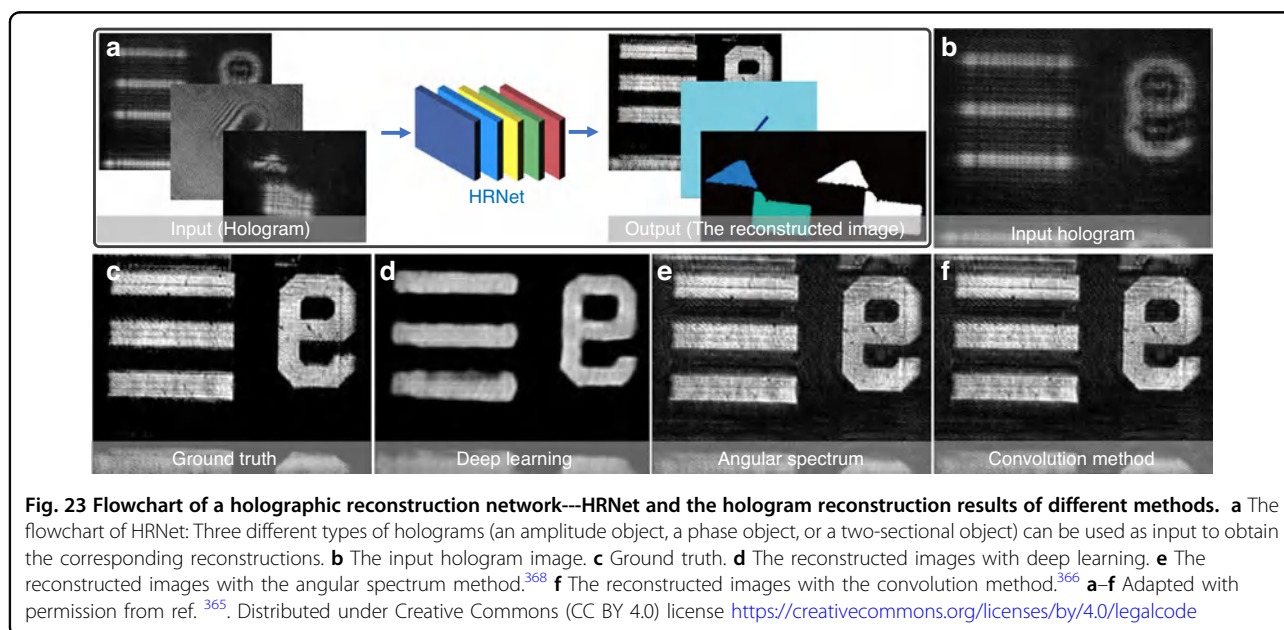
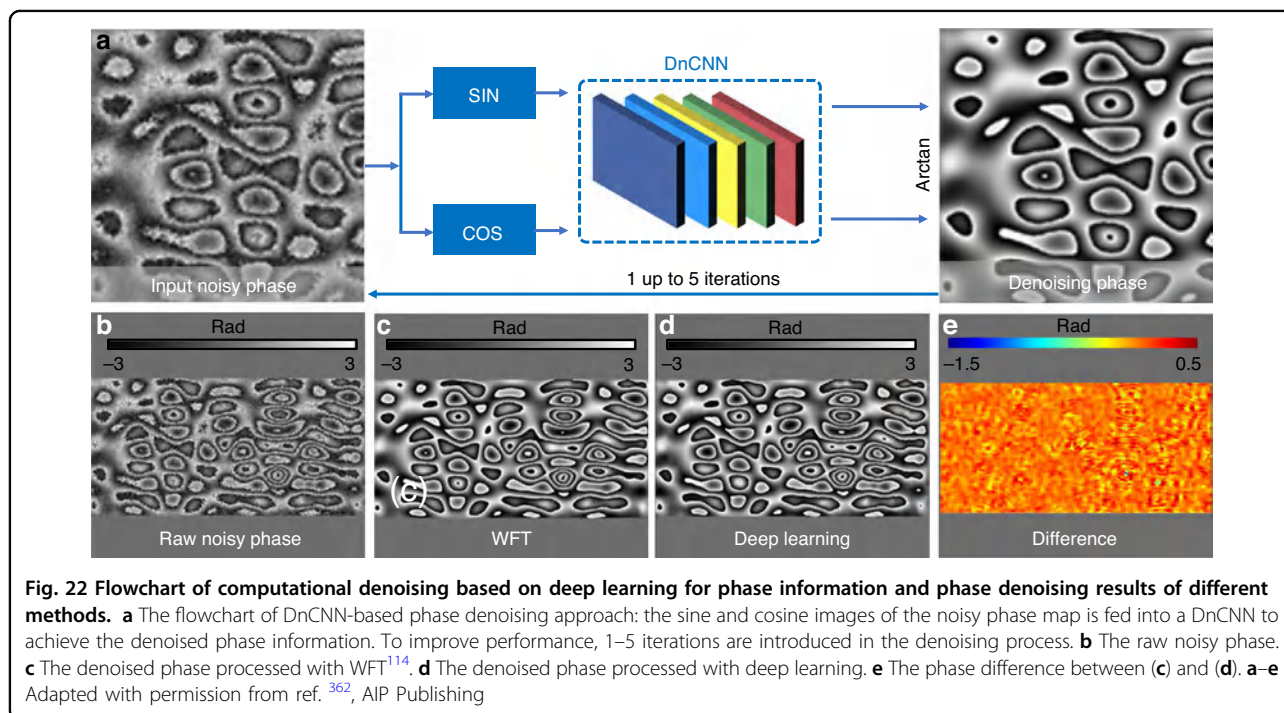
(3) **Postprocessing:** Deep-learning techniques also play an important role in the final postprocessing stage of the image-processing architecture of optical metrology. Examples of applying deep learning for postprocessing are very diverse, including further optimization of the measurement results (e.g., phase denoising, error compensation, and refocusing) and converting the measured intermediate variable to the desired physical quantity (e.g., system calibration and phase-to-height mapping in FPP).

- **Denoising:** Montrésor et al.<sup>362</sup> proposed to use DnCNN for phase denoising. As illustrated in Fig. 22a, the sine and cosine components of the noisy phase map were fed into a DnCNN to produce the corresponding denoised version, and the resultant phase information was calculated by the arctangent

function. The phase was then fed back into and refined by DnCNN again, and this process was repeated several times to achieve a better denoising performance. In order to generate more realistic training datasets via simulation, the additive amplitude-dependent speckle noise was carefully modeled by taking its non-Gaussian statistics, non-stationary properties, and a correlation length into account. Figure 22b–e shows the comparison of the denoising results obtained by WFT<sup>114</sup> and the deep-learning methods, suggesting that DnCNN yielded comparable standard deviation but lower peak-to-valley phase error than WFT. Yan et al.<sup>363</sup> proposed a CNN-based wrapped phase denoising method. By filtering the original numerator and denominator of the arctangent function, phase denoising results can be achieved without tuning any parameters. They also presented a deep-learning-based phase denoising technique for digital holographic speckle pattern interferometry<sup>364</sup>. Their approach could obtain an enhanced wrapped phase map by significantly suppressing the speckle noise, and outperformed traditional phase denoising methods when processing phases with steep spatial gradients.

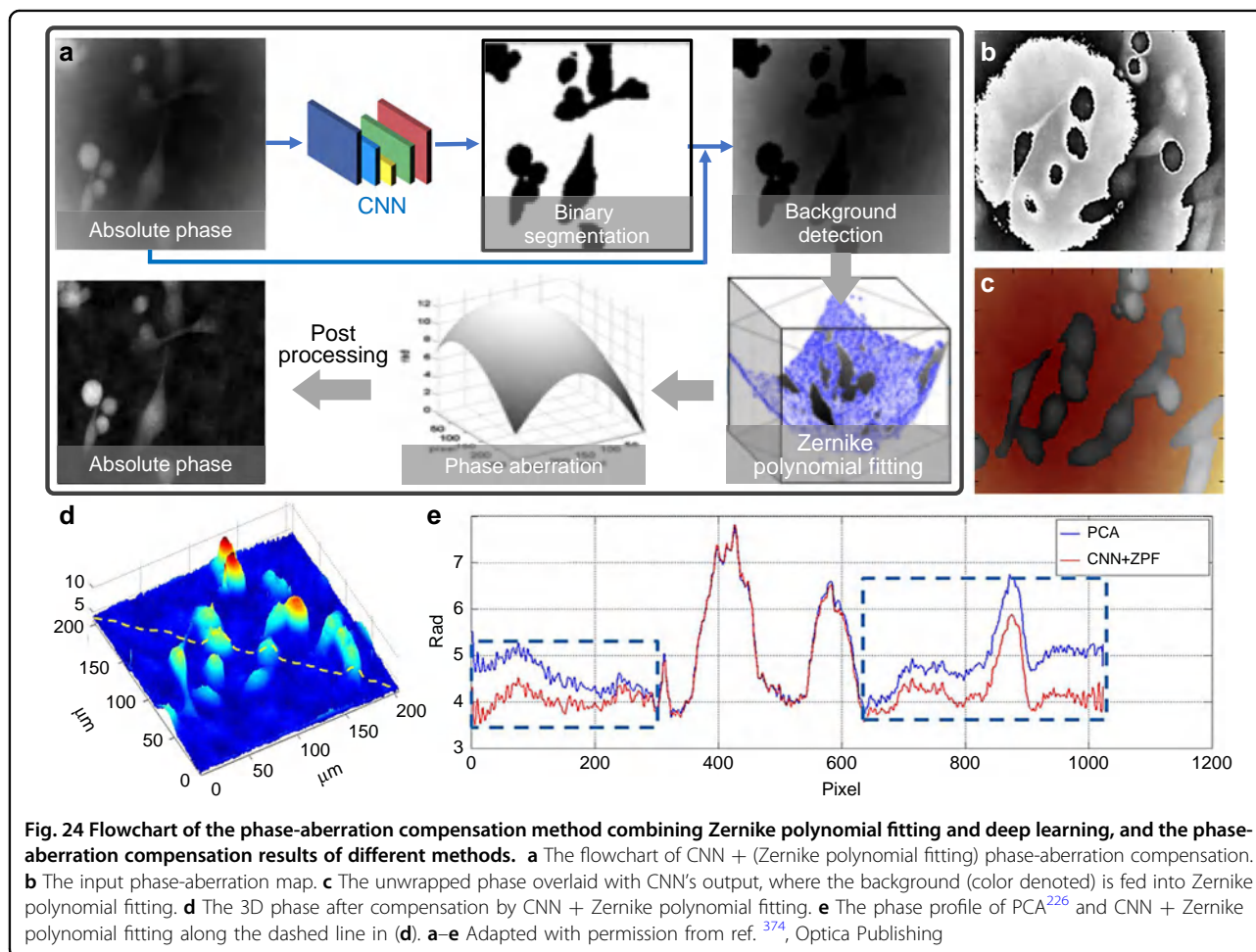
- **Digital refocusing:** Ren et al.<sup>365</sup> proposed the holographic reconstruction network (HRNet) to deal with the holographic reconstruction problem, which could perform automatic digital refocusing without employing any prior knowledge. Figure 23a gives the schematic of their deep-learning workflow, where a hologram input (the first block) was fed into HRNet, and then the reconstructed image (the third





block) corresponding to the specific input was directly predicted. A typical lens-free Mach-Zehnder interferometer was constructed to acquire training input images, and traditional convolution method<sup>366</sup>, PCA aberration compensation<sup>226</sup>, manual artifacts removal, and phase unwrapping<sup>367</sup> were successively employed to obtain the corresponding label images. Figure 23b–f shows the results of refocusing and hologram

reconstruction with different methods, proving that the predicted images by HRNet were precisely in-focus and noise-free, whereas there are significant noises and artifacts in the reconstruction results obtained by traditional convolution and angular spectrum method<sup>368</sup>. Alternatively, the autofocusing problem in DH could be recast as a regression problem, with the focal distance being a continuous response corresponding to a digital

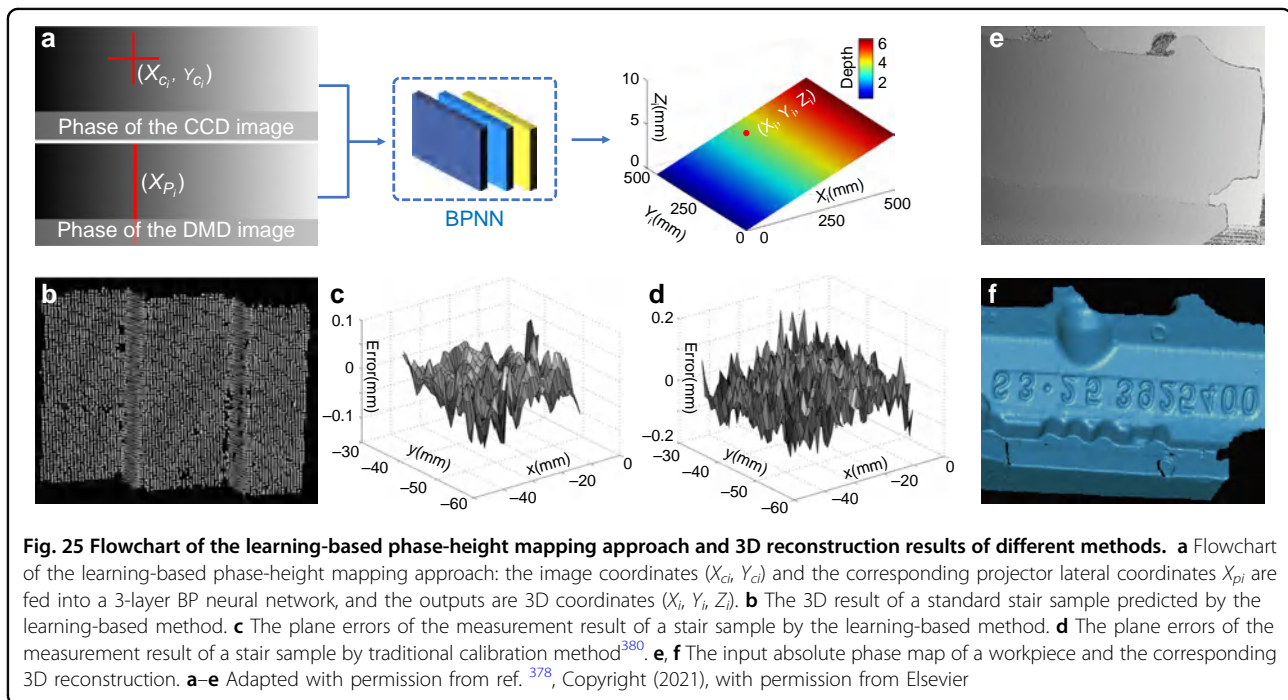


**Fig. 24** Flowchart of the phase-aberration compensation method combining Zernike polynomial fitting and deep learning, and the phase-aberration compensation results of different methods. **a** The flowchart of CNN + (Zernike polynomial fitting) phase-aberration compensation. **b** The input phase-aberration map. **c** The unwrapped phase overlaid with CNN's output, where the background (color denoted) is fed into Zernike polynomial fitting. **d** The 3D phase after compensation by CNN + Zernike polynomial fitting. **e** The phase profile of PCA<sup>226</sup> and CNN + Zernike polynomial fitting along the dashed line in (d). **a–e** Adapted with permission from ref. <sup>374</sup>, Optica Publishing

hologram. Ren et al.<sup>369</sup> constructed a CNN to achieve nonparametric autofocusing for digital holography, which could accurately predict the focal distance without knowing the physical parameters of the optical imaging system. Lee et al.<sup>370</sup> constructed a CNN-based estimator combined with Discrete Fourier Transform (DFT) to realize the automatic focusing of off-axis digital holography. Their method can automatically determine the object-to-image distance rapidly and effectively, and a sharp in-focus image of the object can be reconstructed accurately. Shimobaba et al.<sup>371</sup> used the regression-based CNN for holographic reconstruction, which could directly predict the sample depth position with millimeter accuracy from the power spectrum of the hologram. Jaferzadeh et al.<sup>372</sup> proposed a regression-layer-topped CNN to determine the optimal focus position for numerical reconstruction of micro-sized objects, which can be extended to the study of biological samples such as cancer cells. Pitkäaho et al.<sup>373</sup> constructed a CNN based on AlexNet and

VGG16 to learn the defocus distances from a large number of holograms. The well-trained network can determine the high-accuracy in-focus position of a new hologram without resorting to conventional numerical propagation algorithms.

- **Error compensation:** Nguyen et al.<sup>374</sup> proposed a phase-aberration compensation framework combining CNN and Zernike polynomial fitting, as illustrated in Fig. 24a. The unwrapped phase aberration map of the hologram was fed into a CNN with the U-Net structure to detect the background regions, which were then sent into the Zernike polynomial fitting<sup>375</sup> to determine the conjugated phase aberration. For training data collection/preparation, the PCA method<sup>226</sup> was used for training data collection/preparation. Figure 24b–e gives the phase aberration compensation results of PCA and the deep-learning method, showing that the phase aberrations were completely eliminated by using the deep-learning technique, while they were still visible in the phase results obtained by the PCA



method. In addition, the deep-learning-based technique was fully automatic, and the robustness and accuracy were shown to be superior to PCA. Lv et al.<sup>376</sup> used DNN to compensate projector distortion-induced measurement errors in a FPP system. By learning the mapping between the 3D coordinates of the object and their corresponding distortion-induced error distribution, the distortion errors of the original test 3D data can be accurately predicted. Aguenounon et al.<sup>377</sup> leveraged a DNN with a double U-Net structure to provide the single snapshot of optical properties imaging with the additional function of real-time profile correction. The real-time visualization of the resulting profile-corrected optical property (absorption and reduced scattering) map has the potential to be deployed to guide surgeons.

- **Quantity transformation:** Li et al.<sup>378</sup> proposed an accurate phase-height mapping approach for fringe projection based on a “shallow” (3-layer) BP neural network. The flowchart of their method is shown in Fig. 25a, where the camera image coordinates  $(X_{ci}, Y_{ci})$  and their corresponding horizontal ones  $X_{pi}$  of the projector image were fed into the network to predict the desired 3D information  $(X_i, Y_i, Z_i)$ . To obtain the training data, a standard calibration board with circle marks fixed on a high-precision displacement stage was captured at different  $Z$ -direction positions. With the captured images, the marks’ centers coordinates  $(X_{ci}, Y_{ci})$  with subpixel

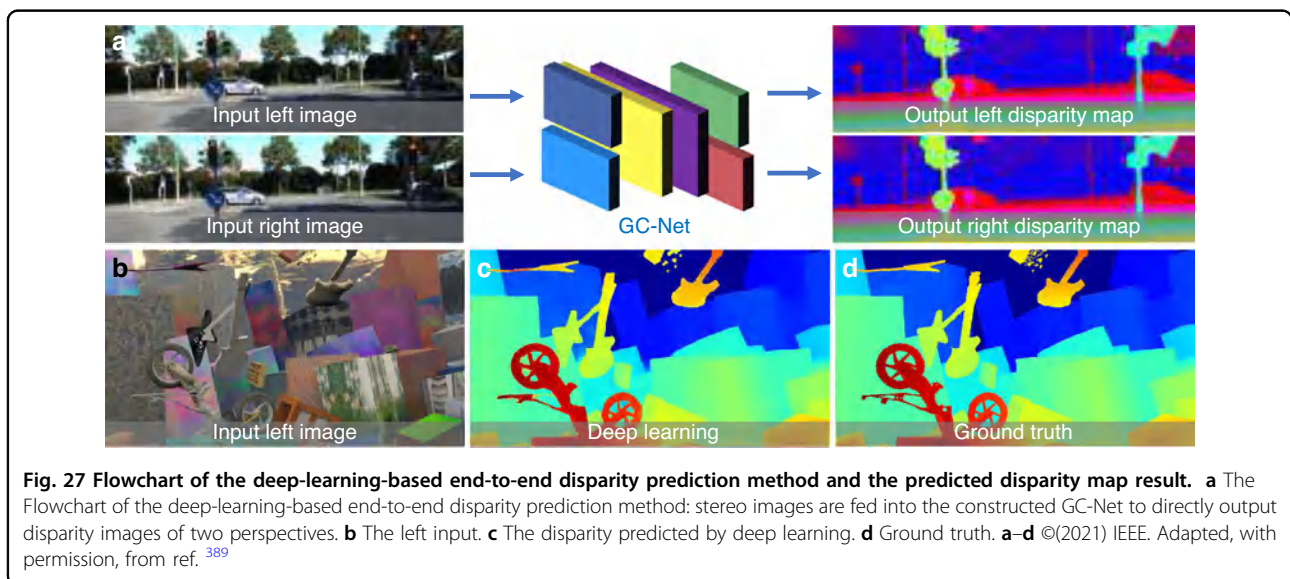
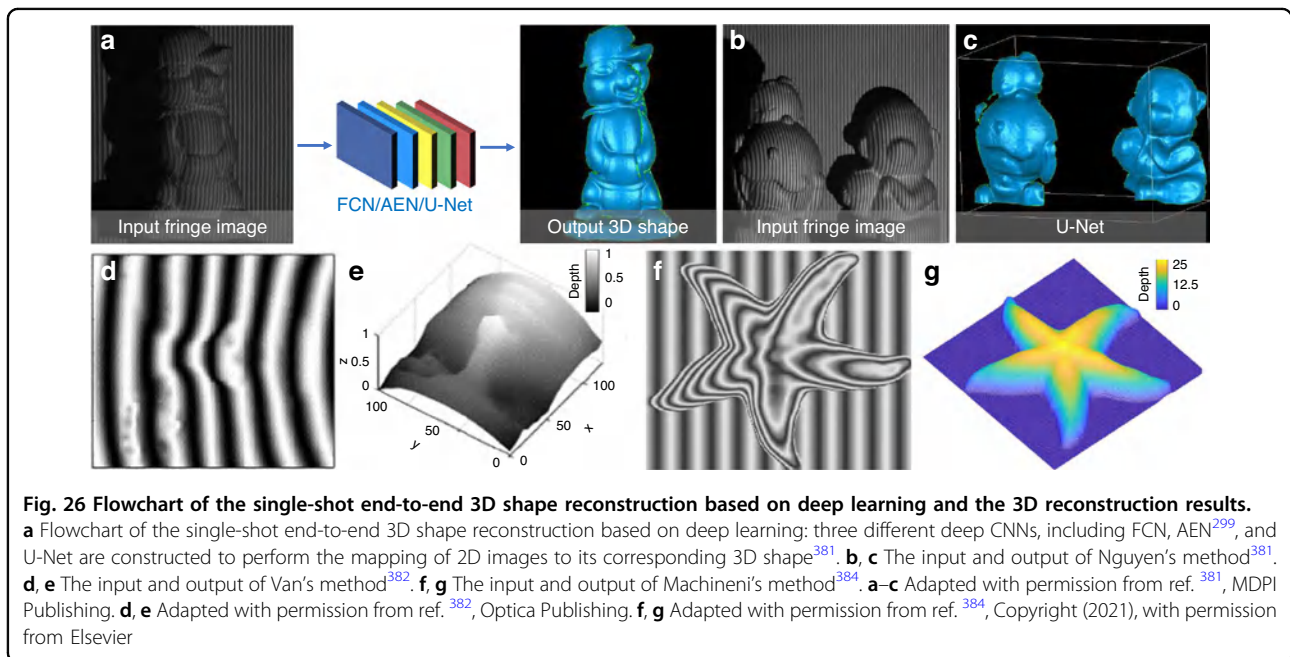
accuracy were extracted with the conventional circle center detection algorithm<sup>379</sup>, and the horizontal coordinates  $X_{pi}$  of the corresponding projector image for each mark center were calculated through the absolute phase value. Figure 25b shows the 3D reconstruction result of a standard stair sample predicted by the neural network. Figure 25c and d shows the error distributions of the measurement results obtained by traditional phase-height conversion method<sup>380</sup> and neural network, showing that the learning-based method was insensitive to the fringe intensity nonlinearity and could recover the 3D shape of a workpiece with high accuracy.

#### End-to-end learning in optical metrology

As mentioned earlier, “divide and conquer” is a core idea of solving complex optical metrology problems by breaking the whole image-processing pipeline into several modules or sub-steps. On a different note, deep learning enables direct mapping between the original input and the desired output, and the whole process can be trained as a whole, in an end-to-end fashion. Although somewhat brute-force, such a straightforward treatment has been extensively used in deep learning, and gradually introduced to many subfields of optical metrology, e.g., FPP and DIC.

- **From fringe to 3D shape:** In FPP, the imaging processing pipeline generally consists of pre-processing, phase demodulation, phase unwrapping,





and phase-to-height conversion. Deep learning provides a viable and efficient way to reconsider the whole problem from a holistic perspective, taking human intervention out of the loop and solving the “fringe to 3D shape” problem in a purely data-driven manner. Based on this idea, Nguyen et al.<sup>381</sup> proposed an end-to-end neural network to directly perform the mapping from a fringe pattern to its corresponding 3D shape, the flowchart of which is shown in Fig. 26a. Three different deep CNNs, including FCN, autoencoder<sup>299</sup>, and U-Net, were

trained based on the datasets obtained by the conventional multi-frequency phase-shifting profilometry method. Figure 26b, c gives an input and its corresponding ground-truth 3D shape. Figure 26c shows the best 3D reconstruction results predicted by the three networks with the depth measurement accuracy of 2mm. Van et al.<sup>382</sup> presented an SRCNN-based DNN to directly extract absolute height information from a single-fringe image. Through simulated fringe and depth image pairs, the trained network was able to obtain

high-accuracy full-field depth information from a single-fringe pattern. Recently, they compared the effect of different loss functions (MAE, MSE, and SSIM) on a modified U-Net for mapping a fringe image to the corresponding depth, and designed a new mixed gradient loss function that yielded higher-quality 3D reconstructions than conventional ones<sup>383</sup>. Machineni et al.<sup>384</sup> constructed a CNN with multiresolution similarity assessment to directly reconstruct the object's shape from the corresponding deformed fringe image. Their proposed method can achieve promising results under various challenging conditions such as low SNR, low fringe density, and high dynamic range. Zheng et al.<sup>385</sup> utilized the calibration matrix from a real-world FPP system to construct its "digital twin", which provided abundant simulation data (fringe pattern and corresponding depth map) required for the model training. The trained U-Net can then be employed to the real-world FPP system to extract the 3D geometry encoded in the fringe pattern in one step. Similarly, Wang et al.<sup>386</sup> constructed a virtual FPP system for the training dataset generation. A modified loss function based on SSIM index was employed, providing improved performance in terms of measurement accuracy and detail preservation.

- **From stereo images to disparity:** Deep learning can also be applied to DIC and stereophotogrammetry to bypass all intermediate image-processing steps in the pipeline for displacement and 3D reconstruction. Mayer et al.<sup>387</sup> presented end-to-end networks for the estimation of disparity (DispNet) and optical flow (FlowNet). In DispNet, a 1D correlation was proposed along the disparity line corresponding to the stereo cost volume. In addition, they also offered a large synthetic dataset, Scene Flow<sup>388</sup>, for training large-scale stereomatching networks. Kendall et al.<sup>389</sup> established an end-to-end Geometry and Context Network (GC-Net) mapping from a rectified pair of stereo images to disparity maps with subpixel accuracy (Fig. 27a). Stereo images were fed into the network to directly output disparity images of two perspectives. Figure 27b–d shows the test results on Scene Flow, where Fig. 27b is the left input, Fig. 27c is the disparity predicted by deep learning, and Fig. 27d is the ground truth. Experimental results show that the end-to-end learning method produced high-resolution disparity images and could tolerate large occlusions. Chang et al.<sup>390</sup> developed a pyramid stereomatching network (PSMNet) to enhance the matching accuracy by using the 3D CNN-based spatial pyramid pooling and multiple hourglass

networks. Zhang et al.<sup>391</sup> proposed a cost aggregation network incorporating the local guided filter and semi-global-matching-based cost aggregation, achieving higher matching quality as well as better network generalization. Recently, our group proposed an end-to-end speckle correlation strategy for 3D shape measurement, where a multiscale residual subnetwork was utilized to obtain feature maps of stereo speckle images, and the 4D cost volume at one-fourth of the original<sup>392</sup>. Besides, a saliency detection network was integrated to generate a pixel-wise mask to exclude the shadow-noised regions. Nguyen et al.<sup>393</sup> used three U-Net-based networks to convert a single speckle image into its corresponding 3D information. It should be mentioned that stereophotogrammetry is a representative field that deep learning has been extensively applied. Many other end-to-end deep-learning structures directly mapping stereo images to disparity have been proposed, such as hybrid CNN-CRF models<sup>394</sup>, Demon (CNN-based)<sup>395</sup>, MVSNet (CNN-based)<sup>396</sup>, CNN-based disparity estimation through feature constancy<sup>397</sup>, Segstereo<sup>398</sup>, EdgeStereo<sup>399</sup>, stereomatching with explicit cost aggregation architecture<sup>400</sup>, HyperDepth<sup>401</sup>, practical deep stereo (PDS)<sup>402</sup>, RNN-based stereomatching<sup>403,404</sup>, and unsupervised learning<sup>405–409</sup>. For DIC, Boukhtache et al.<sup>410</sup> presented an enhanced FlowNet (so-called StrainNet) to predict displacement and strain fields from pairs of deformed and reference images of a flat speckled surface. Their experimental results demonstrated the feasibility of the deep-learning approach for accurate pixel-wise subpixel measurement over full displacement fields. Min et al.<sup>411</sup> proposed a 3D CNN-based strain measurement method, which allowed simultaneous characterization in spatial and temporal domains from the surface images obtained during a tensile test of BeCu thin film. Rezaie et al.<sup>412</sup> compared the performance of conventional DIC method and their deep-learning method based on U-Net for detecting cracks on stone masonry wall images, showing that the learning-based method could detect most visible cracks and better preserve the crack geometry.

It should be mentioned that, not just limited to phase or correlation measurement techniques, deep learning has also been widely adopted in many other fields of optical metrology. However, due to space limitations, it is not possible to describe or discuss all of them. Examples include but are not limited to the time of flight (ToF)<sup>413–418</sup>, photometric stereo<sup>419–425</sup>, wavefront sensing<sup>426–429</sup>, aberrations characterization<sup>430</sup>, and fiber optic imaging<sup>431–435</sup>, etc.

After reviewing hundreds of recent works leveraging deep learning for different optical metrology tasks, readers may still be interested to know to apply these new data-driven approaches to their own problems or projects. To help the reader, we present a step-by-step guide to applying deep learning to optical metrology in the Supplementary Information, taking phase demodulation from a single-fringe pattern as an example. We explain how to build a DNN with fully convolutional network architectures and train it with the experimentally collected training dataset. We also distribute the source code and the corresponding datasets for this example. Based on this example, we demonstrate that a well-trained DNN can accomplish the phase-demodulation task in an accurate and efficient manner, using only a single-fringe pattern as input. Thus, it is capable of combining the single-frame strength of the spatial phase demodulation methods with the high-measurement accuracy of the temporal phase-demodulation methods. The interested reader may refer to the Supplementary Information for the step-by-step tutorial.

### Deep learning in optical metrology: challenges

Our review in the last section shows that the deep-learning solutions in optical metrology are straightforward, but have led to improved performance compared with the state-of-the-art. In this session, we will shift our attention to reveal some challenges of the use of deep learning in optical metrology, which require further attention and careful consideration:

- **High cost of collecting and labeling experimental training data:** Most of the deep-learning techniques reviewed belong to supervised learning, which requires a large amount of labeled data to train the network. To account for real experimental conditions, deep-learning approaches can benefit from large amounts of experimental training data. Since these data serve as ground truth with sufficiently high accuracy, they are usually expensive to collect<sup>436</sup>. In addition, since the optical metrology system is highly customized, training data collected by one system may not be suitable for another system of the same type. This may explain why there were far fewer publicly available datasets in the field of optical metrology (especially compared with the computer vision community). Without such public benchmark datasets, it is difficult to make a fair and standardized comparison between different algorithms. Although some emerging machine learning approaches, such as transfer learning<sup>437</sup>, few-shot learning<sup>438</sup>, unsupervised learning<sup>244</sup>, and weak-supervised learning<sup>439</sup>, can decrease the reliance on the amount of data
- to some extent, their performance is not comparable to that of supervised learning with large data numbers so far.
- **Ground truth inaccessible for experimental data:** In many areas of optical metrology, e.g., fringe or phase denoising, it is infeasible or even impossible to get the actual ground truth of the experimental data. As discussed in previous sections, generating a training dataset by simulating the forward image formation process can bypass this difficulty<sup>362,385</sup>, often at the price of compromised actual performance when the knowledge of the forward image formation model  $\mathcal{A}$  is imprecise or simulated dataset fails to reflect the real experimental system realistically and comprehensively. An alternative approach to this issue is to create a “quasi-experimental” dataset by collecting experimental raw data and then using the conventional state-of-the-art solutions to get the corresponding labels<sup>308–310</sup>. Essentially, the network is trained to “duplicate” the approximate inverse operator  $\tilde{\mathcal{A}}^{-1}$  corresponding to the conventional algorithm that is used to generate the labels. After training, the network is able to emulate the conventional reconstruction algorithm  $\widehat{\mathcal{R}}_{\theta}(\mathbf{I}) \approx \tilde{\mathcal{A}}^{-1}(\mathbf{I})$ , but the improvement in performance over conventional approaches becomes an unreasonable expectation.
- **Empiricism in network design and training:** So far, there is no standard paradigm for selecting appropriate DNN architectures because it requires a comprehensive understanding of the topology, training methods, and other parameters. In practice, we usually determine our network structure by evaluating different available candidate models, or comparing similar task-specific models by training them with different hyperparameters settings (network layers, neural units, and activation functions) on a specific validation dataset<sup>440</sup>. However, the overwhelming number of deep-learning models often limits one to evaluating only a few of the most trustworthy models, which may lead to suboptimal results. Therefore, one should learn how to quickly and efficiently narrow down the range of available models to find those most likely to be best performing on a specific type of problem. In addition, training a DNN is generally laborious and time-consuming, and becomes even worse with repetitive adjustments in the network architecture or hyperparameters to prevent overfitting and convergence issues.
- **Lack of generalization ability after specific sample training:** The generalization ability of



deep-learning approaches is closely related to the size and diversity of training samples. Generally, deep-learning architectures used in optical metrology are highly specialized to a specific domain, and they should be implemented with extreme care and caution when solving issues that do not pertain to the same domain. Thus, we cannot ignore the risk that when a never-before-experienced input differs even slightly from what they encountered at the training stage, the mapping  $\widehat{\mathcal{R}}_\theta$  established by deep networks may quickly stop making sense<sup>441</sup>. This is quite different from the traditional optical metrology solutions in which the reliability of the reconstruction can be secured for diverse types of samples as long as “the forward model  $\mathcal{A}$  is accurate” and “the corresponding reconstruction algorithm  $\tilde{\mathcal{A}}^{-1}$  is effective”.

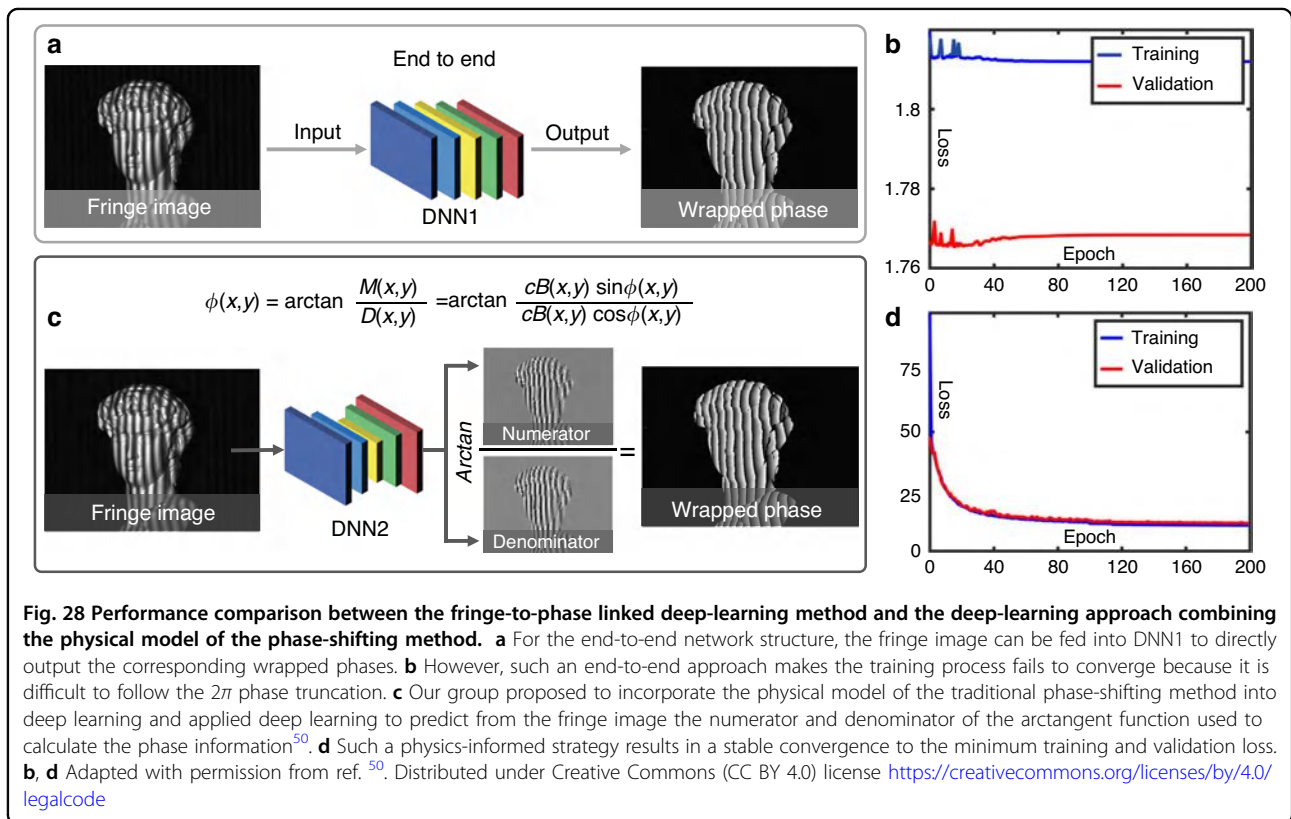
- **“Deep learning in computer vision”  $\neq$  “Deep learning in optical metrology”**: Deep learning is essentially the process of using computers to help us find the underlying patterns within the training dataset. Since the information cannot be “born out of nothing”, DNNs cannot always produce a provably correct solution. Compared with many computer vision tasks, optical metrology concerns more on accuracy, reliability, repeatability, and traceability<sup>442</sup>. For example, surface defect inspection is an indispensable quality-control procedure in manufacturing processes<sup>443</sup>. When using deep learning for optical metrological inspection, one may face the risk that a defect in an industrial component is “smoothed out” and undetected by an overfitted DNN in the inspection stage, which will make the entire production run defective. Since the success of deep learning depends on the “common” features learned and extracted from the training samples, which may lead to unsatisfactory results when facing “rare samples”.
- **“Deep learning” lacks the ability of “deep understanding”**: The “black box” nature of DNNs, which is arguably one of their most well-known disadvantages, prevents us from knowing how the neural network generates expected results from specific inputs by learning a large amount of training data. For example, when we send a fringe pattern into a neural network, and it outputs a poor phase image, it is not easy to comprehend what makes it arrive at such a prediction. Interpretability is critical in optical metrology because it ensures the traceability of the mistake. Consequently, most researchers in optical metrology community use deep-learning approaches in a pragmatic fashion

without the possibility to explain why it provides good results or without the ability to explain the logical bases and apply modifications in the case of underperformance.

### Deep learning in optical metrology: future directions

Although the above challenges have not been adequately addressed, optical metrology is now surfing the wave of deep learning, following a trend similarly being experienced in many other fields. This field is still young, but is expected to play an increasingly prominent role in the future development of optical metrology, especially with the evolution of computer science and AI technology.

- **Hybrid, composite, and automated learning**: It must be admitted that at this stage, deep-learning methods for optical metrology are still limited to some elementary techniques. There is further untapped potential as a number of latest innovations in deep learning can be directly introduced into the context of optical metrology. (1) Hybrid learning methods, such as semi-supervised<sup>242</sup>, unsupervised<sup>244</sup>, and self-supervised learning<sup>444</sup>, are capable of extracting valuable insights from unlabeled data, which is extremely attractive as the availability of ground-truth or labeled data in optical metrology is very limited. For example, GANs utilize two networks in a competitive manner, generator and discriminator, to deceive each other during the training process to generate the final prediction without specific labels<sup>266</sup>. In stereovision, the network models trained by unsupervised methods have been shown to produce better disparity prediction results in real scenes<sup>345</sup>. (2) Composite learning approaches attempt to combine different models pretrained on a similar task to produce a composite model with improved performance<sup>437</sup> or search for the optimal network architecture in the reinforcement learning environment for a certain dataset<sup>445</sup>. They are premised on the idea that a singular model, even very large, cannot outperform a compositional model with several small models/components, each being delegated to specialize in part of the task. As optical metrology tasks are getting more and more complicated, composite learning can deconstruct one huge task into several simpler, or single-function components and make them work together, or against each other, producing a more compressive and powerful model. (3) Automated machine learning (AutoML) approaches, such as Google AutoML<sup>446</sup> and Azure AutoML<sup>447</sup>, is developed to execute tedious modeling tasks that once performed by professional scientists<sup>440,448</sup>. It burns through an

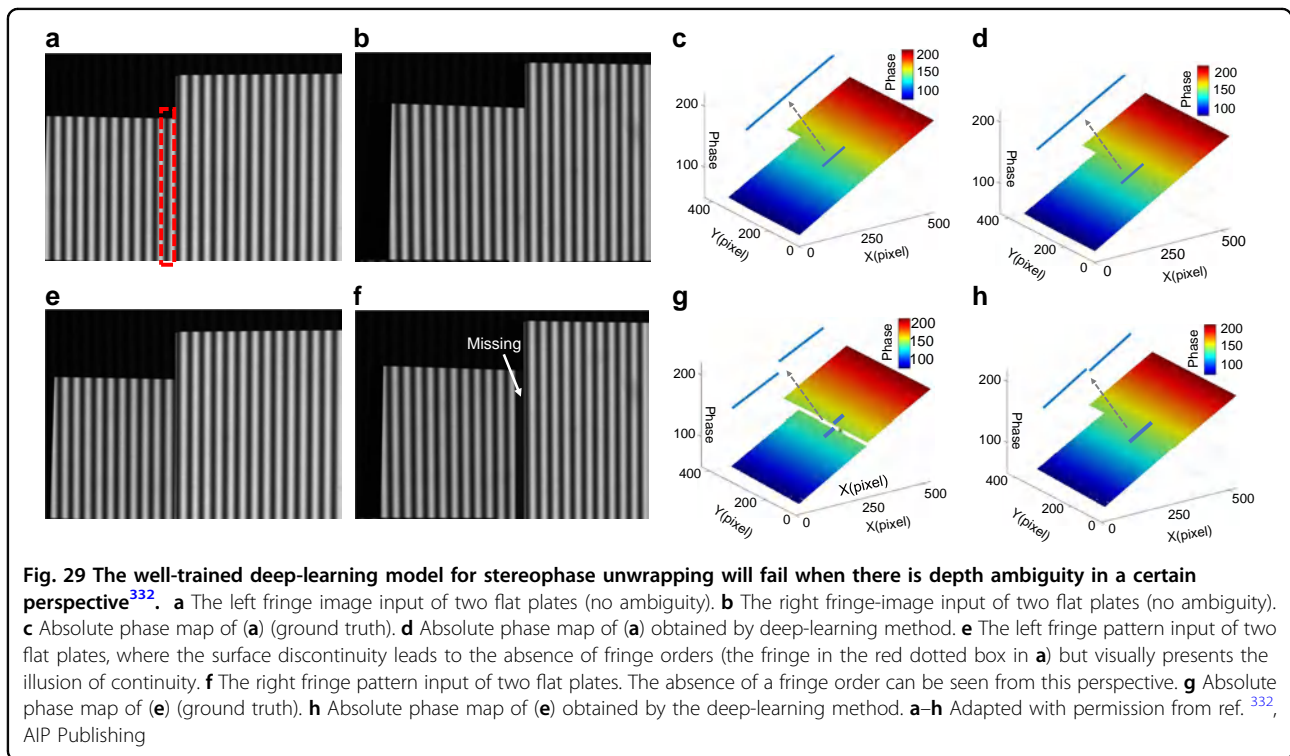


enormous number of models and the associated hyperparameters on the raw input data to decide what model is best applied to it. Consequently, AutoML is expected to permit even “citizen” AI scientists with their background in optical metrology to make streamlined use cases by only utilizing their domain expertise, offering practitioners a competitive advantage with minimum investments.

- **Physics-informed deep learning:** Unlike traditional physics-model-based optical metrology methods for which the domain knowledge is carefully engineered into solutions, most of the current deep-learning-based optical metrology methods do not benefit so much from such prior knowledge but rather learn the solution from scratch by making use of massive training data. In contrast, if the physics laws governing the image formation (the knowledge about the forward image formation model  $\mathcal{A}$ ) are known—even partially, they should be naturally incorporated into the DNN model so that the training data and network parameters are not wasted on “learning the physics”. For example, in fringe analysis, inspired by the conventional phase-shifting techniques, Feng et al.<sup>50</sup> proposed to learn the sine and cosine components of the fringe pattern, based on which the wrapped phase can be calculated by the arctangent function (Fig. 28c, d).

This method shows a significant gain in performance than directly using an end-to-end network structure<sup>50</sup> (Fig. 28a, b). Goy et al.<sup>302</sup> suggested a method for low-photon count phase retrieval where the noisy input image was converted into an approximant. As the approximant obtained by prior knowledge is much closer to the final prediction than the raw low-photon image, the phase reconstruction accuracy by using deep learning can be improved significantly. Wang et al.<sup>449</sup> incorporated the diffraction model of numerical propagation into a DNN for phase retrieval. By minimizing the difference between the actual input image and the predicted input image, DNN learns how to reconstruct the phase that best matches the measurements without any ground-truth data.

- **Interpretable deep learning:** As we have already highlighted in the previous sections, most researchers in optical metrology use deep-learning approaches intuitively without the possibility to explain why it produces such “good” results. This can be very problematic in high-stakes settings such as industrial inspection, quality control, and medical diagnose where the decisions of algorithms must be explainable, or where accountability is required. Academics in deep learning are acutely aware of



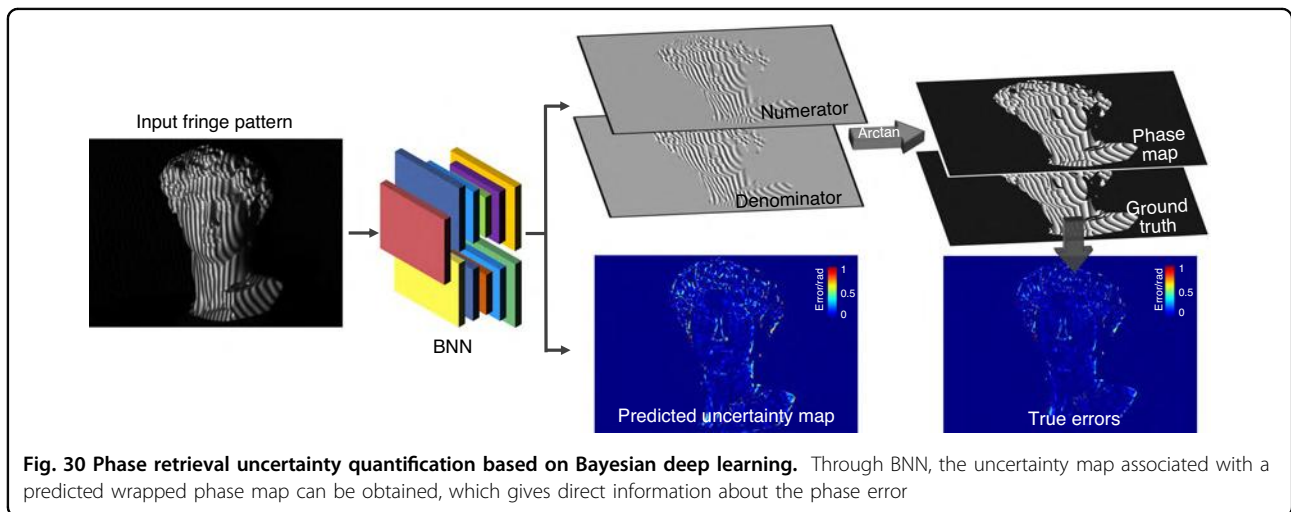
this interpretability problem, and there have been several developments in recent years for visualizing the features and representations they have learned by DNNs<sup>284</sup>. On the other hand, often applied to high-risk scenarios, optical metrology is among the most significant deep-learning challenges—we are dealing with unknown, uncertain, ambiguous, incomplete, noisy, inaccurate, and missing datasets in high-dimensional spaces. The unexplainability and incomprehensibility of deep learning also imply the predictions are at risk of failure. Figure 29 illustrates one such example, where a well-trained deep-learning model for stereophase unwrapping fails when there exists depth ambiguity in a certain perspective<sup>332</sup>. Therefore, explainability will become a key strength in deep-learning techniques to interpret and explain models, which would significantly expand the usefulness of deep-learning methods in optical metrology.

- **Uncertainty quantification:** Characterizing uncertainty in deep-learning solutions can help make better decisions and take precautions against erroneous predictions, which is essential for many optical metrology tasks<sup>450</sup>. However, most deep-learning methods reviewed in this work cannot provide uncertainty estimates. In recent years, Bayesian deep learning has emerged as a unified probabilistic framework that tightly integrates deep

learning with Bayesian models<sup>451</sup>. By using a GAN training framework to estimate a posterior distribution of images fitting a given measurement dataset (or estimation statistics derived from the posterior), Bayesian convolutional neural networks (BNNs) can quantify the reliability of predictions through two predictive uncertainties, including model uncertainty and data uncertainty, akin to epistemic and revelation uncertainty in Bayesian analysis, respectively<sup>452</sup>. It is expected to be adopted in optical metrology applications, e.g., fringe pattern analysis, to give pixel-wise variance estimates and data uncertainty evaluation (Fig. 30)<sup>453</sup>. The latter further allows assessment of the randomness of predictions stemming from data imperfections, including noise, incompleteness of the training data, and other experimental perturbations. Incorporating similar uncertainty quantification into other deep-learning-based optical metrology methods, especially when the ground truth is unavailable, is an interesting direction for future research.

- **Guiding the metrology system design:** Most of the current work using deep learning in optical metrology only considers how to reconstruct the measured data as a postprocessing algorithm while ignoring the way how the image data should be formed. However, an important feature of optical metrology methods is their active nature, especially





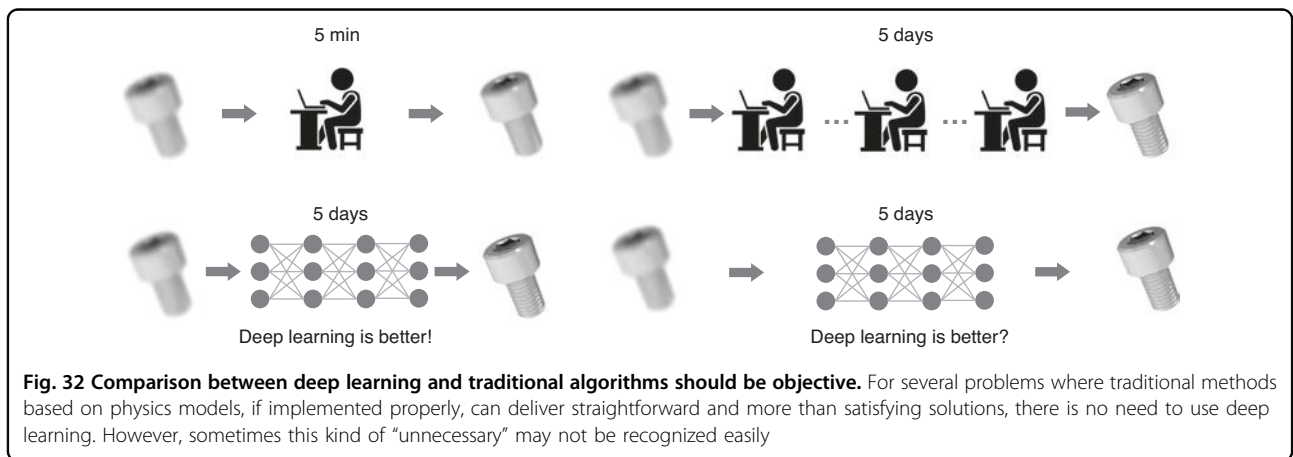
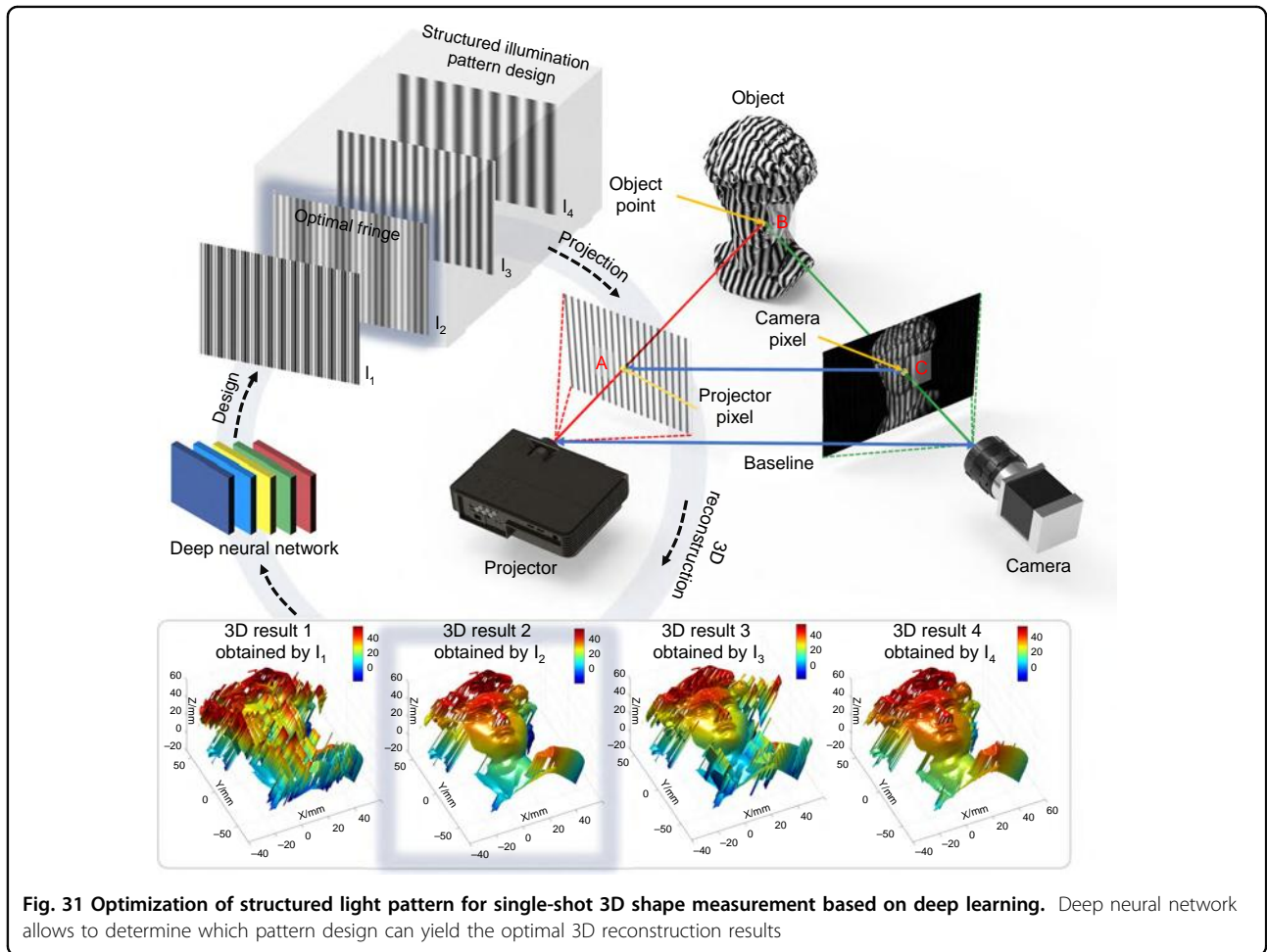
with respect to the way of manipulating the illumination. For example, in FPP, the structure of the illumination is modulated systematically throughout the object surface to deliver high accuracy and robustness in establishing the triangulation. The design of the illumination coding strategy is curial to improving the measurement accuracy removing the ambiguity of the depth reconstruction with a minimum number of image acquisitions. However, this problem has long been tackled using heuristics like composite coding, frequency multiplexing, and color multiplexing, which does not guarantee optimality (in terms of facilitating the recovery of desired information). Deep learning provides a mechanism to optimize the system design in a more principled way. By integrating the image formation model (with trainable parameters controlling the image acquisition) into the reconstruction network, the system design and the reconstruction algorithm (i.e., both  $\mathcal{A}$  and the corresponding  $\widehat{\mathcal{R}}_{\theta}$ ) can be jointly optimized with the training data<sup>454</sup>. It allows us to determine which type of system design can yield the best results for a particular deep-learning-driven task. Such an idea has been successfully demonstrated in designing optimal illumination patterns for computational microscopes<sup>455–457</sup>. We hope that this “joint optimization” network can effectively bridge the gap between how images should be acquired and how these images should be post-processed by deep learning, and can be widely adopted in designing the optical metrology systems, such as the fringe pattern design in FPP (Fig. 31), and the speckle pattern design in DIC, etc.

- **Both “deep” and “in-depth”:** Should we use deep learning or traditional optical metrology algorithms?

It is a tough question to answer because it depends heavily on the problem to be solved. Considering the “no free lunch theorem”, the selection between deep-learning and traditional algorithms should be considered rationally. For several problems where traditional methods based on physics models, if implemented properly, can deliver straightforward and more than satisfying solutions, there is no need to use deep learning. However, sometimes this kind of “unnecessary” may not be recognized easily. While being functionally effective, we should keep in mind that “how best deep learning can do” generally depends on “how reliable the training data we can provide.” For example, though the popular “learning from simulation” scheme used in optical metrology eliminates the dependence on huge labeled experimental data, the inconsistency between the image formation model and actual experimental condition leads to additional challenges of “domain adaptation”. Therefore, our personal view is that deep learning does not (at least at the current stage) make our research easier. On the contrary, it raises the threshold for optical metrology research because it requires researchers not only need to use and understand deep learning deeply but also need to take “in-depth” research in traditional algorithms so as to make an impartial and objective assessment between deep learning and traditional optical metrology algorithms (Fig. 32).

## Conclusions

A brief summary of this review indicates that there has been significant interest in the advancement of optical metrology technologies using deep-learning architectures. The rapid development of deep-learning technology has led to a paradigm shift from physics- and



knowledge-based modeling to data-driven learning for solving a wide range of optical metrology tasks. In general, deep learning is particularly advantageous for many problems in optical metrology whose physical models are complicated and acquired information is limited, e.g., in harsh environments and many challenging applications.

Strong empirical and experimental evidence suggests that using problem-specific deep-learning models outperforms conventional knowledge or physical model-based approaches.

Despite the promising—in many cases pretty impressive—results that have been reported in the literature,

potential problems and challenges remain. For model training, we need to acquire large amounts of experimental data with labels, which, even if available, is laborious and requires professional experts. We have been looking for the theoretical groundwork that would clearly explain the mechanisms and ways to the optimal selection of network structure and training algorithm for a specific task, or to profoundly comprehend why a particular network structure or algorithm is effective in a given task or not. Furthermore, deep-learning approaches have often been regarded as “black boxes”, and in optical metrology, accountability is essential and can cause severe consequences. Combining Bayesian statistics with deep neuron networks to obtain quantitative uncertainty estimates allows us to assess when the network yields unreliable predictions. A synergy of the physics-based models that describe the a priori knowledge of the image formation and data-driven models that learn a regularizer from the experimental data can bring our domain expertise into deep learning to provide more physically plausible solutions to specific optical metrology problems. Leveraging these emerging technologies in the application of deep-learning methods to optical metrology could promote and accelerate the recognition and acceptance of deep learning in more application areas. These are among the most critical issues that will continue to attract the interest of deep-learning research in the optical metrology community in the years to come.

In summary, although for different optical metrology tasks, deep-learning techniques can bring substantial improvements compared to traditional methods, the field is still at the early stage of development. Many researchers are still skeptical and maintain a wait-and-see attitude towards its applications involving industrial inspection and medical care, etc. Shall we accept deep learning as the key problem-solving tool? Or should we reject such a black-box solution? These are controversial issues in the optical metrology community today. Looking on the bright side, it has promoted an exciting trend and fostered expectations of the transformative potential it may bring to the optical metrology society. However, we should not overestimate the power of deep learning by considering it as a silver bullet for every challenge encountered in the future development of optical metrology. In practice, we should assess whether the large amount of data and computational resources required to use deep learning for a particular task is worthwhile, especially when other conventional algorithms may yield comparable performance with lower complexity and higher interpretability. We envisage that deep learning will not replace the role of traditional technologies within the field of optical metrology for the years to come, but will form a cooperative and complementary relationship, which may eventually become a symbiotic relationship in the future.

#### Acknowledgements

This work was supported by National Natural Science Foundation of China (U21B2033, 62075096, 62005121), Leading Technology of Jiangsu Basic Research Plan (BK20192003), “333 Engineering” Research Project of Jiangsu Province (BRA2016407), Jiangsu Provincial “One belt and one road” innovation cooperation project (BZ2020007), Fundamental Research Funds for the Central Universities (30921011208, 30919011222, 30920032101), and Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense (JSGP202105).

#### Author details

<sup>1</sup>Smart Computational Imaging (SCI) Laboratory, Nanjing University of Science and Technology, 210094 Nanjing, Jiangsu Province, China. <sup>2</sup>Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing University of Science and Technology, 210094 Nanjing, Jiangsu Province, China. <sup>3</sup>School of Engineering and Materials Science, Queen Mary University of London, London E1 4NS, UK. <sup>4</sup>School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore

#### Author contributions

C.Z.: conceptualization, writing—original draft, data curation, visualization, supervision, project administration, and funding acquisition. J.Q.: investigation, writing—review, visualization, and editing. S.F.: writing—review and editing. W.Y.: writing—review and editing. Y.L.: writing—review, visualization, and editing. P.F.: writing—review and editing. J.H.: writing—review and editing. K.Q.: writing—review and editing. Q.C.: resources, supervision, project administration, and funding acquisition.

#### Conflict of interest

The authors declare no competing interests.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41377-022-00714-x>.

Received: 11 July 2021 Revised: 3 January 2022 Accepted: 11 January 2022  
Published online: 23 February 2022

#### References

- Gåsvik, K. J. *Optical Metrology*, 3rd edn. (Wiley, 2002).
- Yoshizawa, T. *Handbook of Optical Metrology: Principles and Applications*, 2nd edn. (CRC Press, 2017).
- Sirohi, R. S. *Introduction to Optical Metrology* (CRC Press, 2016).
- Malacara, D. *Optical Shop Testing*, 3rd edn. (John Wiley & Sons, 2007).
- Harding, K. *Handbook of Optical Dimensional Metrology* (CRC Press, 2013).
- Chen, Z. G. & Segev, M. Highlighting photonics: looking into the next decade. *eLight* **1**, 2 (2021).
- Kleppner, D. On the matter of the meter. *Phys. Today* **54**, 11–12 (2001).
- Kulkarni, R. & Rastogi, P. Optical measurement techniques—a push for digitization. *Opt. Lasers Eng.* **87**, 1–17 (2016).
- Chen, F., Brown, G. M. & Song, M. M. Overview of 3-D shape measurement using optical methods. *Optical Eng.* **39**, 10–22 (2000).
- Blais, F. Review of 20 years of range sensor development. *J. Electron. Imaging* **13**, 231–243 (2004).
- Rastogi, P. *Digital Optical Measurement Techniques and Applications* (Artech House, 2015).
- Osten, W. Optical metrology: the long and unstoppable way to become an outstanding measuring tool. In *Proceedings of SPIE 10834, Speckle 2018: VII International Conference on Speckle Metrology*. 1083402 (SPIE, Janów Podlaski, Poland, 2018).
- Wyant, J. C. & Creath, K. Recent advances in interferometric optical testing. *Laser Focus* **21**, 118–132 (1985).
- Takeda, M. & Kujawinska, M. Lasers revolutionized optical metrology. <https://spie.org/news/spie-professional-magazine-archive/2010-october/lasers-revolutionized-optical-metrology?SSO=1> (2010).
- Denisyuk, Y. N. On the reflection of optical properties of an object in a wave field of light scattered by it. *Dokl. Akad. Nauk SSSR* **144**, 1275–1278 (1962).



16. Leith, E. N. & Upatnieks, J. Reconstructed wavefronts and communication theory. *J. Optical Soc. Am.* **52**, 1123–1130 (1962).
17. Gabor, D. A new microscopic principle. *Nature* **161**, 777–778 (1948).
18. Reid, G. T. Automatic fringe pattern analysis: a review. *Opt. Lasers Eng.* **7**, 37–68 (1986).
19. Rajshekhkar, G. & Rastogi, P. Fringe analysis: premise and perspectives. *Opt. Lasers Eng.* **50**, iii–x (2012).
20. Rastogi, P. & Hack, E. *Phase Estimation in Optical Interferometry* (CRC Press, 2015).
21. Hariharan, P., Oreb, B. F. & Eiju, T. Digital phase-shifting interferometry: a simple error-compensating phase calculation algorithm. *Appl. Opt.* **26**, 2504–2506 (1987).
22. Schnars, U. & Jüptner, W. *Digital Holography: Digital Hologram Recording, Numerical Reconstruction, and Related Techniques* (Springer Science & Business Media, 2005).
23. Pan, B. et al. Two-dimensional digital image correlation for in-plane displacement and strain measurement: a review. *Meas. Sci. Technol.* **20**, 062001 (2009).
24. Raskar, R., Agrawal, A. & Tumblin, J. Coded exposure photography: motion deblurring using fluttered shutter. *ACM Trans. Graph.* **25**, 795–804 (2006).
25. Ritschl, L. et al. Improved total variation-based CT image reconstruction applied to clinical data. *Phys. Med. Biol.* **56**, 1545 (2011).
26. Edgar, M. P., Gibson, G. M. & Padgett, M. J. Principles and prospects for single-pixel imaging. *Nat. Photonics* **13**, 13–20 (2019).
27. Katz, O. et al. Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations. *Nat. Photonics* **8**, 784–790 (2014).
28. Stuart, A. M. Inverse problems: a Bayesian perspective. *Acta Numerica* **19**, 451–559 (2010).
29. Osher, S. et al. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling Simul.* **4**, 460–489 (2005).
30. Goldstein, T. & Osher, S. The split Bregman method for L1-regularized problems. *SIAM J. Imaging Sci.* **2**, 323–343 (2009).
31. Osten, W. What optical metrology can do for experimental mechanics? *Appl. Mech. Mater.* **70**, 1–20 (2011).
32. Zuo, C. et al. Phase shifting algorithms for fringe projection profilometry: a review. *Opt. Lasers Eng.* **109**, 23–59 (2018).
33. Baraniuk, R. G. Compressive sensing [lecture notes]. *IEEE Signal Process. Mag.* **24**, 118–121 (2007).
34. Zibulevsky, M. & Elad, M. L1-L2 optimization in signal and image processing. *IEEE Signal Process. Mag.* **27**, 76–88 (2010).
35. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
36. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press Cambridge, 2016).
37. Chang, X. Y., Bian, L. H. & Zhang, J. Large-scale phase retrieval. *eLight* **1**, 4 (2021).
38. Fukushima, K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36**, 193–202 (1980).
39. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015).
40. Baccouche, M. et al. Sequential deep learning for human action recognition. In *Proceedings of the 2nd International Workshop on Human Behavior Understanding*. 29–39 (Springer, Amsterdam, 2011).
41. Charles, R. Q. et al. PointNet: deep learning on point sets for 3D classification and segmentation. In *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. 77–85 (IEEE, Honolulu, 2017).
42. Ouyang, W. L. & Wang, X. G. Joint deep learning for pedestrian detection. In *Proceedings of 2013 IEEE International Conference on Computer Vision*. 2056–2063 (IEEE, Sydney, NSW, 2013).
43. Dong, C. et al. Learning a deep convolutional network for image super-resolution. In *Proceedings of 13th European Conference on Computer Vision*. 184–199 (Springer, Zurich, 2014).
44. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
45. Barbastathis, G., Ozcan, A. & Situ, G. On the use of deep learning for computational imaging. *Optica* **6**, 921–943 (2019).
46. Wang, H. D. et al. Deep learning enables cross-modality super-resolution in fluorescence microscopy. *Nat. Methods* **16**, 103–110 (2019).
47. Rivenson, Y. et al. Phase recovery and holographic image reconstruction using deep learning in neural networks. *Light: Sci. Appl.* **7**, 17141 (2018).
48. Wang, F. et al. Learning from simulation: an end-to-end deep-learning approach for computational ghost imaging. *Opt. Express* **27**, 25560–25572 (2019).
49. Li, S. et al. Imaging through glass diffusers using densely connected convolutional networks. *Optica* **5**, 803–813 (2018).
50. Feng, S. J. et al. Fringe pattern analysis using deep learning. *Adv. Photonics* **1**, 025001 (2019).
51. Shi, J. S. et al. Label enhanced and patch based deep learning for phase retrieval from single frame fringe pattern in fringe projection 3D measurement. *Opt. Express* **27**, 28929–28943 (2019).
52. Yin, W. et al. Temporal phase unwrapping using deep learning. *Sci. Rep.* **9**, 20175 (2019).
53. Zhang, T. et al. Rapid and robust two-dimensional phase unwrapping via deep learning. *Opt. Express* **27**, 23173–23185 (2019).
54. Hao, F. G. et al. Batch denoising of ESPI fringe patterns based on convolutional neural network. *Appl. Opt.* **58**, 3338–3346 (2019).
55. Yan, K. T. et al. Fringe pattern denoising based on deep learning. *Opt. Commun.* **437**, 148–152 (2019).
56. Gerchberg, R. W. & Saxton, W. O. A practical algorithm for the determination of phase from image and diffraction plane pictures. *Optik* **35**, 237–246 (1972).
57. Fienup, J. R. Phase retrieval algorithms: a comparison. *Appl. Opt.* **21**, 2758–2769 (1982).
58. Teague, M. R. Deterministic phase retrieval: a Green's function solution. *J. Optical Soc. Am.* **73**, 1434–1441 (1983).
59. Zuo, C. et al. Transport of intensity equation: a tutorial. *Opt. Lasers Eng.* **135**, 106187 (2020).
60. Zhang, F. C., Pedrini, G. & Osten, W. Phase retrieval of arbitrary complex-valued fields through aperture-plane modulation. *Phys. Rev. A* **75**, 043805 (2007).
61. Faulkner, H. M. L. & Rodenburg, J. M. Movable aperture lensless transmission microscopy: a novel phase retrieval algorithm. *Phys. Rev. Lett.* **93**, 023903 (2004).
62. Zheng, G. N. et al. Concept, implementations and applications of Fourier ptychography. *Nat. Rev. Phys.* **3**, 207–223 (2021).
63. Platt, B. C. & Shack, R. History and principles of Shack-Hartmann wavefront sensing. *J. Refractive Surg.* **17**, S573–S577 (2001).
64. Ragazzoni, R. Pupil plane wavefront sensing with an oscillating prism. *J. Mod. Opt.* **43**, 289–293 (1996).
65. Falldorf, C., von Kopylow, C. & Bergmann, R. B. Wave field sensing by means of computational shear interferometry. *J. Optical Soc. Am. A* **30**, 1905–1912 (2013).
66. Fienup, J. R. Phase retrieval for optical metrology: past, present and future. in *Optical Fabrication and Testing* (eds Reinhard, V.) 2017. OW2B-1 (Optical Society of America, 2017).
67. Claus, D. et al. Dual wavelength optical metrology using ptychography. *J. Opt.* **15**, 035702 (2013).
68. Falldorf, C., Agour, M. & Bergmann, R. B. Digital holography and quantitative phase contrast imaging using computational shear interferometry. *Optical Eng.* **54**, 024110 (2015).
69. Creath, K. V. phase-measurement interferometry techniques. *Prog. Opt.* **26**, 349–393 (1988).
70. Hariharan, P. *Basics of Interferometry*, 2nd edn. (Elsevier, 2007).
71. Aben, H. & Guillemet, C. Integrated photoelasticity. in *Photoelasticity of Glass* (eds Aben, H. & Guillemet, C.) 86–101 (Springer, 1993).
72. Asundi, A. Phase shifting in photoelasticity. *Exp. Tech.* **17**, 19–23 (1993).
73. Ramesh, K. & Lewis, G. Digital photoelasticity: advanced techniques and applications. *Appl. Mech. Rev.* **55**, B69–B71 (2002).
74. Sciammarella, C. A. The moiré method—a review. *Exp. Mech.* **22**, 418–433 (1982).
75. Post, D., Han, B. & Ifju, P. *High Sensitivity Moiré: Experimental Analysis for Mechanics and Materials*. (Springer Science & Business Media, 2012).
76. Durelli, A. J. & Parks, V. J. *Moiré Analysis of Strain* (Prentice Hall, 1970).
77. Chiangi, F. P. Moiré methods of strain analysis. *Exp. Mech.* **19**, 290–308 (1979).
78. Post, D., Han, B. & Ifju, P. Moiré interferometry. in *High Sensitivity Moiré: Experimental Analysis for Mechanics and Materials* (eds Post, D., Han, B. & Ifju, P.) 135–226 (Springer, 1994).
79. Rastogi, P. K. *Holographic Interferometry: Principles and Methods* (Springer-Verlag, 1994).
80. Kreis, T. *Handbook of Holographic Interferometry: Optical and Digital Methods* (John Wiley & Sons, 2004).

81. Hariharan, P., Oreb, B. F. & Brown, N. Real-time holographic interferometry: a microcomputer system for the measurement of vector displacements. *Appl. Opt.* **22**, 876–880 (1983).
82. Heflinger, L. O., Wuerker, R. F. & Brooks, R. E. Holographic interferometry. *J. Appl. Phys.* **37**, 642–649 (1966).
83. Khanna, S. M. & Tonndorf, J. Tympanic membrane vibrations in cats studied by time-averaged holography. *J. Acoustical Soc. Am.* **51**, 1904–1920 (1972).
84. Tonndorf, J. & Khanna, S. M. Tympanic-membrane vibrations in human cadaver ears studied by time-averaged holography. *J. Acoustical Soc. Am.* **52**, 1221–1233 (1972).
85. Schnars, U. et al. Digital holography. in *Digital Holography and Wavefront Sensing: Principles, Techniques and Applications* 2nd edn. (eds Schnars, U. et al.) 39–68 (Springer, 2015).
86. CuChe, E., Bevilacqua, F. & Depeursinge, C. Digital holography for quantitative phase-contrast imaging. *Opt. Lett.* **24**, 291–293 (1999).
87. Xu, L. et al. Studies of digital microscopic holography with applications to microstructure testing. *Appl. Opt.* **40**, 5046–5051 (2001).
88. Picart, P. et al. Time-averaged digital holography. *Opt. Lett.* **28**, 1900–1902 (2003).
89. Singh, V. R. et al. Dynamic characterization of MEMS diaphragm using time averaged in-line digital holography. *Opt. Commun.* **280**, 285–290 (2007).
90. Colomb, T. et al. Automatic procedure for aberration compensation in digital holographic microscopy and applications to specimen shape compensation. *Appl. Opt.* **45**, 851–863 (2006).
91. Løkberg, O. J. Electronic speckle pattern interferometry. in *Optical Metrology* (ed. Soares, O. D. D.) 542–572 (Springer, 1987).
92. Rastogi, P. K. *Digital Speckle Pattern Interferometry and Related Techniques* (Wiley, 2001).
93. Hung, Y. Y. Shearography: a new optical method for strain measurement and nondestructive testing. *Optical Eng.* **21**, 213391 (1982).
94. Hung, Y. Y. & Ho, H. P. Shearography: an optical measurement technique and applications. *Mater. Sci. Eng.: R: Rep.* **49**, 61–87 (2005).
95. Gorghi, S. S. & Rastogi, P. Fringe projection techniques: whither we are? *Opt. Lasers Eng.* **48**, 133–140 (2010).
96. Geng, J. Structured-light 3D surface imaging: a tutorial. *Adv. Opt. Photonics* **3**, 128–160 (2011).
97. Knauer, M. C., Kaminski, J. & Hausler, G. Phase measuring deflectometry: a new approach to measure specular free-form surfaces. In *Proceedings of SPIE 5457, Optical Metrology in Production Engineering*. 366–376 (IEEE, Strasbourg, 2004).
98. Huang, L. et al. Review of phase measuring deflectometry. *Opt. Lasers Eng.* **107**, 247–257 (2018).
99. Zhang, Z. H. et al. Three-dimensional shape measurements of specular objects using phase-measuring deflectometry. *Sensors* **17**, 2835 (2017).
100. Xu, Y. J., Gao, F. & Jiang, X. Q. A brief review of the technological advancements of phase measuring deflectometry. *Photonix* **1**, 14 (2020).
101. Chu, T. C., Ranson, W. F. & Sutton, M. A. Applications of digital-image-correlation techniques to experimental mechanics. *Exp. Mech.* **25**, 232–244 (1985).
102. Schreier, H., Orteu, J. J. & Sutton, M. A. *Image Correlation for Shape, Motion and Deformation Measurements: Basic Concepts. Theory and Applications* (Springer, 2009).
103. Verhulst, E., van Rietbergen, B. & Huiskes, R. A three-dimensional digital image correlation technique for strain measurements in microstructures. *J. Biomech.* **37**, 1313–1320 (2004).
104. Sutton, M. A. et al. The effect of out-of-plane motion on 2D and 3D digital image correlation measurements. *Opt. Lasers Eng.* **46**, 746–757 (2008).
105. Pan, B. Digital image correlation for surface deformation measurement: historical developments, recent advances and future goals. *Meas. Sci. Technol.* **29**, 082001 (2018).
106. Marr, D. & Poggio, T. A computational theory of human stereo vision. *Philos. Trans. R. Soc. B: Biol. Sci.* **204**, 301–328 (1979).
107. Luhmann, T. et al. *Close-Range Photogrammetry and 3D Imaging*, 2nd edn. (De Gruyter, 2014).
108. Fusiello, A., Trucco, E. & Verri, A. A compact algorithm for rectification of stereo pairs. *Mach. Vis. Appl.* **12**, 16–22 (2000).
109. Pitas, I. *Digital Image Processing Algorithms and Applications* (Wiley, 2000).
110. Yu, Q. F. et al. Spin filtering with curve windows for interferometric fringe patterns. *Appl. Opt.* **41**, 2650–2654 (2002).
111. Tang, C. et al. Second-order oriented partial-differential equations for denoising in electronic-speckle-pattern interferometry fringes. *Opt. Lett.* **33**, 2179–2181 (2008).
112. Wang, H. X. et al. Fringe pattern denoising using coherence-enhancing diffusion. *Opt. Lett.* **34**, 1141–1143 (2009).
113. Kaufmann, G. H. & Galizzi, G. E. Speckle noise reduction in television holography fringes using wavelet thresholding. *Optical Eng.* **35**, 9–14 (1996).
114. Kemao, Q. Windowed Fourier transform for fringe pattern analysis. *Appl. Opt.* **43**, 2695–2702 (2004).
115. Kemao, Q. Two-dimensional windowed Fourier transform for fringe pattern analysis: principles, applications and implementations. *Opt. Lasers Eng.* **45**, 304–317 (2007).
116. Bianco, V. et al. Quasi noise-free digital holography. *Light.: Sci. Appl.* **5**, e16142 (2016).
117. Kulkarni, R. & Rastogi, P. Fringe denoising algorithms: a review. *Opti. Lasers Eng.* <https://doi.org/10.1016/j.optlaseng.2020.106190> (2020).
118. Bianco, V. et al. Strategies for reducing speckle noise in digital holography. *Light.: Sci. Appl.* **7**, 48 (2018).
119. Zhi, H. & Johansson, R. B. Adaptive filter for enhancement of fringe patterns. *Opt. Lasers Eng.* **15**, 241–251 (1991).
120. Trusiak, M., Patorski, K. & Wielgus, M. Adaptive enhancement of optical fringe patterns by selective reconstruction using FABEMD algorithm and Hilbert spiral transform. *Opt. Express* **20**, 23463–23479 (2012).
121. Wang, C. X., Qian, K. M. & Da, F. P. Automatic fringe enhancement with novel bidimensional sinusoids-assisted empirical mode decomposition. *Opt. Express* **25**, 24299–24311 (2017).
122. Hsung, T. C., Lun, D. P. K. & Ng, W. W. L. Efficient fringe image enhancement based on dual-tree complex wavelet transform. *Appl. Opt.* **50**, 3973–3986 (2011).
123. Awatsuji, Y. et al. Single-shot phase-shifting color digital holography. In *IEEE Lasers and Electro-Optics Society Annual Meeting Conference Proceedings*. 84–85 (IEEE, Lake Buena Vista, FL, 2007).
124. Zhang, Z. H. Review of single-shot 3D shape measurement by phase calculation-based fringe projection techniques. *Opt. Lasers Eng.* **50**, 1097–1106 (2012).
125. Phillips, Z. F., Chen, M. & Waller, L. Single-shot quantitative phase microscopy with color-multiplexed differential phase contrast (cDPC). *PLoS ONE* **12**, e0171228 (2017).
126. Sun, J. S. et al. Single-shot quantitative phase microscopy based on color-multiplexed Fourier ptychography. *Opt. Lett.* **43**, 3365–3368 (2018).
127. Fan, Y. et al. Single-shot isotropic quantitative phase microscopy based on color-multiplexed differential phase contrast. *APL Photonics* **4**, 121301 (2019).
128. Zhang, Z. H., Towers, C. E. & Towers, D. P. Time efficient color fringe projection system for 3D shape and color using optimum 3-frequency selection. *Opt. Express* **14**, 6444–6455 (2006).
129. Zhang, Y. B. et al. Color calibration and fusion of lens-free and mobile-phone microscopy images for high-resolution and accurate color reproduction. *Sci. Rep.* **6**, 27811 (2016).
130. Lee, W. et al. Single-exposure quantitative phase imaging in color-coded LED microscopy. *Opt. Express* **25**, 8398–8411 (2017).
131. Schemm, J. B. & Vest, C. M. Fringe pattern recognition and interpolation using nonlinear regression analysis. *Appl. Opt.* **22**, 2850–2853 (1983).
132. Schreier, H. W., Braasch, J. R. & Sutton, M. A. Systematic errors in digital image correlation caused by intensity interpolation. *Optical Eng.* **39**, 2915–2921 (2000).
133. Bing, P. et al. Performance of sub-pixel registration algorithms in digital image correlation. *Meas. Sci. Technol.* **17**, 1615 (2006).
134. Pan, B. et al. Study on subset size selection in digital image correlation for speckle patterns. *Opt. Express* **16**, 7037–7048 (2008).
135. Bruck, H. et al. Digital image correlation using Newton-Raphson method of partial differential correction. *Exp. Mech.* **29**, 261–267 (1989).
136. Massig, J. H. & Heppner, J. Fringe-pattern analysis with high accuracy by use of the Fourier-transform method: theory and experimental tests. *Appl. Opt.* **40**, 2081–2088 (2001).
137. Roddier, C. & Roddier, F. Interferogram analysis using Fourier transform techniques. *Appl. Opt.* **26**, 1668–1673 (1987).
138. Takeda, M., Ina, H. & Kobayashi, S. Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry. *J. Optical Soc. Am.* **72**, 156–160 (1982).

139. Su, X. Y. & Chen, W. J. Fourier transform profilometry: a review. *Opt. Lasers Eng.* **35**, 263–284 (2001).
140. Kemaq, Q. *Windowed Fringe Pattern Analysis* (SPIE Press, 2013).
141. Zhong, J. G. & Weng, J. W. Spatial carrier-fringe pattern analysis by means of wavelet transform: wavelet transform profilometry. *Appl. Opt.* **43**, 4993–4998 (2004).
142. Larkin, K. G., Bone, D. J. & Oldfield, M. A. Natural demodulation of two-dimensional fringe patterns. I. general background of the spiral phase quadrature transform. *J. Optical Soc. Am. A* **18**, 1862–1870 (2001).
143. Trusiak, M., Wielgus, M. & Patorski, K. Advanced processing of optical fringe patterns by automated selective reconstruction and enhanced fast empirical mode decomposition. *Opt. Lasers Eng.* **52**, 230–240 (2014).
144. Servin, M., Marroquin, J. L. & Cuevas, F. J. Demodulation of a single interferogram by use of a two-dimensional regularized phase-tracking technique. *Appl. Opt.* **36**, 4540–4548 (1997).
145. Servin, M., Marroquin, J. L. & Quiroga, J. A. Regularized quadrature and phase tracking from a single closed-fringe interferogram. *J. Optical Soc. Am. A* **21**, 411–419 (2004).
146. Kemaq, Q. & Soon, S. H. Sequential demodulation of a single fringe pattern guided by local frequencies. *Opt. Lett.* **32**, 127–129 (2007).
147. Wang, H. X. & Kemaq, Q. Frequency guided methods for demodulation of a single fringe pattern. *Opt. Express* **17**, 15118–15127 (2009).
148. Servin, M., Quiroga, J. A. & Padilla, J. M. *Fringe Pattern Analysis for Optical Metrology: Theory, Algorithms, and Applications* (Wiley-VCH, 2014).
149. Massie, N. A., Nelson, R. D. & Holly, S. High-performance real-time heterodyne interferometry. *Appl. Opt.* **18**, 1797–1803 (1979).
150. Bruning, J. H. et al. Digital wavefront measuring interferometer for testing optical surfaces and lenses. *Appl. Opt.* **13**, 2693–2703 (1974).
151. Srinivasan, V., Liu, H. C. & Halioua, M. Automated phase-measuring profilometry of 3-D diffuse objects. *Appl. Opt.* **23**, 3105–3108 (1984).
152. Wizinowich, P. L. Phase shifting interferometry in the presence of vibration: a new algorithm and system. *Appl. Opt.* **29**, 3271–3279 (1990).
153. Schreiber, H. & Bruning, J. H. Phase shifting interferometry. in *Optical Shop Testing*, 3rd edn. (ed. Malacara, D.) 547–666 (Wiley, 2007).
154. Goldstein, R. M., Zebker, H. A. & Werner, C. L. Satellite radar interferometry: two-dimensional phase unwrapping. *Radio Sci.* **23**, 713–720 (1988).
155. Su, X. Y. & Chen, W. J. Reliability-guided phase unwrapping algorithm: a review. *Opt. Lasers Eng.* **42**, 245–261 (2004).
156. Flynn, T. J. Two-dimensional phase unwrapping with minimum weighted discontinuity. *J. Optical Soc. Am. A* **14**, 2692–2701 (1997).
157. Ghiglia, D. C. & Romero, L. A. Minimum  $L^2$ -norm two-dimensional phase unwrapping. *J. Optical Soc. Am. A* **13**, 1999–2013 (1996).
158. Bioucas-Dias, J. M. & Valadao, G. Phase unwrapping via graph cuts. *IEEE Trans. Image Process.* **16**, 698–709 (2007).
159. Zappa, E. & Busca, G. Comparison of eight unwrapping algorithms applied to Fourier-transform profilometry. *Opt. Lasers Eng.* **46**, 106–116 (2008).
160. Zebker, H. A. & Lu, Y. P. Phase unwrapping algorithms for radar interferometry: residue-cut, least-squares, and synthesis algorithms. *J. Optical Soc. Am. A* **15**, 586–598 (1998).
161. Zhao, M. et al. Quality-guided phase unwrapping technique: comparison of quality maps and guiding strategies. *Appl. Opt.* **50**, 6214–6224 (2011).
162. Sansoni, G. et al. Three-dimensional imaging based on Gray-code light projection: characterization of the measuring algorithm and development of a measuring system for industrial applications. *Appl. Opt.* **36**, 4463–4472 (1997).
163. Sansoni, G., Carocci, M. & Rodella, R. Three-dimensional vision based on a combination of gray-code and phase-shift light projection: analysis and compensation of the systematic errors. *Appl. Opt.* **38**, 6565–6573 (1999).
164. Huntley, J. M. & Saldner, H. Temporal phase-unwrapping algorithm for automated interferogram analysis. *Appl. Opt.* **32**, 3047–3052 (1993).
165. Zhao, H., Chen, W. Y. & Tan, Y. S. Phase-unwrapping algorithm for the measurement of three-dimensional object shapes. *Appl. Opt.* **33**, 4497–4500 (1994).
166. Saldner, H. O. & Huntley, J. M. Temporal phase unwrapping: application to surface profiling of discontinuous objects. *Appl. Opt.* **36**, 2770–2775 (1997).
167. Cheng, Y. Y. & Wyant, J. C. Two-wavelength phase shifting interferometry. *Appl. Opt.* **23**, 4539–4543 (1984).
168. Creath, K., Cheng, Y. Y. & Wyant, J. C. Contouring aspheric surfaces using two-wavelength phase-shifting interferometry. *Opt. Acta: Int. J. Opt.* **32**, 1455–1464 (1985).
169. Towers, C. E., Towers, D. P. & Jones, J. D. C. Optimum frequency selection in multifrequency interferometry. *Opt. Lett.* **28**, 887–889 (2003).
170. Gushov, V. I. & Solodkin, Y. N. Automatic processing of fringe patterns in integer interferometers. *Opt. Lasers Eng.* **14**, 311–324 (1991).
171. Takeda, M. et al. Frequency-multiplex Fourier-transform profilometry: a single-shot three-dimensional shape measurement of objects with large height discontinuities and/or surface isolations. *Appl. Opt.* **36**, 5347–5354 (1997).
172. Zhong, J. G. & Wang, M. Phase unwrapping by lookup table method: application to phase map with singular points. *Optical Eng.* **38**, 2075–2080 (1999).
173. Burke, J. et al. Reverse engineering by fringe projection. In *Proceedings of SPIE 4778, Interferometry XI: Applications*. 312–324 (SPIE, Seattle, WA, 2002).
174. Zuo, C. et al. Temporal phase unwrapping algorithms for fringe projection profilometry: a comparative review. *Opt. Lasers Eng.* **85**, 84–103 (2016).
175. Tao, T. Y. et al. Real-time 3-D shape measurement with composite phase-shifting fringes and multi-view system. *Opt. Express* **24**, 20253–20269 (2016).
176. Liu, X. R. & Kofman, J. Background and amplitude encoded fringe patterns for 3D surface-shape measurement. *Opt. Lasers Eng.* **94**, 63–69 (2017).
177. Weise, T., Leibe, B. & Van Gool, L. Fast 3D scanning with automatic motion compensation. In *Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8 (IEEE, Minneapolis, MN, 2007).
178. Zuo, C. et al. Micro Fourier transform profilometry ( $\mu$ FTP): 3D shape measurement at 10,000 frames per second. *Opt. Lasers Eng.* **102**, 70–91 (2018).
179. An, Y. T., Hyun, J. S. & Zhang, S. Pixel-wise absolute phase unwrapping using geometric constraints of structured light system. *Opt. Express* **24**, 18445–18459 (2016).
180. Li, Z. W. et al. Multiview phase shifting: a full-resolution and high-speed 3D measurement framework for arbitrary shape dynamic objects. *Opt. Lett.* **38**, 1389–1391 (2013).
181. Bräuer-Burchardt, C. et al. High-speed three-dimensional measurements with a fringe projection-based optical sensor. *Optical Eng.* **53**, 112213 (2014).
182. Garcia, R. R. & Zakhor, A. Consistent stereo-assisted absolute phase unwrapping methods for structured light systems. *IEEE J. Sel. Top. Signal Process.* **6**, 411–424 (2012).
183. Jiang, C. F., Li, B. W. & Zhang, S. Pixel-by-pixel absolute phase retrieval using three phase-shifted fringe patterns without markers. *Opt. Lasers Eng.* **91**, 232–241 (2017).
184. Liu, X. R. & Kofman, J. High-frequency background modulation fringe patterns based on a fringe-wavelength geometry-constraint model for 3D surface-shape measurement. *Opt. Express* **25**, 16618–16628 (2017).
185. Tao, T. Y. et al. High-precision real-time 3D shape measurement using a bi-frequency scheme and multi-view system. *Appl. Opt.* **56**, 3646–3653 (2017).
186. Tao, T. Y. et al. High-speed real-time 3D shape measurement based on adaptive depth constraint. *Opt. Express* **26**, 22440–22456 (2018).
187. Cai, Z. W. et al. Light-field-based absolute phase unwrapping. *Opt. Lett.* **43**, 5717–5720 (2018).
188. Pan, B., Xie, H. M. & Wang, Z. Y. Equivalence of digital image correlation criteria for pattern matching. *Appl. Opt.* **49**, 5501–5509 (2010).
189. Gruen, A. W. Adaptive least squares correlation: a powerful image matching technique. *J. Photogramm. Remote Sens. Cartogr.* **14**, 175–187 (1985).
190. Altunbasak, Y., Mersereau, R. M. & Patti, A. J. A fast parametric motion estimation algorithm with illumination and lens distortion correction. *IEEE Trans. Image Process.* **12**, 395–408 (2003).
191. Gutman, S. On optimal guidance for homing missiles. *J. Guidance Control* **2**, 296–300 (1979).
192. Zabih, R. & Woodfill, J. Non-parametric local transforms for computing visual correspondence. In *Proceedings of the 3rd European Conference on Computer Vision*. 151–158 (Springer, Stockholm, 1994).
193. Bhat, D. N. & Nayar, S. K. Ordinal measures for image correspondence. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 415–423 (1998).
194. Sara, R. & Bajcsy, R. On occluding contour artifacts in stereo vision. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 852–857 (IEEE, San Juan, PR, 1997).
195. Sutton, M. A. et al. Effects of subpixel image restoration on digital correlation error estimates. *Optical Eng.* **27**, 271070 (1988).
196. Zhang, D., Zhang, X. & Cheng, G. Compression strain measurement by digital speckle correlation. *Exp. Mech.* **39**, 62–65 (1999).
197. Hung, P. C. & Voloshin, A. In-plane strain measurement by digital image correlation. *J. Braz. Soc. Mech. Sci. Eng.* **25**, 215–221 (2003).



198. Davis, C. Q. & Freeman, D. M. Statistics of subpixel registration algorithms based on spatiotemporal gradients or block matching. *Optical Eng.* **37**, 1290–1298 (1998).
199. Zhou, P. & Goodson, K. E. Subpixel displacement and deformation gradient measurement using digital image/speckle correlation. *Optical Eng.* **40**, 1613–1620 (2001).
200. Press, W. H. et al. *Numerical Recipes in Fortran 77: Volume 1, Volume 1 of Fortran Numerical Recipes: The Art of Scientific Computing*, 2nd edn. (Cambridge University Press, 1992).
201. Chapra, S. C., Canale, R. P. *Numerical Methods for Engineers* (McGraw-Hill Higher Education, 2011).
202. Baker, S. & Matthews, I. Equivalence and efficiency of image alignment algorithms. In *Proceedings of 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1 (IEEE, Kauai, HI, 2001).
203. Baker, S. & Matthews, I. Lucas-Kanade 20 years on: a unifying framework. *Int. J. Computer Vis.* **56**, 221–255 (2004).
204. Pan, B., Li, K. & Tong, W. Fast, robust and accurate digital image correlation calculation without redundant computations. *Exp. Mech.* **53**, 1277–1289 (2013).
205. Pan, B. & Li, K. A fast digital image correlation method for deformation measurement. *Opt. Lasers Eng.* **49**, 841–847 (2011).
206. Zhang, L. Q. et al. High accuracy digital image correlation powered by GPU-based parallel computing. *Opt. Lasers Eng.* **69**, 7–12 (2015).
207. Konolige, K. Small vision systems: hardware and implementation. in *Robotics Research: The Eighth International Symposium* (eds Shirai, Y. & Hirose, S.) 203–212 (Springer, 1998).
208. Hirschmüller, H., Innocent, P. R. & Garibaldi, J. Real-time correlation-based stereo vision with reduced border errors. *Int. J. Computer Vis.* **47**, 229–246 (2002).
209. Scharstein, D. & Szeliski, R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Computer Vis.* **47**, 7–42 (2002).
210. Hirschmuller, H. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**, 328–341 (2008).
211. Boykov, Y., Veksler, O. & Zabih, R. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**, 1222–1239 (2001).
212. Hong, C. K., Ryu, H. S. & Lim, H. C. Least-squares fitting of the phase map obtained in phase-shifting electronic speckle pattern interferometry. *Opt. Lett.* **20**, 931–933 (1995).
213. Aebischer, H. A. & Waldner, S. A simple and effective method for filtering speckle-interferometric phase fringe patterns. *Opt. Commun.* **162**, 205–210 (1999).
214. Yatabe, K. & Oikawa, Y. Convex optimization-based windowed Fourier filtering with multiple windows for wrapped-phase denoising. *Appl. Opt.* **55**, 4632–4641 (2016).
215. Huang, H. Y. H. et al. Path-independent phase unwrapping using phase gradient and total-variation (TV) denoising. *Opt. Express* **20**, 14075–14089 (2012).
216. Chen, R. P. et al. Interferometric phase denoising by pyramid nonlocal means filter. *IEEE Geosci. Remote Sens. Lett.* **10**, 826–830 (2013).
217. Langehanenberg, P. et al. Autofocusing in digital holographic phase contrast microscopy on pure phase objects for live cell imaging. *Appl. Opt.* **47**, D176–D182 (2008).
218. Gao, P. et al. Autofocusing of digital holographic microscopy based on off-axis illuminations. *Opt. Lett.* **37**, 3630–3632 (2012).
219. Dubois, F. et al. Focus plane detection criteria in digital holography microscopy by amplitude analysis. *Opt. Express* **14**, 5895–5908 (2006).
220. Pan, B. et al. Phase error analysis and compensation for non-sinusoidal waveforms in phase-shifting digital fringe projection profilometry. *Opt. Lett.* **34**, 416–418 (2009).
221. Feng, S. J. et al. Robust dynamic 3-D measurements with motion-compensated phase-shifting profilometry. *Opt. Lasers Eng.* **103**, 127–138 (2018).
222. Ferraro, P. et al. Compensation of the inherent wave front curvature in digital holographic coherent microscopy for quantitative phase-contrast imaging. *Appl. Opt.* **42**, 1938–1946 (2003).
223. Di, J. L. et al. Phase aberration compensation of digital holographic microscopy based on least squares surface fitting. *Opt. Commun.* **282**, 3873–3877 (2009).
224. Miccio, L. et al. Direct full compensation of the aberrations in quantitative phase microscopy of thin objects by a single digital hologram. *Appl. Phys. Lett.* **90**, 041104 (2007).
225. Colomb, T. et al. Total aberrations compensation in digital holographic microscopy with a reference conjugated hologram. *Opt. Express* **14**, 4300–4306 (2006).
226. Zuo, C. et al. Phase aberration compensation in digital holographic microscopy based on principal component analysis. *Opt. Lett.* **38**, 1724–1726 (2013).
227. Martínez, A. et al. Analysis of optical configurations for ESPI. *Opt. Lasers Eng.* **46**, 48–54 (2008).
228. Wang, Y. J. & Zhang, S. Optimal fringe angle selection for digital fringe projection technique. *Appl. Opt.* **52**, 7094–7098 (2013).
229. Michie, D., Spiegelhalter, D. J. & Taylor, C. C. Machine learning. *Neural Stat. Classification. Neural Stat. Classif.* **13**, 1–298 (1994).
230. Zhang, X. D. Machine learning. in *A Matrix Algebra Approach to Artificial Intelligence* (ed. Zhang, X. D.) 223–440 (Springer, 2020).
231. Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**, 386–408 (1958).
232. Nair, V. & Hinton, G. E. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*. 807–814 (ACM, Haifa, 2010).
233. Gardner, M. W. & Dorling, S. R. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmos. Environ.* **32**, 2627–2636 (1998).
234. Sussillo, D. Random walks: training very deep nonlinear feed-forward networks with smart initialization. Preprint at <https://arxiv.org/abs/1412.6558v2> (2014).
235. Kraus, M., Feuerriegel, S. & Oztekin, A. Deep learning in business analytics and operations research: models, applications and managerial implications. *Eur. J. Operational Res.* **281**, 628–641 (2020).
236. Zhang, Z. L. & Sabuncu, M. R. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 8792–8802 (ACM, Montréal, 2018).
237. Korhonen, J. & You, J. Y. Peak signal-to-noise ratio revisited: is simple beautiful? In *Proceedings of the 4th International Workshop on Quality of Multimedia Experience*. 37–38 (IEEE, Melbourne, VIC, 2012).
238. Girshick, R. Fast R-CNN. In *Proceedings of 2015 IEEE International Conference on Computer Vision*. 1440–1448 (IEEE, Santiago, 2015).
239. Wang, Z. et al. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
240. Wang, Z. & Bovik, A. C. Mean squared error: love it or leave it? A new look at signal fidelity measures. *IEEE Signal Process. Mag.* **26**, 98–117 (2009).
241. Wang, J. J. et al. Deep learning for smart manufacturing: methods and applications. *J. Manuf. Syst.* **48**, 144–156 (2018).
242. Kingma, D. P. et al. Semi-supervised learning with deep generative models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*. 3581–3589 (ACM, Montreal, 2014).
243. Hinton, G. E. et al. The “wake-sleep” algorithm for unsupervised neural networks. *Science* **268**, 1158–1161 (1995).
244. Bengio, Y. et al. Deep generative stochastic networks trainable by backprop. In *Proceedings of the 31th International Conference on Machine Learning*. 226–234 (JMLR, Beijing, 2014).
245. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophysics* **5**, 115–133 (1943).
246. Minsky, M. & Papert, S. A. *Perceptrons: an Introduction to Computational Geometry* (The MIT Press, 1969).
247. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
248. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–154 (1962).
249. LeCun, Y. et al. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1**, 541–551 (1989).
250. Hochreiter, S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int. J. Uncertain., Fuzziness Knowl.-Based Syst.* **6**, 107–116 (1998).
251. Hinton, G. E., Osindero, S. & Teh, Y. W. A fast learning algorithm for deep belief nets. *Neural Comput.* **18**, 1527–1554 (2006).
252. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
253. Hinton, G. E. & Sejnowski, T. J. Learning and relearning in Boltzmann machines. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations* (eds Rumelhart, D. E. & McClelland, J. L.) (MIT Press, 1986) 282–317.

254. Smolensky, P. Information processing in dynamical systems: foundations of harmony theory. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations* (eds Rumelhart, D. E. & McClelland, J. L.) (MIT Press, 1986) 194–281.
255. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*. 1097–1105 (ACM, Lake Tahoe, Nevada, 2012).
256. LeCun, Y. et al. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
257. Hinton, G. E. et al. Improving neural networks by preventing co-adaptation of feature detectors. Preprint at <https://arxiv.org/abs/1207.0580> (2012).
258. Windhorst, U. On the role of recurrent inhibitory feedback in motor control. *Prog. Neurobiol.* **49**, 517–587 (1996).
259. Elman, J. L. Finding structure in time. *Cogn. Sci.* **14**, 179–211 (1990).
260. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
261. Zhou, J. et al. Graph neural networks: a review of methods and applications. *AI Open* **1**, 57–81 (2020).
262. Xu, K. et al. How powerful are graph neural networks? In *Proceedings of the 7th International Conference on Learning Representations*. (OpenReview, New Orleans, LA, 2018).
263. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations*. (DBIP, San Diego, CA, 2014).
264. Szegedy, C. et al. Going deeper with convolutions. In *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*. 1–9 (IEEE, Boston, MA, 2015).
265. Girshick, R. et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*. 580–587 (IEEE, Columbus, OH, 2014).
266. Goodfellow, I. J. et al. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*. 2672–2680 (ACM, Montreal, 2014).
267. He, K. M. et al. Deep residual learning for image recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. 770–778 (IEEE, Las Vegas, NV, 2016).
268. Chen, J. X. The evolution of computing: AlphaGo. *Comput. Sci. Eng.* **18**, 4–7 (2016).
269. Ouyang, W. L. et al. DeepID-Net: object detection with deformable part based convolutional neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 1320–1334 (2017).
270. Lin, L. et al. A deep structured model with radius-margin bound for 3D human activity recognition. *Int. J. Computer Vis.* **118**, 256–273 (2016).
271. Doulamis, N. & Voulodimos, A. FAST-MDL: fast adaptive supervised training of multi-layered deep learning models for consistent object tracking and classification. In *Proceedings of 2016 IEEE International Conference on Imaging Systems and Techniques (IST)*. 318–323 (IEEE, Chania, 2016).
272. Toshev, A. & Szegedy, C. DeepPose: human pose estimation via deep neural networks. In *Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition*. 1653–1660 (IEEE, Columbus, OH, 2014).
273. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of 2005 IEEE Conference on Computer Vision and Pattern Recognition*. 3431–3440 (IEEE, Boston, MA, 2015).
274. Chen, Q. F., Xu, J. & Koltun, V. Fast image processing with fully-convolutional networks. In *Proceedings of 2017 IEEE International Conference on Computer Vision*. 2516–2525 (IEEE, Venice, 2017).
275. Dong, C. et al. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**, 295–307 (2015).
276. Wang, Z. H., Chen, J. & Hoi, S. C. H. Deep learning for image super-resolution: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 3365–3387 (2021).
277. Dai, Y. P. et al. SRCNN-based enhanced imaging for low frequency radar. In *2018 Progress in Electromagnetics Research Symposium (PIERS-Toyama)*. 366–370 (IEEE, Toyama, 2018).
278. Li, Y. J. et al. Underwater image high definition display using the multilayer perceptron and color feature-based SRCNN. *IEEE Access* **7**, 83721–83728 (2019).
279. Umehara, K., Ota, J. & Ishida, T. Application of super-resolution convolutional neural network for enhancing image resolution in chest CT. *J. Digital Imaging* **31**, 441–450 (2018).
280. Noh, H., Hong, S. & Han, B. Learning deconvolution network for semantic segmentation. In *Proceedings of 2015 IEEE International Conference on Computer Vision*. 1520–1528 (IEEE, Santiago, 2015).
281. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Proceedings of 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*. 234–241 (Springer, Munich, 2015).
282. Falk, T. et al. U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Methods* **16**, 67–70 (2019).
283. Badrinarayanan, V., Kendall, A. & Cipolla, R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 2481–2495 (2017).
284. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In *Proceedings of the 13th European Conference on Computer Vision*. 818–833 (Springer, Zurich, 2014).
285. Shi, W. Z. et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 1874–1883 (IEEE, Las Vegas, NV, 2016).
286. Bell, J. B. Solutions of ill-posed problems. by A. N. Tikhonov, V. Y. Arsenin. *Math. Comput.* **32**, 1320–1322 (1978).
287. Figueiredo, M. A. T. & Nowak, R. D. A bound optimization approach to wavelet-based image deconvolution. In *IEEE International Conference on Image Processing 2005*. II-782 (IEEE, Genova, Italy, 2005).
288. Mairal, J. et al. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*. 689–696 (ACM, Montreal, Quebec, 2009).
289. Daubechies, I., Defrise, M. & De Mol, C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. Pure Appl. Math.* **57**, 1413–1457 (2004).
290. Boyd, S. et al. *Distributed Optimization and Statistical Learning Via the Alternating Direction Method of Multipliers* (Now Publishers Inc, 2011).
291. Candès, E. J., Romberg, J. K. & Tao, T. Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* **59**, 1207–1223 (2006).
292. Greivenkamp, J. E. Generalized data reduction for heterodyne interferometry. *Optical Eng.* **23**, 234350 (1984).
293. Morgan, C. J. Least-squares estimation in phase-measurement interferometry. *Opt. Lett.* **7**, 368–370 (1982).
294. Osten, W. Optical metrology: from the laboratory to the real world. in *Computational Optical Sensing and Imaging* (ed. George, B. et al.) 2013. JW2B-4 (Optical Society of America, 2013).
295. Van der Jeught, S. & Dirckx, J. J. J. Real-time structured light profilometry: a review. *Opt. Lasers Eng.* **87**, 18–31 (2016).
296. Jeon, W. et al. Speckle noise reduction for digital holographic images using multi-scale convolutional neural networks. *Opt. Lett.* **43**, 4240–4243 (2018).
297. Lin, B. W. et al. Optical fringe patterns filtering based on multi-stage convolutional neural network. *Opt. Lasers Eng.* **126**, 105853 (2020).
298. Reyes-Figueroa, A., Flores, V. H. & Rivera, M. Deep neural network for fringe pattern filtering and normalization. *Appl. Opt.* **60**, 2022–2036 (2021).
299. Vincent, P. et al. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **11**, 3371–3408 (2010).
300. Qian, J. M. et al. Single-shot absolute 3D shape measurement with deep-learning-based color fringe projection profilometry. *Opt. Lett.* **45**, 1842–1845 (2020).
301. Zhang, Z. H., Towers, D. P. & Towers, C. E. Snapshot color fringe projection for absolute three-dimensional metrology of video sequences. *Appl. Opt.* **49**, 5947–5953 (2010).
302. Goy, A. et al. Low photon count phase retrieval using deep learning. *Phys. Rev. Lett.* **121**, 243902 (2018).
303. Yu, H. T. et al. Deep learning-based fringe modulation-enhancing method for accurate fringe projection profilometry. *Opt. Express* **28**, 21692–21703 (2020).
304. Feng, S. J. et al. Micro deep learning profilometry for high-speed 3D surface imaging. *Opt. Lasers Eng.* **121**, 416–427 (2019).
305. Qiao, G. et al. A single-shot phase retrieval method for phase measuring deflectometry based on deep learning. *Opt. Commun.* **476**, 126303 (2020).
306. Niu, H. B. et al. Structural light 3D reconstruction algorithm based on deep learning. In *Proceedings of SPIE 11187, Optoelectronic Imaging and Multimedia Technology VI*. 111871F (SPIE, Hangzhou, 2019).

307. Yang, T. et al. Single-shot phase extraction for fringe projection profilometry using deep convolutional generative adversarial network. *Meas. Sci. Technol.* **32**, 015007 (2020).
308. Zhou, W. W. et al. Fourier transform profilometry based on convolution neural network. In *Proceedings of SPIE 10819, Optical Metrology and Inspection for Industrial Applications V*. 108191M (SPIE, Beijing, 2018).
309. Wang, K. et al. Y-Net: a one-to-two deep learning framework for digital holographic reconstruction. *Opt. Lett.* **44**, 4765–4768 (2019).
310. Wang, K. Q. et al. Y4-Net: a deep learning solution to one-shot dual-wavelength digital holographic reconstruction. *Opt. Lett.* **45**, 4220–4223 (2020).
311. Li, Y. X. et al. Single-shot spatial frequency multiplex fringe pattern for phase unwrapping using deep learning. In *Proceedings of SPIE 11571, Optics Frontier Online 2020: Optics Imaging and Display*. 1157118 (SPIE, Shanghai, 2020).
312. Nguyen, H. et al. Real-time 3D shape measurement using 3LCD projection and deep machine learning. *Appl. Opt.* **58**, 7100–7109 (2019).
313. Zhang, S. & Huang, P. S. High-resolution, real-time three-dimensional shape measurement. *Optical Eng.* **45**, 123601 (2006).
314. Zuo, C. et al. High-speed three-dimensional profilometry for multiple objects with complex shapes. *Opt. Express* **20**, 19493–19510 (2012).
315. Zhang, Q. N. et al. Deep phase shifter for quantitative phase imaging. Preprint at <https://arxiv.org/abs/2003.03027> (2020).
316. Li, Z. P., Li, X. Y. & Liang, R. G. Random two-frame interferometry based on deep learning. *Opt. Express* **28**, 24747–24760 (2020).
317. Zhang, L. et al. High-speed high dynamic range 3D shape measurement based on deep learning. *Opt. Lasers Eng.* **134**, 106245 (2020).
318. Wu, S. J. & Zhang, Y. Z. Gamma correction by using deep learning. In *Proceedings of SPIE 11571, Optics Frontier Online 2020: Optics Imaging and Display*. 115710V (SPIE, Shanghai, 2020).
319. Yang, Y. et al. Phase error compensation based on Tree-Net using deep learning. *Opt. Lasers Eng.* **143**, 106628 (2021).
320. Feng, S. J. et al. Generalized framework for non-sinusoidal fringe analysis using deep learning. *Photonics Res.* **9**, 1084–1098 (2021).
321. Wang, K. Q. et al. One-step robust deep learning phase unwrapping. *Opt. Express* **27**, 15100–15115 (2019).
322. Pritt, M. D. & Shipman, J. S. Least-squares two-dimensional phase unwrapping using FFT's. *IEEE Trans. Geosci. Remote Sens.* **32**, 706–708 (1994).
323. Spoorthi, G., Gorthi, S. & Gorthi, R. K. S. S. PhaseNet: a deep convolutional neural network for two-dimensional phase unwrapping. *IEEE Signal Process. Lett.* **26**, 54–58 (2019).
324. Spoorthi, G. E., Gorthi, R. K. S. S. & Gorthi, S. PhaseNet 2.0: phase unwrapping of noisy data based on deep learning approach. *IEEE Trans. Image Process.* **29**, 4862–4872 (2020).
325. Zhang, J. C. et al. Phase unwrapping in optical metrology via denoised and convolutional segmentation networks. *Opt. Express* **27**, 14903–14912 (2019).
326. Kando, D. et al. Phase extraction from single interferogram including closed-fringe using deep learning. *Appl. Sci.* **9**, 3529 (2019).
327. Li, P. H. et al. Deep learning based method for phase analysis from a single closed fringe pattern. In *Proceedings of 11523, Optical Technology and Measurement for Industrial Applications 2020*. 115230E (SPIE, Yokohama, 2020).
328. Liu, K. & Zhang, Y. Z. Temporal phase unwrapping with a lightweight deep neural network. In *Proceedings of SPIE 11571, Optics Frontier Online 2020: Optics Imaging and Display*. 115710N (SPIE, Shanghai, 2020).
329. Li, J. S. et al. Quantitative phase imaging in dual-wavelength interferometry using a single wavelength illumination and deep learning. *Opt. Express* **28**, 28140–28153 (2020).
330. Yao, P. C., Gai, S. Y. & Da, F. P. Coding-Net: a multi-purpose neural network for fringe projection profilometry. *Opt. Commun.* **489**, 126887 (2021).
331. Yao, P. C. et al. A multi-code 3D measurement technique based on deep learning. *Opt. Lasers Eng.* **143**, 106623 (2021).
332. Qian, J. M. et al. Deep-learning-enabled geometric constraints and phase unwrapping for single-shot absolute 3D shape measurement. *APL Photonics* **5**, 046105 (2020).
333. Yu, H. T. et al. Dynamic 3-D measurement based on fringe-to-fringe transformation using deep learning. *Opt. Express* **28**, 9405–9418 (2020).
334. Žbontar, J. & LeCun, Y. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.* **17**, 2287–2318 (2016).
335. Mei, X. et al. On building an accurate stereo matching system on graphics hardware. In *Proceedings of 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 467–474 (IEEE, Barcelona, 2011).
336. Luo, W. J., Schwing, A. G. & Urtasun, R. Efficient deep learning for stereo matching. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. 5695–5703 (IEEE, Las Vegas, NV, 2016).
337. Yin, W. et al. Composite deep learning framework for absolute 3D shape measurement based on single fringe phase retrieval and speckle correlation. *J. Phys.: Photonics* **2**, 045009 (2020).
338. Hartmann, W. et al. Learned multi-patch similarity. In *Proceedings of 2017 IEEE International Conference on Computer Vision*. 1595–1603 (IEEE, Venice, 2017).
339. Žbontar, J. & LeCun, Y. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*. 1592–1599 (IEEE, Boston, MA, 2015).
340. Zagoruyko, S. & Komodakis, N. Learning to compare image patches via convolutional neural networks. In *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*. 4353–4361 (IEEE, Boston, MA, 2015).
341. Chen, Z. Y. et al. A deep visual correspondence embedding model for stereo matching costs. In *Proceedings of 2015 IEEE International Conference on Computer Vision*. 972–980 (IEEE, Santiago, 2015).
342. Du, Q. C. et al. Stereo-matching network for structured light. *IEEE Signal Process. Lett.* **26**, 164–168 (2019).
343. Yang, G. S. et al. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5510–5519 (IEEE, Long Beach, CA, 2019).
344. Guo, X. Y. et al. Group-wise correlation stereo network. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3273–3282 (IEEE, Long Beach, CA, 2019).
345. Zhou, C. et al. Unsupervised learning of stereo matching. In *Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV)*. 1576–1584 (IEEE, Venice, 2017).
346. Kim, S. et al. Unified confidence estimation networks for robust stereo matching. *IEEE Trans. Image Process.* **28**, 1299–1313 (2019).
347. Pang, J. H. et al. Cascade residual learning: a two-stage convolutional neural network for stereo matching. In *Proceedings of 2017 IEEE International Conference on Computer Vision Workshops*. 887–895 (IEEE, Venice, 2017).
348. Khamis, S. et al. StereoNet: guided hierarchical refinement for real-time edge-aware depth prediction. In *Proceedings of the 15th European Conference on Computer Vision*. 596–613 (Springer, Munich, 2018).
349. Moo Yi, K. et al. Learning to find good correspondences. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2666–2674 (IEEE, Salt Lake City, UT, 2018).
350. Huang, P. H. et al. DeepMVS: learning multi-view stereopsis. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2821–2830 (IEEE, Salt Lake City, UT, 2018).
351. Yao, Y. et al. Recurrent MVSNet for high-resolution multi-view stereo depth inference. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5520–5529 (IEEE, Long Beach, CA, 2019).
352. Chhabra, R. et al. StereoDRNet: dilated residual stereoNet. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11778–11787 (IEEE, Long Beach, CA, 2019).
353. Duggal, S. et al. DeepPruner: learning efficient stereo matching via differentiable patchmatch. In *Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 4383–4392 (IEEE, Seoul, 2019).
354. Kim, S. et al. LAF-Net: locally adaptive fusion networks for stereo confidence estimation. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 205–214 (IEEE, Long Beach, CA, 2019).
355. Yee, K. & Chakrabarti, A. Fast deep stereo with 2D convolutional processing of cost signatures. In *Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision*. 183–191 (IEEE, Snowmass, CO, 2020).
356. Tonioni, A. et al. Real-time self-adaptive deep stereo. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 195–204 (IEEE, Long Beach, CA, 2019).
357. Wang, Y. et al. UnoS: unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8063–8073 (IEEE, Long Beach, CA, 2019).
358. Jie, Z. Q. et al. Left-right comparative recurrent model for stereo matching. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3838–3846 (IEEE, Salt Lake City, UT, 2018).
359. Poggi, M. & Mattoccia, S. Learning from scratch a confidence measure. In *Proceedings of the British Machine Vision Conference 2016*. (BMVC, York, 2016).



360. Yin, W. et al. High-speed 3D shape measurement with the multi-view system using deep learning. In *Proceedings of SPIE 11189, Optical Metrology and Inspection for Industrial Applications VI*. 111890B (SPIE, Hangzhou, 2019).
361. Fanello, S. R. et al. UltraStereo: efficient learning-based matching for active stereo systems. In *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6535–6544 (IEEE, Honolulu, HI, 2017).
362. Montrésor, S. et al. Computational de-noising based on deep learning for phase data in digital holographic interferometry. *APL Photonics* **5**, 030802 (2020).
363. Yan, K. T. et al. Wrapped phase denoising using convolutional neural networks. *Opt. Lasers Eng.* **128**, 105999 (2020).
364. Yan, K. T. et al. Deep learning-based wrapped phase denoising method for application in digital holographic speckle pattern interferometry. *Appl. Sci.* **10**, 4044 (2020).
365. Ren, Z. B., Xu, Z. M. & Lam, E. Y. M. End-to-end deep learning framework for digital holographic reconstruction. *Adv. Photonics* **1**, 016004 (2019).
366. Goodman, J. W. *Introduction to Fourier Optics*, 3rd edn. (Roberts and Company Publishers, 2005).
367. Bioucas-Dias, J. et al. Absolute phase estimation: adaptive local denoising and global unwrapping. *Appl. Opt.* **47**, 5358–5369 (2008).
368. Kreis, T. M., Adams, M. & Jüepfner, W. P. O. Methods of digital holography: a comparison. In *Proceedings of SPIE 3098, Optical Inspection and Micro-measurements II*. 224–233 (SPIE, Munich, 1997).
369. Ren, Z. B., Xu, Z. M. & Lam, E. Y. Learning-based nonparametric autofocusing for digital holography. *Optica* **5**, 337–344 (2018).
370. Lee, J. et al. Autofocusing using deep learning in off-axis digital holography. in *Digital Holography and Three-Dimensional Imaging* (ed. Yoshio, H. et al.) 2018. Dth1C.4 (Optical Society of America, 2018).
371. Shimobaba, T., Kakue, T. & Ito, T. Convolutional neural network-based regression for depth prediction in digital holography. In *Proceedings of the IEEE 27th International Symposium on Industrial Electronics (ISIE)*. 1323–1326 (IEEE, Cairns, QLD, 2018).
372. Jaferzadeh, K. et al. No-search focus prediction at the single cell level in digital holographic imaging with deep convolutional neural network. *Biomed. Opt. Express* **10**, 4276–4289 (2019).
373. Pitkäaho, T., Manninen, A. & Naughton, T. J. Focus prediction in digital holographic microscopy using deep convolutional neural networks. *Appl. Opt.* **58**, A202–A208 (2019).
374. Nguyen, T. et al. Automatic phase aberration compensation for digital holographic microscopy based on deep learning background detection. *Opt. Express* **25**, 15043–15057 (2017).
375. Nguyen, T. et al. Accurate quantitative phase digital holographic microscopy with single-and multiple-wavelength telecentric and nontelecentric configurations. *Appl. Opt.* **55**, 5666–5683 (2016).
376. Lv, S. Z. et al. Projector distortion correction in 3D shape measurement using a structured-light system by deep neural networks. *Opt. Lett.* **45**, 204–207 (2020).
377. Aguénoun, E. et al. Real-time, wide-field and high-quality single snapshot imaging of optical properties with profile correction using deep learning. *Biomed. Opt. Express* **11**, 5701–5716 (2020).
378. Li, Z. W. et al. Complex object 3D measurement based on phase-shifting and a neural network. *Opt. Commun.* **282**, 2699–2706 (2009).
379. Ouellet, J. N. & Hebert, P. A simple operator for very precise estimation of ellipses. In *Proceedings of the 4th Canadian Conference on Computer and Robot Vision (CRV07)*. 21–28 (IEEE, Montreal, QC, 2007).
380. Li, Z. W. et al. Accurate calibration method for a structured light system. *Optical Eng.* **47**, 053604 (2008).
381. Nguyen, H., Wang, Y. Z. & Wang, Z. Y. Single-shot 3D shape reconstruction using structured light and deep convolutional neural networks. *Sensors* **20**, 3718 (2020).
382. Van der Jeught, S. & Dirckx, J. J. J. Deep neural networks for single shot structured light profilometry. *Opt. Express* **27**, 17091–17101 (2019).
383. Van Der Jeught, S., Muylshondt, P. G. G. & Lobato, I. Optimized loss function in deep learning profilometry for improved prediction performance. *J. Phys.: Photonics* **3**, 024014 (2021).
384. Machineni, R. C. et al. End-to-end deep learning-based fringe projection framework for 3D profiling of objects. *Computer Vis. Image Underst.* **199**, 103023 (2020).
385. Zheng, Y. et al. Fringe projection profilometry by conducting deep learning from its digital twin. *Opt. Express* **28**, 36568–36583 (2020).
386. Wang, F. Z., Wang, C. X. & Guan, Q. Z. Single-shot fringe projection profilometry based on deep learning and computer graphics. *Opt. Express* **29**, 8024–8040 (2021).
387. Mayer, N. et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. 4040–4048 (IEEE, Las Vegas, NV, 2016).
388. Menze, M. & Geiger, A. Object scene flow for autonomous vehicles. In *Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*. 3061–3070 (IEEE, Boston, MA, 2015).
389. Kendall, A. et al. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV)*. 66–75 (IEEE, Venice, 2017).
390. Chang, J. R. & Chen, Y. S. Pyramid stereo matching network. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5410–5418 (IEEE, Salt Lake City, UT, 2018).
391. Zhang, F. H. et al. GA-Net: guided aggregation net for end-to-end stereo matching. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 185–194 (IEEE, Long Beach, CA, 2019).
392. Yin, W. et al. Single-shot 3D shape measurement using an end-to-end stereo matching network for speckle projection profilometry. *Opt. Express* **29**, 13388–13407 (2021).
393. Nguyen, H. et al. Three-dimensional shape reconstruction from single-shot speckle image using deep convolutional neural networks. *Opt. Lasers Eng.* **143**, 106639 (2021).
394. Knöbelreiter, P. et al. End-to-end training of hybrid CNN-CRF models for stereo. In *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1456–1465 (IEEE, Honolulu, HI, 2017).
395. Ummerhofer, B. et al. DeMoN: depth and motion network for learning monocular stereo. In *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. 5622–5631 (IEEE, Honolulu, HI, 2017).
396. Yao, Y. et al. MVSNet: depth inference for unstructured multi-view stereo. In *Proceedings of the 15th European Conference on Computer Vision*. 785–801 (Springer, Munich, 2018).
397. Liang, Z. F. et al. Learning for disparity estimation through feature constancy. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2811–2820 (IEEE, Salt Lake City, UT, 2018).
398. Yang, G. R. et al. SegStereo: exploiting semantic information for disparity estimation. In *Proceedings of the 15th European Conference on Computer Vision (ECCV)*. 660–676 (Springer, Munich, 2018).
399. Song, X. et al. EdgeStereo: a context integrated residual pyramid network for stereo matching. In *Proceedings of the 14th Asian Conference on Computer Vision*. 20–35 (Springer, Perth, 2018).
400. Yu, L. D. et al. Deep stereo matching with explicit cost aggregation sub-architecture. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. (AAAI, New Orleans, LA, 2018).
401. Fanello, S. R. et al. HyperDepth: learning depth from structured light without matching. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5441–5450 (IEEE, Las Vegas, NV, 2016).
402. Tulyakov, S., Ivanov, A. & Fleuret, F. Practical deep stereo (PDS): toward applications-friendly deep stereo matching. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 5871–5881 (ACM, Montréal, 2018).
403. Nie, G. Y. et al. Multi-level context ultra-aggregation for stereo matching. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3278–3286 (IEEE, Long Beach, CA, 2019).
404. Zhong, Y. R., Li, H. D. & Dai, Y. C. Open-world stereo video matching with deep RNN. In *Proceedings of the 15th European Conference on Computer Vision (ECCV)*. 101–116 (Springer, Munich, 2018).
405. Tonioni, A. et al. Unsupervised adaptation for deep stereo. In *Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV)*. 1614–1622 (IEEE, Venice, 2017).
406. Tonioni, A. et al. Unsupervised domain adaptation for depth prediction from images. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 2396–2409 (2020).
407. Chen, X. Y. Non-destructive three-dimensional measurement of hand vein based on self-supervised network. *Measurement* **173**, 108621 (2020).
408. Zhang, Y. D. et al. ActiveStereoNet: end-to-end self-supervised learning for active stereo systems. In *Proceedings of the 15th European Conference on Computer Vision (ECCV)*. 802–819 (Springer, Munich, 2018).

409. Tonioni, A. et al. Learning to Adapt for Stereo. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9653–9662 (IEEE, Long Beach, CA, 2019).
410. Boukhache, S. et al. When deep learning meets digital image correlation. *Opt. Lasers Eng.* **136**, 106308 (2021).
411. Min, H. G. et al. Strain measurement during tensile testing using deep learning-based digital image correlation. *Meas. Sci. Technol.* **31**, 015014 (2020).
412. Rezaie, A. et al. Comparison of crack segmentation using digital image correlation measurements and deep learning. *Constr. Build. Mater.* **261**, 120474 (2020).
413. Son, K., Liu, M. Y. & Taguchi, Y. Learning to remove multipath distortions in time-of-flight range images for a robotic arm setup. In *Proceedings of 2016 IEEE International Conference on Robotics and Automation (ICRA)*. 3390–3397 (IEEE, Stockholm, 2016).
414. Marco, J. et al. DeepToF: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. *ACM Trans. Graph.* **36**, 219 (2017).
415. Song, S. & Shim, H. Depth reconstruction of translucent objects from a single time-of-flight camera using deep residual networks. In *Proceedings of the 14th Asian Conference on Computer Vision*. 641–657 (Springer, Perth, 2018).
416. Su, S. C. et al. Deep end-to-end time-of-flight imaging. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6383–6392 (IEEE, Salt Lake City, UT, 2018).
417. Chen, Y. et al. A learning method to optimize depth accuracy and frame rate for Time of Flight camera. *IOP Conf. Ser.: Mater. Sci. Eng.* **563**, 042067 (2019).
418. Chen, Y. et al. Very power efficient neural time-of-flight. In *Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision*. 2246–2255 (IEEE, Snowmass, CO, 2020).
419. Santo, H. et al. Deep photometric stereo network. In *Proceedings of 2017 IEEE International Conference on Computer Vision Workshops*. 501–509 (IEEE, Venice, 2017).
420. Ikehata, S. CNN-PS: CNN-based photometric stereo for general non-convex surfaces. In *Proceedings of the 15th European Conference on Computer Vision (ECCV)*. 3–19 (Springer, Munich, 2018).
421. Taniai, T. & Maehara, T. Neural inverse rendering for general reflectance photometric stereo. In *Proceedings of the 35th International Conference on Machine Learning*. 4864–4873 (PMLR, Stockholm, 2018).
422. Xu, Z. X. et al. Deep image-based relighting from optimal sparse samples. *ACM Trans. Graph.* **37**, 126 (2018).
423. Li, J. X. et al. Learning to minify photometric stereo. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7560–7568 (IEEE, Long Beach, CA, 2019).
424. Chen, G. Y. et al. Self-calibrating deep photometric stereo networks. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8731–8739 (IEEE, Long Beach, CA, 2019).
425. Sang, L., Haefner, B. & Cremers, D. Inferring super-resolution depth from a moving light-source enhanced RGB-D sensor: a variational approach. In *Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision*. 1–10 (IEEE, Snowmass, CO, 2020).
426. Nishizaki, Y. et al. Deep learning wavefront sensing. *Opt. Express* **27**, 240–251 (2019).
427. Hu, L. J. et al. Learning-based Shack-Hartmann wavefront sensor for high-order aberration detection. *Opt. Express* **27**, 33504–33517 (2019).
428. DuBose, T. B., Gardner, D. F. & Watnik, A. T. Intensity-enhanced deep network wavefront reconstruction in Shack-Hartmann sensors. *Opt. Lett.* **45**, 1699–1702 (2020).
429. Hu, L. J. et al. Deep learning assisted Shack-Hartmann wavefront sensor for direct wavefront detection. *Opt. Lett.* **45**, 3741–3744 (2020).
430. Rodin, I. A. et al. Recognition of wavefront aberrations types corresponding to single Zernike functions from the pattern of the point spread function in the focal plane using neural networks. *Computer Opt.* **44**, 923–930 (2020).
431. Moran, O. et al. Deep, complex, invertible networks for inversion of transmission effects in multimode optical fibres. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 3284–3295 (ACM, Montréal, 2018).
432. Borhani, N. et al. Learning to see through multimode fibers. *Optica* **5**, 960–966 (2018).
433. Fan, P. F., Zhao, T. R. & Su, L. Deep learning the high variability and randomness inside multimode fibers. *Opt. Express* **27**, 20241–20258 (2019).
434. Caramazza, P. et al. Transmission of natural scene images through a multimode fibre. *Nat. Commun.* **10**, 2029 (2019).
435. Fan, P. F. et al. Speckle reconstruction with corruption through multimode fibers using deep learning. In *Proceedings of 2020 Conference on Lasers and Electro-Optics (CLEO)*. 1–2 (IEEE, San Jose, CA, 2020).
436. Sun, C. et al. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of 2017 IEEE International Conference on Computer Vision*. 843–852 (IEEE, Venice, 2017).
437. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
438. Sung, F. et al. Learning to compare: relation network for few-shot learning. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1199–1208 (IEEE, Salt Lake City, UT, 2018).
439. Goh, G. B. et al. Using rule-based labels for weak supervised learning: a ChemNet for transferable chemical property prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 302–310 (ACM, London, 2018).
440. Hutter, F., Kotthoff, L. & Vanschoren, J. *Automated Machine Learning: Methods, Systems, Challenges* (Springer, 2019).
441. Neyshabur, B. et al. Exploring generalization in deep learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 5949–5958 (ACM, Long Beach, CA, 2017).
442. Ledig, C. et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition*. 105–114 (IEEE, Honolulu, HI, 2017).
443. Qian, J. M. et al. High-resolution real-time 360° 3D surface defect inspection with fringe projection profilometry. *Opt. Lasers Eng.* **137**, 106382 (2021).
444. Jing, L. L. & Tian, Y. L. Self-supervised visual feature learning with deep neural networks: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 4037–4058 (2021).
445. Baker, B. et al. Designing neural network architectures using reinforcement learning. In *Proceedings of the 5th International Conference on Learning Representations*. (OpenReview, Toulon, 2017).
446. Bisong, E. Google AutoML: cloud vision. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners* (ed. Bisong, E.) 581–598 (Springer, 2019).
447. Barnes, J. *Microsoft Azure Essentials Azure Machine Learning* (Microsoft Press, 2015).
448. Feurer, M. et al. Efficient and robust automated machine learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*. 2755–2763 (ACM, Montreal, 2015).
449. Wang, F. et al. Phase imaging with an untrained neural network. *Light: Sci. Appl.* **9**, 77 (2020).
450. Abdar, M. et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf. Fusion* **76**, 243–297 (2021).
451. Korattikara, A. et al. Bayesian dark knowledge. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*. (ACM, Montreal, 2015).
452. Shekhovtsov, A. & Flach, B. Feed-forward propagation in probabilistic neural networks with categorical and max layers. In *Proceedings of the 7th International Conference on Learning Representations*. (OpenReview, New Orleans, LA, 2019).
453. Feng, S. J. et al. Deep-learning-based fringe-pattern analysis with uncertainty estimation. *Optica* **8**, 1507–1510 (2021).
454. Chakrabarti, A. Learning sensor multiplexing design through back-propagation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 3089–3097 (ACM, Barcelona, 2016).
455. Horstmeyer, R. et al. Convolutional neural networks that teach microscopes how to image. Preprint at <https://arxiv.org/abs/1709.07223> (2017).
456. Kellman, M. R. et al. Physics-based learned design: optimized coded-illumination for quantitative phase imaging. *IEEE Trans. Comput. Imaging* **5**, 344–353 (2019).
457. Muthumbi, A. et al. Learned sensing: jointly optimized microscope hardware for accurate image classification. *Biomed. Opt. Express* **10**, 6351–6369 (2019).
458. Kim, Y. et al. Evaluation for snowfall depth forecasting using neural network and multiple regression models. *J. Korean Soc. Hazard Mitig.* **13**, 269–280 (2013).
459. Geiger, A. et al. Vision meets robotics: the KITTI dataset. *Int. J. Robot. Res.* **32**, 1231–1237 (2013).
460. Hirschmuller, H. & Scharstein, D. Evaluation of cost functions for stereo matching. In *Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8 (IEEE, Minneapolis, MN, 2007).

·特邀综述·

# 深度学习下的计算成像:现状、挑战与未来

左超<sup>1,2</sup>, 冯世杰<sup>1,2</sup>, 张翔宇<sup>1,2</sup>, 韩静<sup>2</sup>, 陈钱<sup>2\*</sup>

<sup>1</sup>南京理工大学电子工程与光电技术学院, 智能计算成像实验室(SCILab), 江苏 南京 210094;

<sup>2</sup>南京理工大学江苏省光谱成像与智能感知重点实验室, 江苏 南京 210094

**摘要** 近年来,光学成像技术已经由传统的强度、彩色成像发展进入计算光学成像时代。计算光学成像基于几何光学、波动光学等理论对场景目标经光学系统成像再到探测器采样这一完整图像生成过程建立精确的正向数学模型,再求解该正向成像模型所对应的“逆问题”,以计算重构的方式来获得场景目标的高质量图像或者传统技术无法直接获得的相位、光谱、偏振、光场、相干度、折射率、三维形貌等高维度物理信息。然而,计算成像系统的实际成像性能也同样极大程度地受限于“正向数学模型的准确性”以及“逆向重构算法的可靠性”,实际成像物理过程的不可预见性与高维病态逆问题求解的复杂性已成为这一领域进一步发展的瓶颈问题。近年来,人工智能与深度学习技术的飞跃式发展为计算光学成像技术开启了一扇全新的大门。不同于传统计算成像方法所依赖的物理驱动,深度学习下的计算成像是一类由数据驱动的方法,它不但解决了许多过去计算成像领域难以解决的难题,还在信息获取能力、成像的功能、核心性能指标(如成像空间分辨率、时间分辨率、灵敏度等)上都获得了显著提升。基于此,首先概括性介绍深度学习技术在计算光学成像领域的研究进展与最新成果,然后分析了当前深度学习技术在计算光学成像领域面临的主要问题与挑战,最后展望了该领域未来的发展方向与可能的研究方向。

**关键词** 成像系统; 计算成像; 深度学习; 光学成像; 光信息处理

中图分类号 O436

文献标志码 A

doi: 10.3788/AOS202040.0111003

## Deep Learning Based Computational Imaging: Status, Challenges, and Future

Zuo Chao<sup>1,2</sup>, Feng Shijie<sup>1,2</sup>, Zhang Xiangyu<sup>1,2</sup>, Han Jing<sup>2</sup>, Qian Chen<sup>2\*</sup>

<sup>1</sup>Smart Computational Imaging Laboratory (SCILab), School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China;

<sup>2</sup>Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China

**Abstract** In recent years, optical imaging techniques have entered into the era of computational optical imaging from the traditional intensity and color imaging. Computational optical imaging, which is based on geometric optics, wave optics, and other theoretical foundations, establishes an accurate forward mathematical model for the whole image formation process of the scene imaged through the optical system and then sampled by the digital detector. Then, the high-quality reconstruction of the image and other high dimensional information, such as phase, spectrum, polarization, light field, coherence, refractive index, and three-dimension profile, which cannot be directly accessed using traditional methods, can be obtained through computational reconstruction method. However, the actual imaging performance of the computational imaging system is also limited by the “accuracy of the forward mathematical model” and “the reliability of inverse reconstruction algorithm”. Besides, the unpredictability of real physical imaging process and the complexity of solving high dimensional ill-posed inverse problems have become the bottleneck of further development of this field. In recent years, the rapid development of artificial intelligence and deep learning for the technology opens a new door for computational optical imaging technology. Unlike “physical driven” model that traditional computational imaging method is based on,

收稿日期: 2019-11-06; 修回日期: 2019-11-22; 录用日期: 2019-12-05

基金项目: 国家自然科学基金(61722506, 61705105, 11574152)、总装“十三五”装备预研项目(30102070102)、总装“十三五”领域基金(61404150202)、国防科技项目基金(0106173)、江苏省杰出青年基金(BK20170034)、江苏省重点研发计划(BE2017162)、江苏省“333工程”科研项目资助计划(BRA2016407)、江苏省光谱成像与智能感知重点实验室开放基金(3091801410411)

\* E-mail: chenqian@njjust.edu.cn



computational imaging based on deep learning is a kind of “data-driven” method, which not only solves many problems considered quite challenge to be solved in this field, but also achieves remarkable improvement in information acquisition ability, imaging functions, and key performance indexes of imaging system, such as spatial resolution, temporal resolution, and detection sensitivity. This review first briefly introduces the current status and the latest progress of deep learning technology in the field of computational optical imaging. Then, the main problems and challenges faced by the current deep learning method in computational optical imaging field are discussed. Finally, the future developments and possible research directions of this field are prospected.

**Key words** imaging systems; computational imaging; deep learning; optical imaging; optical information processing

**OCIS codes** 110.0180; 100.5070; 180.6900

## 1 引 言

视觉是人类获得客观世界信息的主要途径,而人眼受限于视觉性能,在时间、空间、灵敏度等方面均存在局限性。光学成像技术由此应运而生,其利用各种光学成像系统,如显微镜、望远镜等,实现光信息的可视化,同时延伸扩展人眼的视觉特性。然而,一方面,传统光学成像系统因受强度成像机理、探测器技术水平、光学系统设计、成像衍射极限等因素制约,在空间分辨、时间分辨、光谱分辨、信息维度与探测灵敏度等方面仍存在一定局限性,难以满足人们对成像系统功能与性能的进一步需求,以及军民领域日益增长的高分辨、高灵敏度和多维高速成像的应用需求。采用传统光学成像系统的设计思路想要获得成像性能的少量提升,通常意味着硬件成本的急剧增加,甚至难以实现工程化应用。另一方面,光探测器规模尺寸、像元大小、响应灵敏度等均已接近物理极限,很难满足这些极具挑战性的需求。

随着成像电子学的发展,计算机数据处理能力的增强,光场调控、孔径编码、压缩感知、全息成像等光电信息处理技术取得了重大进展;此外,经过成千上万年的发展自然界已经演化出多类能够满足不同生存需求的生物视觉系统,从生物视觉系统中获得灵感无疑可以给新一代光学成像技术的发展带来有益的启示。在此背景下,20世纪90年代中期,光学成像界和图像处理界的许多研究人员不约而同地探索出了一种新型成像模式,即图像形成不再仅仅依赖于光学物理器件,还依赖于前端光学和后端探测信号处理的联合设计。这种技术就是现在广为人知的“计算成像”(Computational Imaging)技术<sup>[1]</sup>,它将光学调控与信息处理有机结合,为突破上述传统成像系统中的诸多限制性因素提供了新手段与新思路。

计算光学成像是一种通过联合优化光学系统和信号处理来实现特定成像功能与特性的新兴研究领

域。其建立在几何光学、波动光学,甚至光量子模型的基础上,采用照明与光学系统调制等方式,建立目标场景与观测图像之间的变换或调制模型,然后利用逆问题求解等数学手段,通过计算反演来进行成像。这种计算成像方法实质上就是在场景和图像之间建立某种特定的联系,这种联系可以是线性的也可以是非线性的。它突破了传统成像技术点对点一一对应的强度直接采样形式,采用了更加灵活的非直接的采样形式,更能充分发挥成像系统中各组件的特点与性能。这种灵活的设计模式可以改变光学测量的性质以获得所需的结果,并平衡物理域和计算域之间图像生成和信息提取所依赖的资源。基于信息论的概念,计算光学成像设计师不仅可以借助于传统光学设计的优势,还可以充分利用物理光学在光信号处理中的潜力来设计成像系统。这种新型的成像方式将有望改变成像系统获取信息的方式,提升其获取信息的能力,增强资源利用,赋予其诸多传统光学成像技术难以获得甚至无法获得的革命性的优势;例如,突破探测器制造工艺、工作条件、功耗成本等因素的限制,有效提高成像质量(信噪比、对比度、动态范围),简化系统(无透镜、小体积、低成本),突破光学系统与图像采集设备的分辨率限制(超像素分辨、超衍射极限),并使其功能(相位、光谱、偏振、光场、相干度、折射率、三维形貌、景深开拓、模糊复原、数字重聚焦、改变观测视角)、性能(空间分辨、时间分辨、光谱分辨、信息维度与探测灵敏度)、可靠性、可维护性等获得显著提高,有助于实现成像设备的高性能、微型化、智能化。

现如今,计算光学成像已发展为一门集几何光学、信息光学、计算光学、计算机视觉、现代信号处理等理论于一体的新兴交叉技术研究领域,成为光学成像领域的一大国际研究重点和热点。然而,隐藏在计算成像华丽外衣之下的是其所必须付出的额外成本与代价:用于进行非传统测量的物理实体器件相关的成本、多次测量产生的时间成本、数据量以及

物理模型和校准对处理性能的影响。更重要的是,计算成像技术的实际成像性能极大程度地受限于“正向数学模型的准确性”以及“逆向重构算法的可靠性”,实际物理成像过程中的不可预见性与高维病态逆问题求解的复杂性已成为这一领域进一步发展亟需解决的瓶颈问题。

近几年,DeepMind 公司研制的人工智能机器人 AlphaGo 战胜顶尖围棋棋手李世石<sup>[2]</sup>、先进图像分类算法在具有挑战性的数据集 ImageNet 上的正确率超过人类<sup>[3]</sup>等令人振奋消息一个接一个地传来,人工智能已经成为我们身边一个耳熟能详的词汇,国际上也开始迎来这一技术的研究热潮。当下谈到人工智能,“机器学习”、“深度学习”和“神经网络”便是经常浮现在人们脑海里的高频词汇。借助于数学中集合的概念,它们之间的关系可以理解为一种包含关系,也就是“机器学习”包含“深度学习”,“深度学习”包含“神经网络”。深度学习已经成为目前最为热门的一种机器学习方案。深度学习这一名称中的“深度”一词表示其使用的神经网络结构多于四层。一般而言,随着神经网络层数的增加,神经网络的性能会更强,学习的效果也会更佳。

互联网技术的蓬勃发展指引着大数据时代的来临,以数据推动的深度学习技术无疑是大数据时代的算法利器。相比于传统的机器学习技术:首先,深度学习技术可利用不断增多的数据不断提升其性能,而传统机器学习技术无法做到这一点;其次,有别于传统方法需要手动提取特征,深度学习技术是一项全自动的技术,它可以从海量数据中直接抽取特征,并且,对于不同的任务,不再需要设计独特的特征提取器,所有工作都可由深度学习自动完成。这是智能机器逐渐代替人工操作的一个显著体现,因此深度学习技术已成为大数据时代的一项热点技术,无论学术界还是工业界都对这项技术产生了浓厚的兴趣。特别是在计算机视觉领域,深度学习作为近年来兴起的一种“数据驱动”的技术,在图像分类、物体检测及识别等诸多应用上均取得了巨大成功。

自 2017 年初,深度学习技术逐渐走入计算成像领域研究者的视野,并在短短的两三年内已在数字全息成像<sup>[4-9]</sup>、傅里叶叠层成像技术<sup>[10-13]</sup>、鬼成像/单像素成像<sup>[14-16]</sup>、超分辨显微成像技术<sup>[17-22]</sup>、光学相干层析成像(OCT)<sup>[23-27]</sup>、散射介质成像<sup>[28-32]</sup>、极弱光成像<sup>[33-34]</sup>、跨模态染色成像<sup>[35-36]</sup>、光栅条纹分析<sup>[37-39]</sup>与快速三维成像<sup>[40-42]</sup>等成像体制上得以成

功应用,取得了一系列令人瞩目的开创性研究成果。令人欣喜的是,对比传统物理模型驱动的计算成像技术,样本数据驱动的深度学习方法下的计算成像技术发生了思想观念上的根本变革,它不但解决了许多过去计算成像领域难以解决的难题,还在信息获取能力、成像的功能、核心性能指标(如成像空间分辨率、时间分辨率、灵敏度等)上获得了显著提升。如今,以深度学习为主题的计算成像相关方面的论文喷井而出,呈指数式增长趋势。

在此背景下,本文概括性地介绍深度学习技术在计算光学成像领域的研究现状与最新进展。简要讨论计算成像技术与深度学习技术的基本概念,并按照深度学习技术的“目的与动机”或者说“深度学习技术为传统计算成像技术带来了哪些新的要素”进行细分,对现有深度学习计算成像技术的研究现状及其典型应用进行概述。值得注意的是,深度学习是一把“双刃剑”,它给计算成像领域研究带来了惊喜的同时也引入了一系列亟待解决的问题。本文分析了当前深度学习技术在计算光学成像领域面临的主要问题与挑战,这亦是本文重要的组成部分。最后,对深度学习在计算成像领域未来的发展方向与可能的研究方向进行讨论并展望,并给出了总结性评论。

## 2 深度学习下的计算成像:现状

一个典型的光学成像系统主要由光源、光学镜头组、光探测器三部分组成。其通过将三维场景中目标发出的光线聚焦在光探测器上进行“点对点”成像。然而这种“所见即所得”的成像方式因其单视角、平面投影等因素的限制,导致高维度场景信息存在缺失。除此之外,日益复杂庞大的光学成像系统也限制着其应用场景。为了解决传统光学成像系统所面临的问题,计算成像技术应运而生,其采用“先调制,再拍摄,最后解调”的成像方式。将光学系统(照明、光学器件、光探测器)与数字图像处理算法作为一个整体考虑,并在设计时一同进行综合优化,前端成像元件与后端数据处理二者相辅相成,构成一种“混合光学-数字计算成像系统”,如图 1 所示。不同于传统光学成像的“所见即所得”,计算成像建立在几何光学、波动光学,甚至光量子模型的基础上,采用照明与光学系统调制等方式,建立目标场景与观测图像之间的变换或调制模型,然后利用逆问题求解等数学手段,通过计算反演来进行成像,以获得场景目标的高质量图像与高维度物理信息。

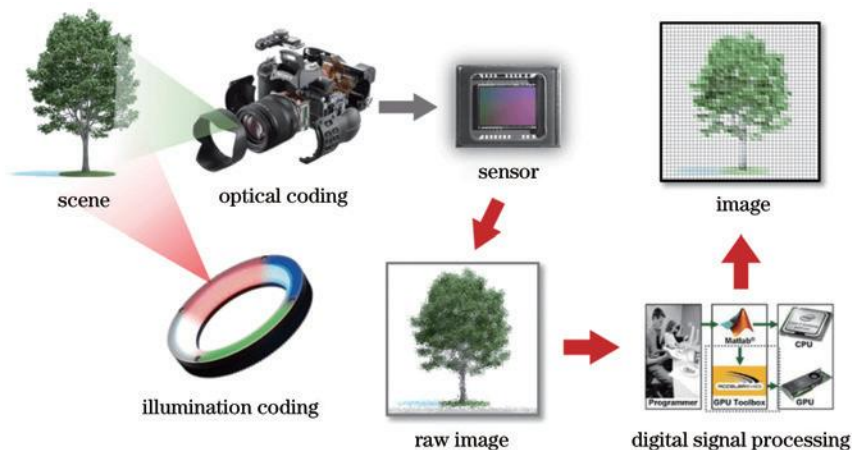


图 1 计算光学成像系统的成像过程

Fig. 1 Imaging process of computational optical imaging system

俗话说“天下没有免费的午餐”，任何事物的存在都具有两面性，计算成像技术亦是如此。当设计计算成像系统时，必须权衡计算成像和传统成像相关的成本代价与预期的改进效用。计算成像技术所能带来的功能与性能上的提升往往是以复杂昂贵的系统硬件、大量额外的数据采集、复杂耗时的算法处理等为代价而换取的。更重要的是，计算光学成像华丽的外衣下还掩盖了其所依赖的两大关键问题：

第一，如何准确地建立场景与图像之间的正向模型，并使得采样数据包含场景中所感兴趣的物理信息，这是最关键的问题；

第二，如何利用逆问题优化求解来重构图像，这也是个核心问题。它既要与非直接采样的自身特点相匹配以建立优化模型，又要在重构过程中保证重构图像的准确度。

然而通过已有的实验结果可以发现，在某些领域（特别是间接成像），目前通过计算成像算法重构的图像和基于传统透镜直接拍摄到的图像在成像质量和保真度上仍然有一些差距。其根本原因在于计算成像要求算法中采用的光学系统的数学模型能够真实全面地反映实际的物理成像过程，如果该模型不能真实地反映光学系统的复杂性，计算成像将有可能得不到理想的成像结果，其优势也可能会完全丧失。例如，如果数学模型对光学成像过程中的光的波动性质、像差或系统对温度变化的敏感度的客观参数进行忽略或者简化，则很可能会出现这种情况。此外在数据采集的过程中还会受到各种环境不确定因素的影响（噪声、振动等），从而会导致所建立的成像模型并不准确。即使完全知道这些影响因素的存在，设计人员仍然会面临着这样一个难题：简单

化的模型可能无法产生精确有效的结果，而更加真实的模型可能需要大量的系统参量、很高的处理负载和很长的处理时间。实际考虑因素包括正向模型成立的条件和逆向重构算法的复杂程度、测量过程对噪声及环境扰动的敏感性，以及算法后处理引入的伪影(Artifacts)水平等。

深度学习作为近年来兴起的一种“数据驱动”的技术，其在图像分类、物体检测及识别乃至“看图作文(Image Captioning)”等诸多计算机视觉任务上均取得了巨大成功。并且由于其其在“黑盒子般的盲建模”与“高维非线性特征拟合”方面的卓越表现，深度学习自 2017 年初也逐渐得到计算成像领域研究者的广泛关注，并在短短的两三年内取得了一系列令人瞩目的开创性研究成果。深度学习的成功不仅仅带来了人工智能相关技术的快速进步，还解决了计算成像领域许多过去被认为是难以解决甚至无法解决的难题，更重要的是它给该领域带来了思想观念上的根本变革：

#### 1) 从“物理模型驱动”到“样本数据驱动”

在深度学习兴起之前，成像物理模型和经验驱动主宰了计算成像领域多年。一个典型的计算光学显微成像系统由照明、样品、成像系统、探测器四部分构成。照明光与样品发生作用后，成为其本质信息（如吸收、相位、光谱、三维、折射率等）的载体，通过对照明与成像系统进行光学调控使物体的本质信息转化为光强信号并由探测器离散采集，最后通过相应的重构算法对样品本质信息进行反演，获得样品的图像或其他所感兴趣的高维物理信息。为了对整个成像过程进行数学建模，通常需要基于标量衍射理论或部分相干理论对照明光产生与自由传播以



及与被测物体相互作用进行建模。例如:部分相干光场需要利用交叉谱密度/互强度或者相空间光学理论中的维格纳函数来对其进行表征;在空域利用交叉谱密度/互强度/维格纳函数所满足的传输方程,或在频域引入衍射的角谱理论去描述待测物体对照明光波的散射作用;利用 van Cittert-Zernike 定理、部分相干光学传递函数理论等去描述光学系统对成像过程的影响;最终完成从产生照明光到传感器上产生低维耦合离散光强信号的整体过程的正向数学建模。模型初步建立后,通常还需要利用 VirtualLab、Comsol 等光学仿真软件基于严格的麦克斯韦方程求解算法对成像过程进行模拟计算,并在现有成像系统上利用已知物体实测加以比对,对成像的正向模型进行进一步修正与优化。整个正向物理建模过程依赖于大量的专家知识和经验驱动,严重影响了计算成像技术的通用性和可重用性。

深度学习彻底颠覆了这种“物理模型驱动”的范式,开启了“样本数据驱动”的学习范式。具体体现在两点:第一,所谓的经验和知识也在样本数据中,在数据量足够大时无需显式的经验或知识的嵌入,直接从数据中可以学到;第二,基于深度神经网络特有的“高维特征自动提取”能力,可以直接从原始信号进行学习,而无需借助人为的特征变换或提取。数据驱动的学习范式使得科研人员无需根据经验和知识针对不同的成像问题设计不同的处理流程,从而大大提高了算法的通用性,也大大降低了解决新问题的难度。

### 2) 从“分步/分治”到“端到端学习”

分治或分步法,即将复杂的问题分解为若干简单子问题或子步骤,这曾经是解决复杂问题的常用思路。分步法在计算成像领域,也是被广泛采用的方法论。比如,为了解决数字全息图重构问题,过去经常将其分为预处理、相位解调、衍射计算(数值传播)、焦面判断等若干步骤。再如,为了解决非线性优化问题,可以采用分段线性方式来逼近全局的非线性。这样做的动机虽然很清晰,即子问题或子步骤变得简单、可控、更易解决,但从深度学习的视角来看,其劣势也同样明显:子问题最优未必意味着全局的最优。相反,深度学习更强调端到端的学习,即不去人为地分步骤或者划分子问题,而是完全交给神经网络直接学习从原始输入到期望输出的映射。相比分治策略,端到端的学习有协同增效的优势,有更大的可能获得全局上更优的解。当然,如果一定要把分层看作是“子步骤或者子问题”也是可以的,

但这些分层各自完成什么功能并不是预先设置好的,而是通过基于数据的全局优化来自动学习的。

### 3) 从“病态非线性逆问题”到“直接(伪)正向非线性建模”

计算成像中所涉及的众多复杂逆问题本质上是高度病态且非线性的,而深度学习实现了从输入到输出的非线性变换,这是深度学习在众多复杂问题上取得突破的原因之一。在深度学习出现之前,众多线性模型求解或非线性迭代优化算法是计算成像图像重构的主流技术。对于可通过近似手段线性化的逆问题(如傍轴近似下的相位恢复问题可通过光强传输方程线性化直接求解,Born 或 Rytvo 近似下的某些逆散射问题也可以实现线性化求解),相应的病态方程组求逆、反卷积与偏微分方程求解等是求解这类问题的核心算法;对于无法线性化的逆问题(如非傍轴条件或复杂照明情况下相位重构问题),可基于凸集投影与梯度搜索的优化算法进行迭代求解(尽管解空间往往是非凸的,但事实证明这些优化算法往往是奏效的)。一般而言,基于某些限制性假设的线性化求解方法所得的解可以作为更为一般条件下非线性问题求解的初值,以提高迭代算法的收敛速度与求解的稳定性。针对逆问题的病态性,通常通过引入被测物体的先验作为正则化手段限定解空间以使其良态化。这里值得一提的是压缩感知技术,它由于在解决病态逆问题方面的突出表现成为了计算成像领域中一个耳熟能详的专业词汇。压缩感知的核心假设在于已知解具有稀疏性(Sparsity),因此可以使用少量的数据来接近完美地恢复原始信号。稀疏性可以作为约束或者正则项,提供额外的先验信息。而大部分信号本身并不是稀疏的(即在自然基下的表达不是稀疏的),但是经过适当的线性变换后是稀疏的(即在另一组基下是稀疏的),如离散余弦变换与小波变换等。该领域曾经非常热门的一个研究课题是字典学习(Dictionary Learning)和变换学习(Transform Learning),通过大量的信号实例,自适应地学习最优的稀疏性表达(自建完备字典)。为了求解稀疏约束下的最小化问题(最常用的是总变分最小化),需要进一步确定最小化能量泛函的数值求解算法(最陡下降法、非线性共轭梯度法、迭代阈值收缩法等),以及相应的自适应正则化参数的选取方法(对噪声进行统计建模,并对其局部方差进行准确估计),以获得稳定且有意义的解。

而深度学习则利用复杂的网络结构与非线性激活函数,在提取样本高维特征的同时,不断去除与目



标特征无关的信息,最终获得了足以适配足够复杂系统的非线性变换能力。实际上,深度学习同样隐含了稀疏性先验,但它认为问题本身就是稀疏的,可以不断将输入数据进行“降维”,把高维的数据空间投影到抽象但低维的认知空间。相比于传统方法“先正向建模,再求解该模型下的逆问题”的思路,深度学习技术不需要这种数据表达过程的可逆性。它直接建立了从图像到待恢复信息的“伪正向模型”——将光学系统拍摄到的图像作为“网络输入”,将待恢复的期望信息作为“网络输出”,巧妙越过“非线性病态逆问题”求解这一大障碍,直接通过高维度特征拟合实现图像与信息的提取与重建。

正是由于深度学习为计算成像技术从思想观念上带来了重大变革,且其研究内容也是极其多样与发散的,目前还没有一个比较明确的分类方法。如果按研究的问题或者成像的体制来分的话可能会较为琐碎。因此,在本节中我们将按照采用深度学习技术的“目的与动机”或者说“深度学习技术为传统

计算成像技术带来了哪些新的要素”进行细分。一般而言,引入深度学习均是为了从不同的角度、采用不同的方法来解决传统计算光学成像三方面的问题:

1) 提升传统计算成像技术的信息获取能力:突破传统计算成像技术的“信息量守恒”准则,从极少量原始图像数据中解耦并挖掘出更多场景的本质信息。

2) 降低传统计算成像技术对“正向物理模型”或“逆向重构算法”的过度依赖:绕过精确物理建模与病态非线性逆问题求解的障碍,使计算成像技术实施起来更加简单智能。

3) 突破传统计算成像技术所能够达到的功能/性能疆界:实现传统计算成像技术因物理模型所限而无法实现的功能与无法达到的性能指标。

按此方式分类的整个框架如图 2 所示。下面我们就依据此分类方式,对现有深度学习计算成像技术以及典型应用进行概述。

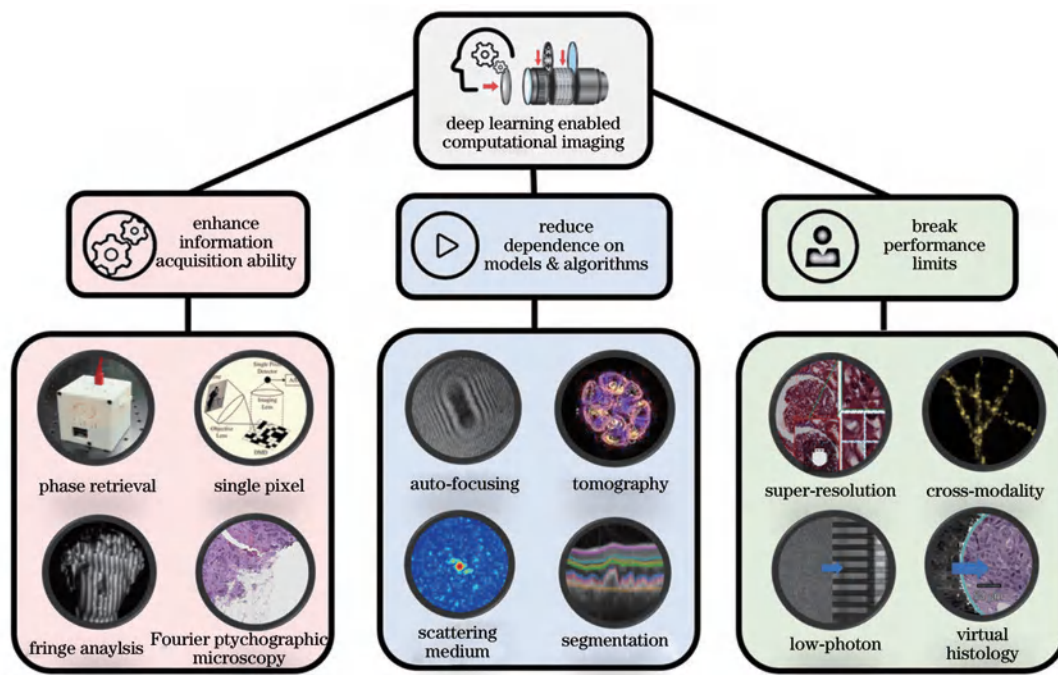


图 2 基于“目的与动机”对典型深度学习计算成像技术所作的分类

Fig. 2 Classification of typical deep learning based computational imaging techniques according to their objectives and motivations

### 2.1 提升传统计算成像技术的信息获取能力

从信息论的角度而言,数据不会凭空产生,但计算成像技术将光学系统的信息获取能力与计算机的信息处理能力相结合,通过光学调控与相应的信息处理技术从原始图像数据中解耦并挖掘出更多场景的本质信息。然而,对于传统的计算成像技术而言,

研究者们都会或多或少地在直觉上遵守着“信息量守恒”准则,而深度学习技术的出现为突破这一准则的约束提供了可能性。

实例 1——相位恢复(无透镜显微/同轴全息)

在 Gerchberg 等<sup>[43-44]</sup>所提出的迭代相位恢复算法(G-S 算法, Gerchberg-Saxton 算法)中,为了确保

解的存在性与唯一性(排除孪生像)<sup>[45-47]</sup>,避免因迭代算法陷于局部极小值造成的收敛停滞<sup>[48-50]</sup>等问题,往往需要采集两幅甚至多幅不同离焦距离上的衍射图像(同轴全息图),从而利用更多测量值的约束来提高算法的收敛性和可靠性<sup>[51]</sup>。然而,代价是需要额外获取大量的原始数据,且成像系统依赖于高精度的轴向位移装置。仅仅依靠单幅衍射图虽然也可以实现图像重构,但一般仅限于尺寸较小且分布稀疏的样本(采用空域支持域约束),对于一般的较大尺度的样品而言,由相位缺失造成的“孪生像”会在物体成像周围产生自干涉“伪影”,极大地影响成像质量。为了解决这一问题,2018年 Rivenson 等<sup>[4]</sup>基于

深度学习提出了单帧相位恢复技术,该方法的思想在于仅使用一幅相机拍摄到的离焦强度图像进行相位恢复。将拍摄到的离焦图像直接反向传播至焦平面,即将传播得到的复振幅作为网络输入,以使用8幅离焦图像通过传统 G-S 迭代算法得到的样品清晰相位为目标,利用深度神经网络来模拟相位提取算法的过程,成功地在孪生像与物体伪影的干扰中从单幅同轴全息图提取并分离出了待测样品的真实信息,获得了准确的振幅与相位信息(图3)。相比于传统的 G-S 迭代算法,不仅避免了复杂的迭代优化过程,还大大降低了成像所需的图像数目,单幅重建也使得该系统不再需要复杂的轴向机械位移装置。

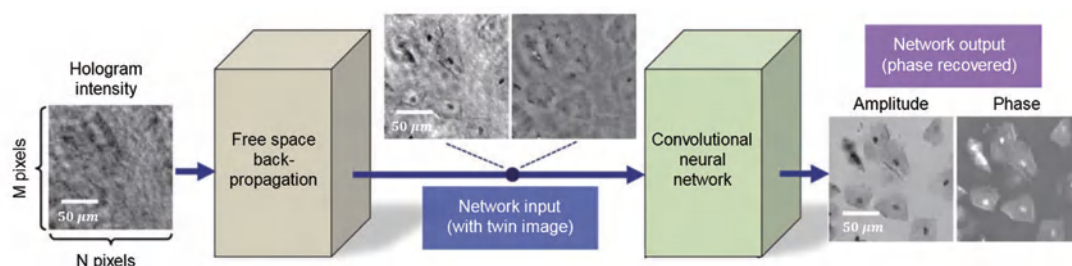


图3 使用深度学习进行单帧无透镜相位恢复<sup>[4]</sup>

Fig. 3 Single-frame lensless phase recovery using deep learning<sup>[4]</sup>

#### 实例2——傅里叶叠层显微成像

2013年,Zheng等<sup>[52]</sup>提出了一种基于相位恢复与合成孔径的计算显微成像技术——傅里叶叠层显微成像(FPM)。在该技术中,样品被不同角度的照明光束(通常是一个LED阵列)依次照射后,由低数值孔径物镜拍摄一系列低分辨率图像。由于成像系统有限孔径效应对频域的低通滤波,改变了照明光束角度,实现了物体频谱在频域子孔径的交叠扫描。傅里叶叠层成像的核心优势在于其不仅仅能获得待测样品的相位信息,还能在基于最优化的交叠更新过程中实现频域内的合成孔径,有效促进了成像分辨率的提高<sup>[53-55]</sup>。与传统的频域合成孔径超分辨率算法不同,傅里叶叠层成像交叠更新算法的相位恢复与频域合成孔径是同时完成的,这也正是傅里叶叠层成像技术本身的优美之处。通常情况下,低数值孔径的低倍率物镜本身具有很大的观察视场,再加之利用大角度照明光束依次照射样品,并在频域进行合成孔径,最终将成像的等效数值孔径提升到物镜与照明数值孔径之和,即保持低倍率物镜的大视场的同时,又达到很大的成像空间带宽积。然而,傅里叶叠层显微成像中空间带宽积的提升往往是以大量(数百幅)低分辨率图像数据采集与高度复杂的频域变换与空域约束反复迭代为代价的。为

了解决这一问题,2018年,Nguyen等<sup>[10]</sup>提出将深度学习技术应用于FPM领域。该方法思想在于将使用传统FPM技术获得的相位作为目标真值,将在自行设计的照明模式下拍摄到的5幅图像作为输入图像,利用深度神经网络模拟FPM中图像信息的提取与合成,从输入图像中提取物体的相位信息,整个过程如图4所示。经过训练的网络在保证重构成像质量的前提下,将FPM成像所需的图像数目大大降低,并在减少数据量的需求的同时,避免了传统重构算法繁琐的迭代优化过程。

#### 实例3——条纹相位分析(相位测量)

在条纹相位分析领域,光学相位测量技术已被广泛应用于光学干涉测量、数字全息、电子散斑干涉、莫尔轮廓术及条纹投影轮廓术等。这些方法的一大共性在于通过干涉或者投影的方式在物体表面形成周期性的结构条纹,从而使所测量物体的相关物理量直接或间接地反映在条纹的相位信息中。所以,从根本上而言,这些光学技术的测量精度直接取决于条纹图案的相位解调精度。因此,条纹图案分析是光学相位测量技术中最核心的步骤,也通常是最困难的部分。经典的条纹分析技术大致可分为两类:

1) 时域相移解调法:采用多幅具有相对相位差的条纹图像进行相位提取。该方法能够实现像素级



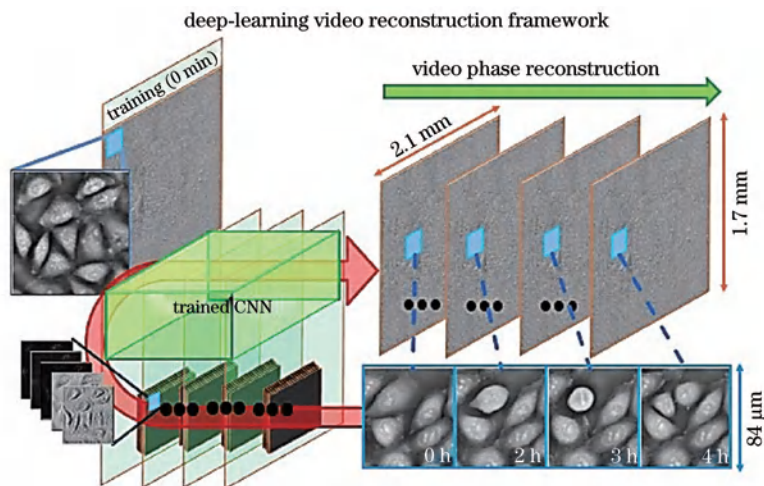


图 4 使用深度学习技术进行少图快速 FPM 成像<sup>[10]</sup>

Fig. 4 Fast FPM imaging with few images using deep learning technology<sup>[10]</sup>

的高分辨率相位测量,但需要采集多幅条纹图像,容易受到物体运动/环境振动等外界干扰的影响,通常难以应用于动态测量。

2) 空域相位解调法:仅采用单幅条纹图像(通常是包含载频的高频条纹)实现相位信息的提取,如傅里叶变换法(FT)、加窗傅里叶变换法(WFT)等。但对条纹陡变、不连续以及物体细节丰富的区域较为敏感,难以实现高精度、高分辨率的相位测量。且算法一般具有较多的参数(如滤波窗尺寸等)需手动调节,难以实现全自动化操作。

针对这一问题,本课题组首次将深度学习技术应用在条纹分析中,并有效提高了条纹投影轮廓术的三维测量精度<sup>[37]</sup>。该方法的思想在于仅采用一幅条纹图像作为输入,利用深度神经网络来模拟相

移法的相位解调过程。如图 5 所示,构建两个卷积神经网络(CNN1 和 CNN2)。CNN1 负责从输入条纹图像( $I$ )中提取背景信息( $A$ );随后,CNN2 利用提取的背景图像和原始输入图像生成所需相位的正弦部分( $M$ )与余弦部分( $D$ );最后,将该输出的正余弦结果代入反正切函数计算得到最终的相位分布。相比于傅里叶变换法与加窗傅里叶变换法,该方法能够更为准确地提取相位信息,特别是针对具有丰富细节的物体表面,相位精度可提升 50% 以上,仅采用一幅输入条纹图像但总体测量效果接近于 12 步相移法[如图 5(a)~(d)所示]。该技术目前已被成功应用于高速三维成像,实现了速度高达 20000 frame/s 的高精度三维面型测量<sup>[40]</sup>。

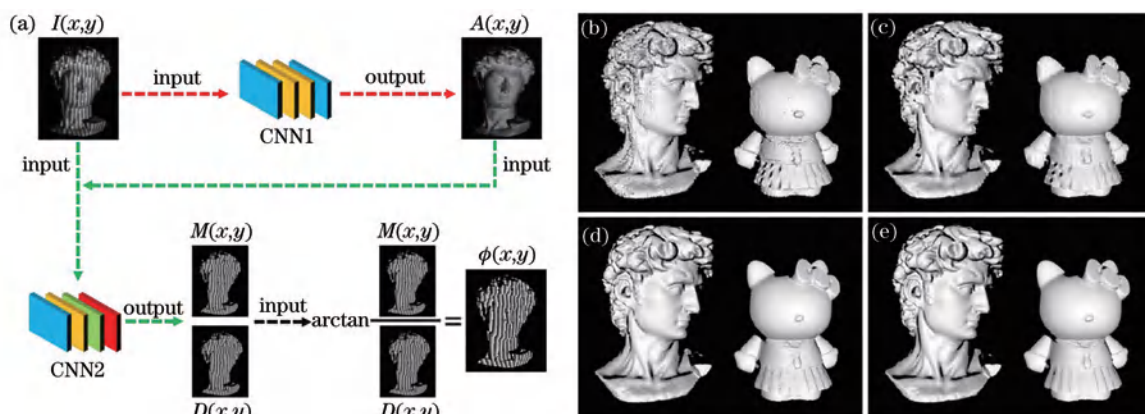


图 5 基于深度学习的条纹分析方法原理与相位重构结果对比<sup>[37]</sup>。(a)基于深度学习的条纹分析方法原理图; (b)傅里叶变换法重构结果;(c)加窗傅里叶变换法重构结果;(d)深度学习法重构结果;(e)12 步相移法重构结果

Fig. 5 Principle of fringe analysis method based on deep learning and comparison of phase reconstruction results<sup>[37]</sup>. (a) Principle of fringe analysis method based on deep learning; (b) reconstruction result of FT; (c) reconstruction result of WFT; (d) reconstruction result of proposed deep-learning method; (e) reconstruction result of 12-step phase-shifting profilometry



实例 4——单像素成像(计算“鬼”成像/“对偶摄影”)

在很多重要的成像领域,如极弱光成像、远红外成像、深紫外成像等,很难制造出具有高空间采样率并且成本低廉的阵列探测器。而制造满足相同技术指标的单像素探测器,尤其是制作非可见光波段的单像素探测器,要容易且成本低廉得多。因此,单像素成像成为了简化成像系统、降低成本的一个良好选择。单像素成像最早起源于双光子纠缠鬼成像<sup>[56]</sup>,利用纠缠态光子对的空间信息相关性来探测目标物体的空间信息,从而实现了对物体图像的重建,随后该项技术被拓展到了热光源<sup>[57]</sup>与赝热光源<sup>[58]</sup>。而“计算鬼成像”技术通过空间光调制器(SLM)产生随机散斑以模拟光子的随机性,因此无需再使用面阵探测器来探测散斑图案,即仅需要使用一个单像素探测器(作为唯一的探测器),就可以实现真正意义上的“单像素成像”技术。2005年,Sen等<sup>[59]</sup>提出了“对偶摄影”(dual photography),利用投影仪与摄像机的互换性实现了一系列新奇的成像功能,如场景渲染与绕墙成像。或许很多光学成像领域的研究人员现在还不知道:当下十分热门的单像素成像技术<sup>[60-61]</sup>(起源于2006年)与计算鬼

成像技术<sup>[62]</sup>(起源于2012年)其实就是对偶摄影的一种特殊形式。不论是单像素成像、计算“鬼”成像还是“对偶摄影”,都需要对场景进行多次图案投影并用单像素探测器收集散射光场,故往往需要上万次甚至数十万次的原始数据测量,十分繁琐耗时,难以实现动态成像。压缩感知技术<sup>[61]</sup>利用“先压缩,后采样”的压缩感知思想来得到物体信息在空间域的欠采样数据,并以稀疏性先验为约束,以较少的测量数据重建出物体的图像(实验中仅使用1500次测量,就重建出像素数为原图像像素数的2%的图像)。该方法虽然大大降低了采样数据量,运算却极其复杂耗时,并且很难准确恢复出图像细节部分。针对此问题,Lyu等<sup>[14]</sup>于2017年首次提出了一种基于深度学习的单像素技术(如图6所示)。该方法利用空间光调制器显示图像数据集,并获取了相机拍摄得到的原始图像。之后以空间光调制器上显示的图像作为训练目标,分别在以原始图像和传统鬼成像法得到的相位图作为输入的情况下对神经网络进行训练,从而在5%的信息采样率下获得了显著超过传统压缩感知鬼成像的重构结果,并且该方法在噪声鲁棒性方面也得到了明显提升。

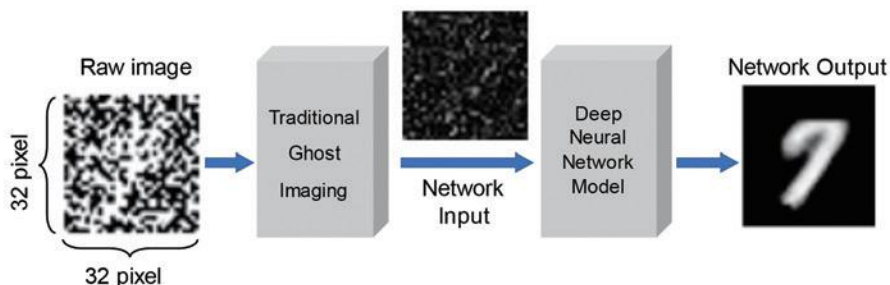


图6 使用深度神经网络的单像素技术框架<sup>[14]</sup>

Fig. 6 Framework of single-pixel technique using deep neural network<sup>[14]</sup>

## 2.2 降低传统计算成像技术对“正向物理模型”或“逆向重构算法”的过度依赖

正如前文所言,计算成像的两大核心内容之一正是如何准确地建立场景与图像之间的正向模型,并精确地设计成像系统,使得采样数据能够包含全部的场信息。这就要求所涉及的正向物理模型能够真实地反映真实世界成像的物理过程。然而,由于实际成像过程的复杂性与不可预知性,精确的正向物理模型通常难以获得,从而就会使得我们使用计算成像技术获得的图像信息与真实情况相差甚远。而深度学习则利用复杂的网络结构与非线性激活函数,直接建立了从原始图像到待恢复信息的“伪

正向模型”,打破了传统计算成像技术中“正向物理模型精确可知”这一限制性条件。作为一个强大的高维特征提取工具,神经网络是直接从大量样本数据中学习输入与输出之间的复杂高维关联,难以预知的不确定因素(如噪声、像差等)也自然而然地被纳入其中,最终获得足以适配足够复杂的真实成像系统的非线性变换能力。

除了如何设计精确可知的正向物理模型外,计算成像技术另一大核心内容便是如何利用逆问题优化求解来重构图像。这要求在重构过程中保证重构图像的准确度。然而,传统的数值优化算法与物理模型中大量的“中间参数”往往需要相关从业人员进

行手动调节选取,最终的图像重构质量很大程度上依赖于参数的人为选取,难以实现无人工干预下的全自动处理。而深度学习技术具有“端到端”的特殊映射机制及高维特征“自行提取”的特点,模型一旦训练完成后就没有任何自由参数需要调节,从而可完全实现“无参数”与“全自动”。

#### 实例 1——(穿透)散射介质成像

光在均匀介质中是沿直线传播的。然而当经过浑浊媒介、生物组织等介质时,光会在这些介质内发生多重散射,出射后的光场将变为散斑场。这是一种不可逆的扩散过程,严重影响了目标的可见性。在传统计算光学成像范畴,实现(穿透)散射介质成像的方法包括反馈波前调制<sup>[63]</sup>、传输矩阵<sup>[64]</sup>、相位共轭技术<sup>[65-66]</sup>、散斑相关<sup>[67-68]</sup>等。然而这些方法的有效性往往被限制在光学记忆效应区,即介质的散射作用不可过强并在一定入射角内可以被视为一个线性移不变系统。而复杂强散射介质形成的散斑空间分布是散射体微观排列和入射光场波前的复杂函数,难以对其建立全面精确的物理模型并给出简单直接的逆散射解决方案,且算法对于不同类型的散

射介质的可迁移性较差。

深度学习技术为解决这些问题提供了很多新的思路。Li 等<sup>[28]</sup>提出了一种具有统计特征的“一对多”深度学习技术(如图 7),该技术从大型数据集中识别出隐藏的统计不变性,其封装了多个微观结构不同的散射介质系统的一系列统计变化,使神经网络模型能够适应散斑的去相关(decorrelation)。卷积神经网络能够学习在具有相同宏观参数的散射体上捕获的散斑强度图案中包含的统计信息(如图 7 所示)。经过训练后,该网络能够迁移至未经训练的散射介质环境中进行成像,且可对不同类型物体生成高质量的目标预测,在数据类型、系统结构等方面表现出了良好的泛化性。Lü 等<sup>[69]</sup>构建了混合神经网络(HNN)模型,在强散射情形下实现了隐藏物体的恢复。实验中使用的散射介质是 3 mm 厚的白色聚苯乙烯平板,其记忆效应范围小于  $0.01^\circ$ ,光学厚度为 13.4,远超记忆效应区。说明了基于深度学习的散射成像方法可以不受“正向物理模型精确可知”的约束,突破传统技术中依赖的记忆效应视场角的限制。

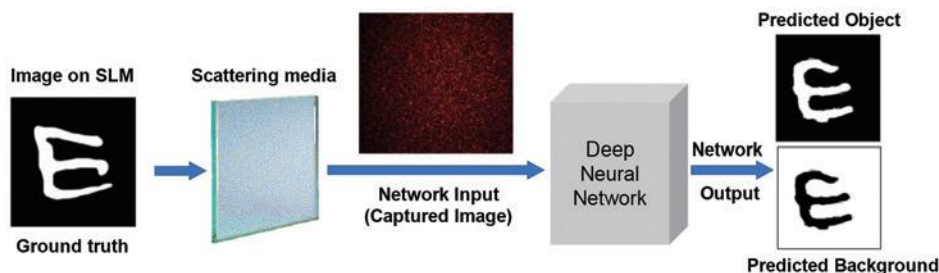


图 7 基于深度学习进行散射介质成像的网络原理图<sup>[28]</sup>

Fig. 7 Network of deep learning based imaging through scattering medium<sup>[28]</sup>

#### 实例 2——三维衍射层析成像

在讨论相位成像技术时,通常都会假设大部分待测物体属于二维(薄)物体,可以将其表示为由吸收与相位构成的二维的复透射率分布,透射光场的复振幅分布即为入射光场复振幅与物体复透射率的乘积。然而,相位延迟其实是样品三维折射率在一个二维平面上的投影(俗称 2.5D 成像),这是一个沿光传播方向的从入射表面到出射表面各个平面的相位延迟的积分变化量,并不是“真三维”的立体信息<sup>[70-73]</sup>。三维衍射层析技术<sup>[70,72,74-75]</sup>可以有效解决这一问题,其可以对三维样品内部各点的折射率实现全方位(横向+轴向)高分辨率成像,从而获取样品三维折射率分布。该项技术通常需将相位测量技术(数字全息或相位恢复技术)与计算机断层扫描技术相结合,通过旋转物体<sup>[76-77]</sup>或改变照明方向<sup>[78-81]</sup>

等方式得到多组定量相位信息,然后结合反投影滤波<sup>[82]</sup>、逆 Radon 变换(忽略衍射效应)<sup>[78,83]</sup>或是 Wolf 的衍射层析理论(考虑衍射效应)<sup>[70,79-80]</sup>,重建出物体的空间三维折射率分布。图 8 给出基于深度学习进行三维衍射层析重建的基本框图。近年来,“强度衍射层析技术”——一种基于非干涉强度测量原理的衍射层析技术逐渐崭露头角。相比于传统光学衍射层析技术,该方法只需要直接拍摄物体不同焦面或者不同照明角度的强度图像,再利用图像重构算法就可以反演出物体的三维折射率分布,这有效避免了传统衍射层析技术干涉测量与光束机械扫描的难题。强度衍射层析成像技术主要分为两类:基于轴向扫描的三维光强传输技术<sup>[84-86]</sup>与基于角度扫描的三维傅里叶叠层成像技术<sup>[87-89]</sup>。不管是哪种技术都依赖对光与三维物体相互作用最终形成图像

的物理过程的准确数学模型:这通常需要假设样品满足纯/弱/缓变相位近似、Born 或 Rytvo 近似的弱散射近似、多层叠加近似(multi-slice)或非负折射率近似等。然而,不论哪种近似都存在一定的局限性,特别地,在强散射、多次散射、后向散射、大数值孔径

照明情况下的普适性普遍较差,因此目前文献中大多数衍射层析成像的实验结果并不十分理想。另一方面,三维层析成像往往还需要巨量的原始数据,这也对其后续算法的高效重建(速度与存储上)提出了巨大挑战。

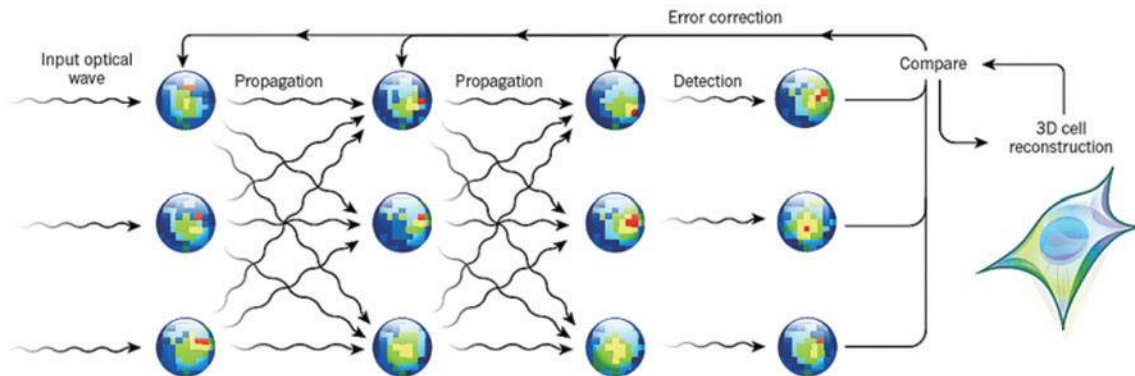


图 8 基于深度学习进行三维衍射层析重建的基本框图<sup>[26]</sup>

Fig. 8 Basic framework of 3D diffraction tomography reconstruction based on deep learning<sup>[26]</sup>

借助于机器学习算法则有望巧妙地“规避”上述问题。Kamilov 等<sup>[90]</sup>将神经网络应用于衍射层析成像,他们利用神经网络构建了一个类似于 multi-slice 的三维衍射传播模型,将目标对象按一组薄片切片建模:每个切片由一个网络层表示,三维对象的每个像素由网络节点表示。因为入射光与三维物体的散射过程及其复杂,难以通过理论推导得到完善的数学模型,所以通过神经网络强大的高阶拟合特性去“学习”是一种巧妙且有效的替代方案(注意,由于采用的网络结构较为简单、层数较少,他们的工作严格来说并不算深度学习,但其实核心思想别无二致)。神经网络的训练数据由一组从不同角度捕获的三维物体的二维全息图组成,使用“反向传播”最小化训练数据和模型解之间的差异来预测物体的三维折射率。他们根据实验获得的数据直接通过该方法训练后的神经网络成功恢复出了 HeLa 细胞的三维折射率结构。2018 年,Nguyen 等<sup>[27]</sup>利用深度学习技术和衍射强度分布直接对样品三维折

射率进行重建。为了获取相应的训练数据集,他们首先仿真了一个具有不同空间折射率分布的物体,并生成了一系列不同角度投影的二维的相位图像。再将二维图像显示在空间光调制器上并通过相机拍摄到了一组原始的强度图,利用这种“仿真与实验结合”的方式完成了数据集的构建。然后将相机拍摄到的一组原始光强信息经过逆 Radon 变换后作为网络输入,通过深度学习网络使其直接匹配至样品的三维折射率分布(图 9)。尽管这种处理方式的合理性与最终实验结果的准确性仍然有待商榷,但这的确为三维衍射层析数据重建提供了一种新颖的思路。

### 实例 3——数字全息成像自动聚焦

在数字全息图的重建过程中,具体样品位置具有不可预知性,一种通常的做法是分步处理并循环搜索:如首先对全息图进行预处理去除噪声提升信噪比与对比度(可选),然后通过相位求解算法获取衍射场的相位分布,再经过衍射计算(数值传播)实现不同焦面图像的重建,最后经过经典的清晰度判据(如梯度、

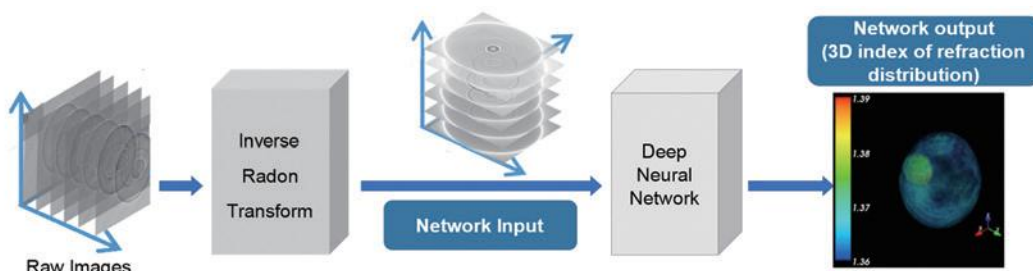


图 9 基于深度学习进行光学衍射层析的网络原理图<sup>[27]</sup>

Fig. 9 Schematic of network of optical diffraction tomography based on deep learning<sup>[27]</sup>



图像熵等)判断当前面是否是真实物体所在的平面,对上述过程反复进行迭代搜索以获取焦面的最优估计。然而这一过程不仅繁琐费事,且依据传统聚焦判据得到的结果还有可能不够准确。为了简化这一复杂过程,2018年Ren等<sup>[6]</sup>提出利用深度学习网络对数字全息图的离焦距离进行预测。该方法主要思想在于利用高精度位移台控制物体位置以在多个不同离焦距离下拍摄到相应的数字全息图,从而构建了数字全息图与其相应离焦距离的数据集,并利用深度学习神经网络进行训练,直接建立离焦距离和衍射图之间的对应关系,输出参数只有一个数字,即对应了样品的离焦距离。如图10所示,经过训练的深度学习神经网络无需迭代搜索,可以直接根据输入的数字全息图输出相应的离焦距离。更进一步,Zhang等<sup>[9]</sup>利用深度学习技术从离焦的离轴干涉全息图中直接恢复得到聚焦状态下的相位与振幅,极大简化了传统数字全息技术重建过程中对于物理模型参数的调整及获取过程。

实例4——光学相干层析成像图像分割

光学相干层析成像因可以获得微米级分辨率的人体组织三维截面图像,被广泛应用于医学与工业

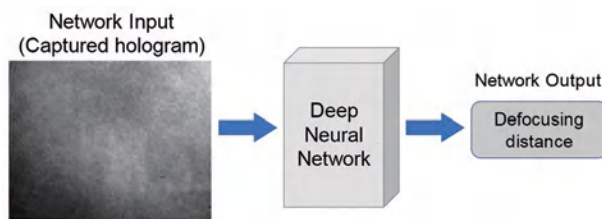


图10 使用深度神经网络的数字全息离焦距离计算框架<sup>[6]</sup>  
Fig. 10 Framework of defocusing distance calculation in digital holography based on deep neural network<sup>[6]</sup>

成像领域中。在许多视网膜疾病的研究中,光学相干层析成像的图像信息的准确量化(如视网膜图像的边界分割)对于提高病灶识别及致病过程等因素的分析至关重要。然而,光学相干断层图像中视网膜的边界分割往往依赖于医生的经验,难以实现全自动处理。Fang等<sup>[23]</sup>提出了一种结合卷积神经网络和图形搜索方法的视网膜光学相干断层图像边界自动分割框架,原理如图11所示。得益于深度学习神经网络对特殊视网膜层特征的准确提取,该方法可对九层视网膜边界进行准确分割,有效避免了人工分割时的主观性和时间成本。

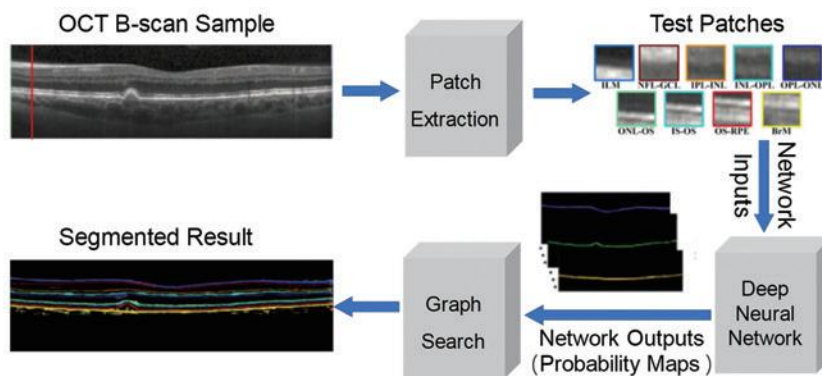


图11 针对视网膜光学相干断层图像的边界自动分割原理图<sup>[23]</sup>

Fig. 11 Schematic of automatic boundary segmentation framework for retinal OCT image<sup>[23]</sup>

2.3 突破传统计算成像技术所能够达到的功能/性能疆界

计算成像技术通过将光学系统的信息获取能力与计算机的信息处理能力相结合,突破了传统光学成像系统对于成像器件硬件的过度依赖。然而“物理模型驱动”的计算成像技术不可避免地受到物理模型的制约,这些制约既包括光学系统与成像条件的制约,如衍射极限、空间带宽积、成像光子数、器件灵敏度等“先天不足”,也包括重构物体信息与光信息为载体之间的关联性制约,如相位、光谱、三维等与所获取的强度信息间需要具有“显式”关联。由于信息不会凭空产生,因此当所获取的图像数据因受

这些制约而无法直接与目标信息相关联时,传统计算成像技术是无能为力的。而在深度学习技术中,由于最终的输出不仅仅取决于输入数据,还和神经网络从大量训练数据中学习到的成百上千万个权值参数紧密相关,这些参数不仅为图像重建提供了强大而完备的“先验数据库”,还建立了输入与输出间难以通过简单公式表示的“隐性高维”关联,从而有望实现数据“无中生有”与“点石成金”,为突破传统计算成像技术的物理模型限制和拓宽其功能/性能疆界提供了可能。

实例1——超分辨成像(突破衍射极限)

由于成像系统具有有限孔径效应,一个理想

物点发出的光在图像平面并不会形成一个理想的几何点,而是会形成一个弥散斑(艾里斑)。对于非相干的衍射受限系统而言,艾里斑的半径为 $1.22\lambda/NA$ ,其中 $\lambda$ 为成像光波的波长, $NA$ 为成像系统的数值孔径,这被称为“阿贝衍射极限”<sup>[91]</sup>。为了提升光学显微镜的分辨率,往往需要采用高数值孔径的油浸物镜,使用起来非常不便。为了解决这一问题,Rivenson等<sup>[20]</sup>于2017年提出了基于深度学习的显微成像超分辨算法。该方法首先利用40倍物镜(干)和100倍物镜(油)下拍摄的同一样品图像生成数据集,而后利用深度神经网

络学习低分辨率图像到高分辨率图像之间的“映射”关系。从傅里叶光学理论上,这种映射关系是没有科学依据的。因为低分辨率图像在频域是没有高频细节信息的,不论如何处理,数据也不会“无中生有”。但深度学习的能力的确让人惊叹,经过大量样本训练后的神经网络的确只需输入一幅低分辨率图像(图12左图)即可成功地突破像素分辨率及衍射极限的限制,生成相应的高倍物镜下的高分辨图像(图12右图)。即利用40倍干镜获得了100倍油镜的成像效果,省去复杂油浸物镜观测时的诸多不便。

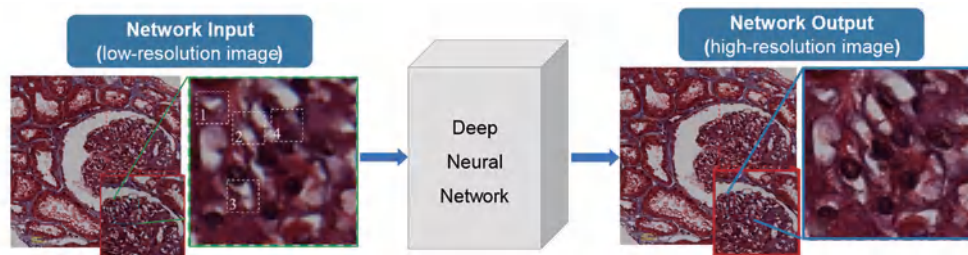


图12 基于深度学习进行超分辨率成像的网络框架示意图<sup>[20]</sup>

Fig. 12 Network framework of super-resolution imaging based on deep learning<sup>[20]</sup>

高数值孔径油浸物镜虽然能够提升成像分辨率,但受阿贝衍射极限所限,最终能达到的分辨率也不会超过光波波长的一半。随着人类对微观世界的探索逐步深入,需要观测的微观尺度越来越小,传统光学显微镜的分辨率已无法满足科学研究的需要,人们迫切需要分辨率更高的显微技术。2014年,诺贝尔化学奖的三位得主使用荧光分子和特殊的光物理原理,巧妙地突破了普通光学显微镜无法突破的“阿贝极限”,其开创性的成就使得人们能够窥探纳米世界,这些技术包括受激发射损耗(STED)技术<sup>[92]</sup>、光激活定

位显微技术(PALM)<sup>[93]</sup>、随机光学重建显微技术(STORM)<sup>[94]</sup>等。但这些技术依赖于复杂昂贵的硬件系统,且实际操作和使用起来非常复杂不便,STED技术逐点扫描的成像机理使其对环境扰动非常敏感,难以实现动态成像,PALM和STORM还依赖于特殊荧光分子标记,成像过程需要成千上万次的图像采集,复杂耗时。针对这一问题,Wang等<sup>[17]</sup>直接通过深度神经网络实现了传统聚焦显微镜图像的超分辨,不借助于任何额外的物理硬件获得了与STED技术相当的成像分辨率,结果如图13所示。

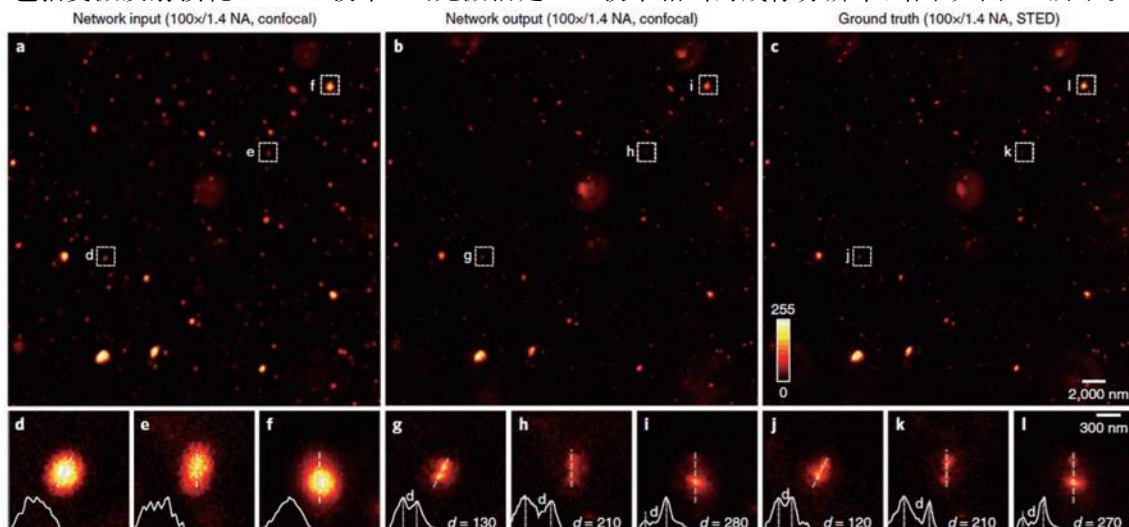


图13 基于深度学习进行STED超分辨率成像的实验结果<sup>[17]</sup>

Fig. 13 Experimental results of STED super-resolution imaging based on deep learning<sup>[17]</sup>



实例 2——高灵敏度(低照度、低光子数)成像

随着电子设备的快速发展和专业摄像器材应用的普及,人们能拍摄到质量越来越高的图片。但是在实际摄影过程中,总会存在各种不可控因素,致使获得的图片存在各种缺陷。尤其是拍摄环境较暗或光照条件不足等会导致图像信噪比较差,随后的转换、传输、存储等操作进一步降低了低照度图像的质量。然而,低照度图像被广泛存在于诸如地质勘测、水下探测与生物学等领域中,如何提高低照度条件下的成像质量就成为了当前一大研究热点。当前,主要从两方面着手:提升探测器元件的响应灵敏度;对探测器获得的信号进行图像增强。然而,目前光探测器响应灵敏度指标正逐渐逼近物理极限(已实现单光子探测),难以进一步满足低照度成像的要求。而简单的低照度图像增强处理算法(如直方图均衡化等)仅能简单提升视觉效果,难以应用于图像保真度要求较高的医疗或科研领域。2018年,Chen等<sup>[34]</sup>将深度学习技术应用于极弱光成像领域。其主要思想为在使用相同相机的情况下,先在极低照度下拍摄到短曝光(约 1/30 s)图像,而后再在长曝光时间下拍摄到的图像作为深度神经网络的匹配目标。经过训练的深度学习网络可以在照度低于 0.1 lx 的情况下,仅根据一幅极弱光条件下拍摄到的短曝光时间(约 1/30 s)图像[如图 14(a)所示]恢复得到一张细节清晰的正常图像[如图 14(c)所示],相比于图 14(b)所示的使用高感光灵敏度 CCD

拍摄到的图像,深度学习所得结果无论是色彩、细节还是阴影中的背景均得到了更好的还原。

实例 3——跨模态成像

对组织标本进行显微成像观察是对临床上多种疾病进行诊断的基本工具,也是组织病理学与生物科学的必备工具。通过临床手段获得组织切片的标准染色图像通常需要一系列复杂工序:福尔马林固定和石蜡包埋(FFPE)、切片(通常为 2~10 μm)、标记染色、风干封片等多个步骤,整个过程极其繁琐耗时。为了简化切片染色流程、降低染色成本,Rivenson等<sup>[35-36]</sup>于 2018 年利用深度学习技术对虚拟组织染色技术进行了研究。该方法的主要思想在于利用组织切片染色前后的图像构建训练数据集,利用对抗神经网络学习未染色切片与染色切片之间的映射关系。虽然,从现有认知范畴而言,这种映射关系是没有科学依据的,因为未染色样品本身不具备化学染料所存在的生化反应过程,更不会具备组织样本各组分特异性。但深度学习却似乎能够从大量测试样本数据中发现这些看似无关的数据集的隐性复杂关联,而这种关联是无法利用我们现有知识体系来建立甚至理解的。实验发现,经过训练的神经网络可以根据一张未染色的切片图像的“自发荧光图像”或者“定量相位图像”直接生成其染色后的结果(如图 15 所示),使得组织学切片分析“绕开”了切片染色这一复杂繁琐的过程,有望为“即时病理诊断”打开一扇新的大门。



图 14 基于深度学习进行极弱光成像的结果<sup>[34]</sup>。(a)摄像机输出(ISO 8000);(b)摄像机输出(ISO 409600);(c)由原始数据(a)恢复得到的结果

Fig. 14 Results of imaging using very weak light based on deep learning<sup>[34]</sup>. (a) Camera output with ISO 8000; (b) Camera output with ISO 409600; (c) recovered result from raw data of Fig. 14(a)

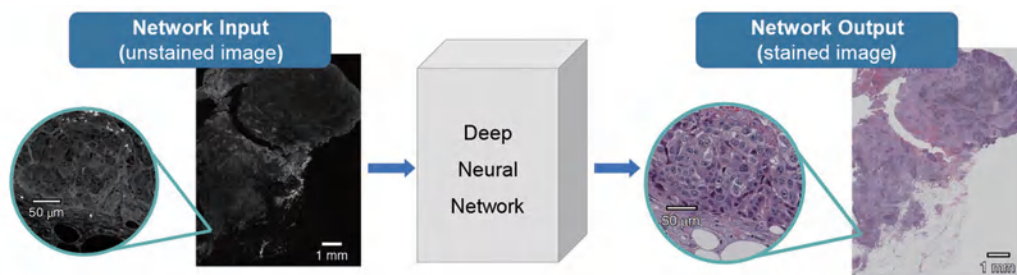


图 15 基于深度学习进行虚拟染色成像的网络框架示意图<sup>[35]</sup>

Fig. 15 Network framework of virtual staining imaging based on deep learning<sup>[35]</sup>



### 3 深度学习下的计算成像:挑战

第二章从提升信息获取能力、降低对“正向物理模型”与“逆向重构算法”的依赖,以及突破传统计算成像技术的性能极限这三个方面简述了深度学习下计算成像的研究现状与最新进展。由于篇幅限制,仅列举了一些该领域内的代表性案例,但从这些典型案例中已经可以看出,深度学习下的计算成像领域取得了许多令人振奋的进展,这些定将为计算成像领域的进一步发展注入新的活力。本章中,将把注意力进一步向前转移,探讨深度学习在计算成像领域所面临的诸多挑战,这将有助于为我们未来的下一步研究制定大胆的战略构想。

#### 3.1 (实测)训练数据的获取与标注成本高

近年来,深度学习技术的发展及其相关应用呈现爆炸式增长。值得注意的是,神经网络的想法并非最近几年才提出,它已有几十年的历史。但是直到最近它才受到如此广泛的关注。这其中重要的因素在于神经网络是一种以数据为导向的算法,卓越的网络输出取决于大量/巨量的数据训练。正是近年来互联网时代积累的大量数据和算力,释放了深度学习神经网络的潜力,也为人工智能应用的飞跃式发展打下了坚实基础。但就目前而言,大部分在计算成像领域中取得成功应用的深度学习技术的案例中,往往需要借助实际成像系统,通过实验获得大量的实测训练数据来进行标注(获取所期待的真值)。这里就存在两个关键问题:

1) 耗时费力:大部分光学成像实验的数据采集过程都是极其复杂且耗时的,而深度学习所依赖的大规模训练数据的获取与正确标注无疑让这一问题雪上加霜,因此需要耗费大量的人力和物力成本。加之光学成像领域的公开数据集稀少,这又增加了深度学习计算成像技术的实施难度。虽然有一些方法可以减少其对数据的依赖,比如迁移学习、少样本学习、无监督学习和弱监督学习,但是到目前为止,它们的性能还没法与大样本的监督学习相比。

2) 真值难知:数据采集后无法获得目标准确可靠的理想真实值,是数据集构建的另一大限制性因素。例如,进行物体相位信息的获取时,往往需要使用传统算法进行信息提取。然而这不仅制约了深度神经网络的表现,即网络的输出质量很难超越用于训练的数据,还使得多种类、大数据量的信息获取十分困难、繁琐。因此,在进行深度网络训练时,如何获取真实、有效并且具有良好代表性的数据集仍是

计算成像领域面临的一大问题。

值得注意的是,对于某些正向物理模型精确可知的光学成像应用而言,一种可能的做法是直接通过物理建模并在计算机中进行仿真来获取网络训练所需的大量训练数据。这种方式虽然规避了上述两个问题,但也丧失了深度学习技术的最大优势:深度学习能够从大量实验样本数据中学习实际成像系统输入与输出之间的复杂高维关联,获得足以适配足够复杂的真实成像系统的非线性变换能力。而计算机仿真是无法准确还原实际系统中难以预知的不确定因素(如噪声、像差等),所以所学习到的模型也不一定能够真实全面地反映实际的物理成像过程,从而将有可能得不到理想中的成像结果。

#### 3.2 目标合作度与环境稳定性要求高

在计算成像领域,为了能够获得足以适配足够复杂的真实成像系统的映射能力,深度学习训练数据往往通过实际的成像系统获取。这往往需要复杂耗时的数据采集过程,不仅使得训练集构建繁杂耗时,更为深度学习在一些特定领域的实际应用,如穿透散射介质成像(军事、生物成像等),带来了一定的困难。这主要是由以下两方面问题导致的:

1) 目标合作度要求高:传统的监督学习方法往往需要由大量具备不同特征目标的样本组成的数据集进行训练。然而方法本身隐含着目标侧灵活可控的假设,只有满足这一假设才能稳定地获得对应样本的真值信息。然而,在某些实际情况下,满足这一条件是困难的。例如,穿透散射介质成像中通常需获得大量已知标准样本的散斑场图像,这就需要在物体侧频繁替换目标物。而在实际军事(如穿透雾霾、遮障等)成像环境中,很难直接对目标侧进行自由操控。生物医学成像(深层穿透成像)应用也存在类似的问题,在人体皮下或者脏器内置放(大量)合作目标通常也是不切实际的。

2) 环境稳定性要求高:在构建深度神经网络训练所需的数据集时,往往需要假设成像系统的物理模型是趋于稳定甚至不随时间变化的,这样所采集到的大量实验样本数据才能够集中体现成像系统真实复杂的物理成像过程。然而样本采集过程往往十分繁杂耗时(可能长达数小时甚至数天),这就必须保证数据采集过程中的系统环境要尽可能地保持稳定一致。然而,在许多实际应用中,该条件往往难以满足。仍以穿透散射介质成像为例:雾霾、遮障、大气湍流、水下浑浊等外界环境往往都会随时间而改变,从而使得获取的数据难以真实有效地反映实际

成像时散射介质对成像系统的影响。生物医学成像应用中也存在类似的问题:人体的呼吸、血流及新陈代谢活动都是永不停息的,这也意味着散射介质并不会在较长的时间内保持一成不变,这种系统与环境的稳定性给深度学习技术的实际应用带来了相当大的挑战。

### 3.3 网络结构的选取趋于经验主义

针对特定的成像需求,到底选择什么样结构的神经网络合适?这是初次尝试深度学习的研究人员经常面对的一个问题。尽管从前人的相同或者相似工作中能找到网络结构设计的灵感,但是在神经网络后期的调试与优化过程中,如何调整超参数(Hyper Parameters,如神经网络的层数、CNN中滤波器的大小、特征的数量等)使得该网络能够在既定的应用中表现出色仍是一个难以回答的问题。通过试错法进行超参的调整尽管有一定效果,但这一手段依赖于从业者本身对深度学习调参的理解,且试错过程中时间成本过高。

此外,深度神经网络的规模同样也是研究人员在进行网络设计时需要考量的因素。随着神经网络层数的增多,其非线性拟合能力也就越强,往往训练结果的精度也会得到提升,但是当网络层数、参数数量达到一定规模时,不仅训练过程会变得复杂,还会给网络能否快速输出结果提出了硬件性能上的挑战。快速输出神经网络的运算对于算力强大的服务器与工作站而言可能算不上太大的负担,但对于移动终端或穿戴设备(如手机、平板等)而言,其往往难以承担规模过大的神经网络的部署,这时需要在设计阶段考虑对网络结构、尺寸进行合理限制,合理权衡成像性能与运算资源。

### 3.4 “调参好比炼丹”式的试错法训练机制

对于尝试使用深度学习技术的许多人来说,深度学习方法预测最终结果的过程往往是难以理解的。虽然深度神经网络所基于的基本运算过程,如卷积运算、激活函数运算、梯度求解等,十分简单易懂,但是随着深度神经网络规模的扩大与参数的增加,大量乃至巨量的参数迭代过程、梯度优化过程互相耦合,使得整个训练过程难以理解,导致研究人员在很多情况下只能通过最终的测试结果来对神经网络性能的优劣进行判断,使得优化和提升神经网络性能这一目标充斥着大量的试错过程。但是类似“试错式”的研究方法往往意味着大量的计算资源与时间成本。尤其是对于大规模的深度神经网络,巨量的参数使得完成一次训练甚至可能需要占用数十

乃至上百块高性能TPU、耗费数个昼夜甚至更久的时间,这种多次且无明确目标方向的试错极易带来时间与算力的大量浪费。

近年来,越来越多的研究人员也逐步意识到这个问题的重要性与严重性。为了解释神经网络的学习过程,Zeiler等<sup>[95]</sup>提出了一种针对卷积神经网络的可视化方法。该方法对神经网络学习的特征进行了可视化,为优化网络结构、提升预测的准确性提供了思路。而Shwartz-Ziv等<sup>[96]</sup>于2017年也尝试使用“信息瓶颈理论”解释深度学习的训练过程,其发现了深度学习训练过程存在的“特征拟合”和“特征压缩”两个阶段,并进行了相应的可视化分析。

### 3.5 特定样本训练后的网络缺乏泛化能力

泛化能力评价的是一个神经网络完成训练后,在处理“从未遇见过”的输入数据时的表现。虽然传统计算成像技术的实际成像性能受限于“正向数学模型的准确性”以及“逆向重构算法的可靠性”,但只要“模型全面准确”且“算法稳定可靠”,对于不同的观测对象都可获得较为理想的成像结果。但对于基于数据驱动的深度计算成像,神经网络输入与输出之间关系的建立主要依赖于对大量样本数据的反复训练过程。显而易见,对于训练数据中常见的图像特征,神经网络更容易学习到从该特征到输出结果的映射。而对于训练过程中出现较少的图像特征或者实际成像中遇到的一个区别于训练集的全新样本,神经网络一般难以给出正确的输出。因此泛化能力通常与训练数据样本的规模和多样性密切相关。而正如前文所说,在计算光学成像领域,深度学习技术所依赖的数据集通常获取起来较为困难,仿真得到的数据集与真实的成像过程总是存在偏差,而多类样本的大量实验数据又难以获取,这一困难成为了制约深度神经网络泛化能力的一大因素。

值得注意的是,对于神经网络泛化能力本身需要一分为二来看待。这好比是关于“通才”与“专才”的思考。通才的知识面广,但深度较为欠缺。而专才尽管知识面相对较窄,但能精通一到两项专业特长。对于社会的发展而言,通才和专才都是不可或缺的。回到光学成像的范畴,由于不少应用面向的对象本身就较为单一,因此不断增加同类型的训练数据,可对面向特定应用的成像系统性能提升起到积极的作用。因此,从实际成像需求出发,辩证地看待深度学习辅助下光学成像方法的泛化能力,在某些场合缺乏泛化能力反倒不是一件坏事。



### 3.6 “深度学习下的计算机视觉”≠“深度学习下的计算成像”

在过去的7年中,深度学习技术以其特有的“高维特征自动提取”的功能避免了传统机器学习极其依赖的人工“特征工程”选取工作,迅速成为了机器学习领域的主流。而在这其中,最为令人瞩目的便是其在计算机视觉领域的重大进展:目标识别、三维视频渲染、图像去模糊、图像超分辨率等技术都因为深度学习技术的出现彻底改头换面。然而,当关注到其在计算机视觉领域的成功并为之欣喜时,也同样要认识到其潜在的缺陷:深度学习并不是魔术,数据也并不会“无中生有”,由大量训练数据中学习“先验数据”并不能与某个待测样品所获取“真实信息”画上等号。对于“深度学习所获得的结果是真实有效且准确可靠的”,相信目前谁也没有办法拍胸脯保证,至少现在还无法做到。

可能有人会提出质疑,这真的那么重要吗?不管白猫黑猫,能抓到老鼠的就是好猫。的确在某些领域,这并不算是个问题。计算机视觉所面向的很大一部分应用就是满足人类观测的需要(如消费电子、影视娱乐),人眼是最终的受体也是评价者,“看起来”好看、“看起来”清楚、“看起来”真实足矣(一个更为恰当的英文单词叫 photorealistic)! 而“它归根到底是不是真的”,其实(似乎)并不是那么重要。例如,深度学习技术可基于单张图像实现图像(像素)超分辨率重建,这种数据的无中生有显然是违背经典的信息论的。换言之,无法保证经过深度神经网络所“长出来的”图像细节与真实高分辨率场景中的完全一致。但是“who cares”? 只要知道图像的确变清楚了,马赛克的确消失不见了,这就足矣!

然而,在光学范畴的计算成像领域,上述深度学习技术的潜在缺陷或许是“致命的”。不仅仅是为了满足视觉观测的需要,计算成像技术往往还和工业测量、医疗诊断、科学发现等领域密不可分,这就意味着不仅需要得到一个“看起来还不错”的结果,更需要确保它们的“准确、可靠、可重复、可溯源”。而这些要素都很不幸的是深度学习技术的“软肋”。“如果我都不能保证结果是真的,我还要它有什么意义?”这要求看似有些苛刻,但在某些特定领域的确也是必须的。因为谁都不希望自己的产品在质检阶段被深度学习算法的“存在某个瑕疵”而打上不合格的标签,更不希望自己的体检报告中由于深度学习算法的“长出了某个病灶”而被诊断为得了不治之症。更进一步说,深度学习技术的成功所依赖的是

从大量训练样本中学习并提取的“共性”信息(特征),这恰恰导致其在面对“罕见样本”(与训练数据集差异较显著)时,所得结果的准确性往往并不理想。而对这类“罕见样本”的正确检测与可靠识别往往也是工业检测与医学诊断领域最有意义且最具挑战性的一部分。毫不夸张地说,“异常”更是一切科学新发现的起源。因此,在当前深度学习技术所获得的巨大成功面前,还是应当对其在计算成像领域的应用保持清醒且理性的态度。

### 3.7 “深度学习”缺乏“深入理解”的能力

目前,深度学习技术仍然在很大程度上依赖大量数据进行特征信息提取,换言之,当深度神经网络面对一项截然不同的任务时,需要使用新的数据进行相应训练。正如图灵奖得主、贝叶斯网络之父 Judea Pearl 所说,当前的深度学习不过只是“曲线拟合”,清华大学的张钹院士也曾指出现在的人工智能基本方法有缺陷,而我们必须走向具有理解能力的 AI,这才是真正的人工智能。需要明确的是,现有的深度学习缺乏理解和推理能力的原因在于它缺乏常识信息。举例来说,利用深度学习来进行条纹分析以计算条纹图中蕴含的包裹相位信息,目前的方法是两步走:先利用深度学习技术求解条纹的正弦和余弦部分,然后将它们带入反正切函数计算包裹相位。由于缺乏推理能力,深度学习技术无法预知包裹相位具有不连续的空间跳变这一常识(而陡变区域的高精度拟合往往对于卷积神经网络是非常具有挑战性的),使得难以训练出准确可靠的端对端(条纹到相位的直接映射)的神经网络。为了改善这一问题,需要建立常识库,将常识信息引入到深度学习,使神经网络在预测时既考虑已看到的样本又与有关真实世界的常识相联系。

## 4 深度学习下的计算成像:未来

### 4.1 搭上深度学习技术发展的顺风车

当了解到计算成像系统的性能、功能与成像能力因物理模型(如衍射极限、逆问题模型)等受限时,深度学习技术为计算成像所带来的性能优势就显而易见了。毫无疑问,深度学习技术与深度神经网络模型仍会在接下来的若干年不停地向前发展,而计算光学成像技术也必定将搭着这列顺风车继续快速前行。

#### 4.1.1 对抗学习(GAN)

GAN 是从 Goodfellow 等<sup>[97]</sup>的研究工作里演化出来的一个深度学习分支。图灵奖获得者 Yann



LeCun 曾评价对抗学习为“Adversarial training is the coolest thing since sliced bread(对抗训练是自切片面包以来最酷的事情)”。这个由博弈论启发而产生的技术包含两个算法,一个是生成器算法,一个是鉴别器算法,它们的目标是在训练的过程中欺骗对方。当这两者的博弈达到平衡时,模型训练结束。此时利用生成器即可输出最终的结果。

GAN 的优势在于它是一种以半监督方式训练的方法,适合标签较少的训练数据。而且 GAN 模型只用到了反向传播,不需要复杂的马尔可夫链。此外,GAN 不仅可用作图像生成,还可以用于图像分类。从计算机视觉顶会 CVPR 2018 年的论文统计数据来看,以 GAN 为关键词的论文数量占比已接近论文总量的 10%。且纵观 CVPR 2014 至 CVPR 2018,与 GAN 有关的论文数量呈现逐年翻倍的情况。这足以说明 GAN 这项技术正不断地影响着深度学习技术未来的发展。

#### 4.1.2 迁移学习与少样本学习

迁移学习专注于利用已有问题的解决模型求解其他不同但相关问题。比如说,用辨识轿车的模型来提升识别卡车的能力。迁移学习的初衷是节省人工标注样本的时间,让模型可以通过已有的标记数据向未标记数据迁移,从而训练出适用于未标记数据的运算模型。具体来说,迁移学习算法先在一个拥有更大的数据集的任务(源任务)上训练,然后再被迁移为学习另一个只有较少数据集的任务(目标任务)。如果存在一个与目标任务有相关性的任务,且该任务具有丰富的数据,那么可先训练一个针对该任务的模型,然后在我们的目标任务中重用这个模型,或者将这个模型作为我们目标任务模型的训练起始点。这将有利于加速训练的过程,提升神经网络性能。

少样本学习是迁移学习的一个分支,它的产生依赖于人类非常擅长通过少量的样本识别一个新物体。比如,小孩只需要学习几幅图片就能辨别“狗”、“猫”、“牛”等动物。受人类这种快速学习能力的启发,少样本学习在机器学习一定类别的大量数据后,对于新的物体,只需要少量的样本就能迅速完成学习。

#### 4.1.3 自动化机器学习(AutoML)

深度学习算法的性能受许多决策的影响。对于没有丰富计算机技术背景的研究人员,深度学习神经网络的设计总是给他们带来不小的困扰。研究人员需要选择相应的神经网络架构、正则化方法、超参

数等。所有的这些操作对神经网络的性能都有很大的影响。自动机器学习的目标就是使用自动化的方式做出上述的决策。使用者只需提供训练数据,自动机器学习系统就能自动地决定最佳的训练方案。让不同领域的研究人员不必苦恼于学习各种机器学习的算法。

目前,自动化机器学习的实现方式包括:超参数优化(Hyper-parameter Optimization)、元学习(Meta Learning)、神经网络架构搜索(Neural Architecture Search)等。对于超参数优化,常用的方法有网格搜索(Grid Search)、随机搜索(Random Search)和贝叶斯优化。对于元学习,它的主要任务是让机器学习“如何学习”。通过对现有的学习任务之间的性能差异进行系统观测,然后让机器学习已有的经验和元数据,用于更好地执行新的学习任务。从某种意义上来说,元学习的过程蕴含了超参数优化。因为它学习了超参数、流水线构成、神经网络架构、模型构成与元特征等。对于神经网络架构搜索,伴随着深度学习的流行,神经网络的架构变得越来越复杂。利用主观经验来确定合适的神经网络架构难度也越来越大,神经网络架构搜索就是为了解决这个问题。通过定义搜索空间(Search Space)、确定搜索策略(Search Strategy)、性能评价(Performance Estimation Strategy)这三个阶段,机器可根据反馈进行每一轮的架构搜索。自动化机器学习将大幅降低机器学习技术的使用门槛,进一步推动其在光学成像领域中的应用。

## 4.2 物理模型驱动数据与数据驱动物理模型

### 4.2.1 物理模型驱动数据

物理模型驱动是当前深度学习发展的一个重要方向,即在深度学习中嵌入或内蕴特征规则先验,代替单一的纯数据驱动式学习。众所周知,深度学习是一种标准的数据驱动型方法,它将深度网络作为黑箱,依赖于大量数据解决现实问题;而模型驱动方法则是从目标、机理、先验出发,首先形成学习的一个代价函数,然后通过极小化代价函数来解决问题。模型驱动方法的最大优点是只要模型足够精确,解的质量可预期甚至能达到最优,而且求解方法是确定的;模型驱动方法的缺陷是在应用中难以精确建模。模型驱动深度学习方法有效结合了模型驱动和数据驱动方法的优势。

2018 年, Xu 等<sup>[98]</sup>提出一种模型驱动与数据驱动相结合的深度学习方法(图 16),给出了模型驱动深度学习标准流程:1)根据问题,建立模型族

(Family of Models); 2) 根据模型族, 设计算法族 (Family of Algorithms) 并建立算法族的收敛性理论; 3) 将算法族展开 (unfold) 成深度网络并实施深度学习。这种方法将物理模型、逻辑规则作为先验

引入到深层神经网络中, 利用人类意图和领域知识对神经网络模型进行引导, 包括特征规则约束、网络架构设计等, 可有效提升网络大样本学习效率、小样本/零样本学习能力、数据泛化能力。

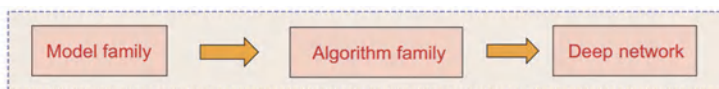


图 16 模型驱动深度学习的方法<sup>[98]</sup>

Fig. 16 Model-driven deep-learning approach<sup>[98]</sup>

#### 4.2.2 数据驱动物理模型

深度学习的方法一般来说只能学习到数据集中已有的知识, 比较擅长于归纳, 而不擅长演绎, 也就是说网络难以推演物理定律, 但是可以在一定约束/条件下拟合物理定律。例如, 概率生成模型<sup>[97]</sup>可用于自然图像的生成, 训练 1000 万张图片后生成的模型可以自动学习到其内部分布, 能够解释给定的训练图片并同时生成新的图片。与庞大的真实数据相比, 概率生成模型的参数量远远小于数据量, 在训练过程中生成模型会被迫去发现数据背后更为简单的统计规律, 从而能够生成这些数据。2017 年, Lin 等<sup>[99]</sup>给出深度学习定性的物理解释(图 17): 1) 基本

的物理学定律都是 2 到 4 阶, 而且拥有对称性等, 这些约束使得解空间变小, 因此 DNN 可以近似得到这个解; 2) 所有物质由简单的基本单元构成, 这个分层结构与 DNN 相似, DNN 的层数越多, 生成的结果越复杂。

然而未知定律必定隐含在数据中, 既然深度学习能够对已知分布/规律进行可靠判别筛选, 那么可以只利用深度学习分析数据并试图找到突出/奇异点, 而不一定要找到特定规则或者新规律。在检测器物理学中, NOvA 中微子实验的研究人员将 CNN 用于粒子识别和分类、粒子轨迹重建、粒子的相互作用分析等<sup>[100]</sup>; 在天体物理学方面, CNN 被用来发现

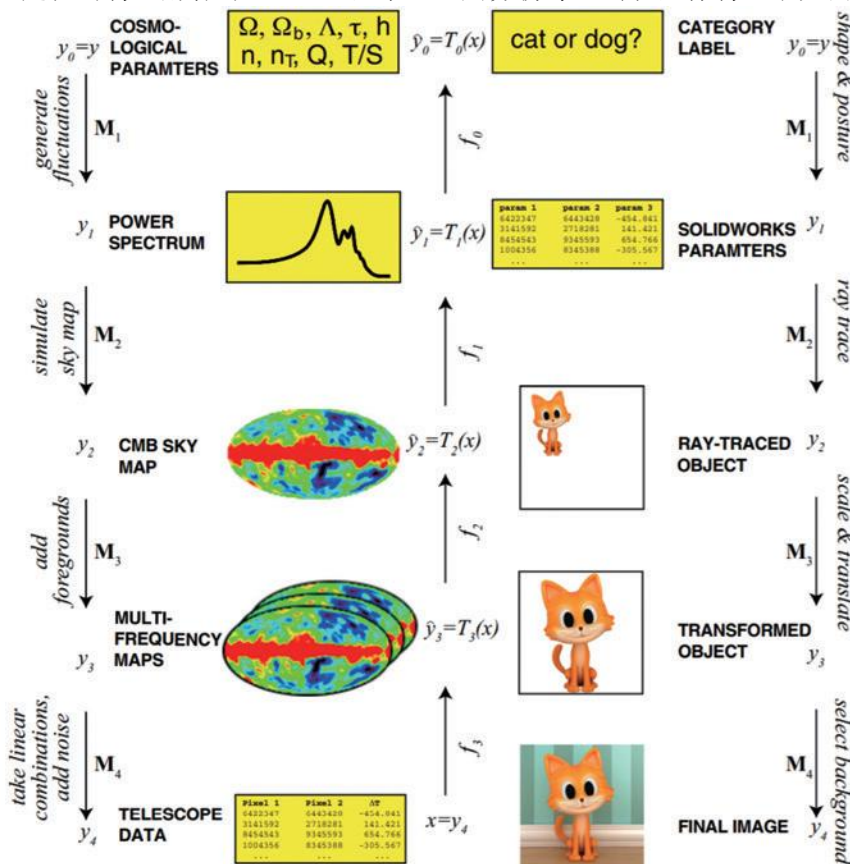


图 17 物理(左)和图像分类(右)关联的因果层次结构<sup>[99]</sup>

Fig. 17 Causal hierarchy structure relevant to physics (left) and image classification (right)<sup>[99]</sup>

引力透镜(引力透镜是指可以扭曲来自它们后面的遥远星系的光的大型天体),加速对望远镜数据扫描以寻找引力透镜扭曲现象的过程<sup>[101]</sup>;在机器视觉应用中也提出了诸多新奇检测方法,采用深度学习对数据中新奇的或未观测到的数据进行检测识别<sup>[102-103]</sup>。

### 4.3 深度学习的可解释性有待进一步探究

所谓可解释性是指在我们需要了解或解决一件事情的时候,可以获得所需要的足够的可以理解的信息。反过来说,如果在一些情境中我们无法得到相应的足够的信息,那么这些事情对人们来说都是不可解释的。正如前文所述其依赖于多层简单的线形运算的组合,最终实现了高度非线性化的高维特征提取功能并获得了极高的模型表现能力。但是,虽然人们创造了准确度极高的网络模型,但最后只得到了一堆看起来“毫无意义”的模型参数

与匹配度非常高的判定结果。尽管如此,但模型本身也意味着知识,我们希望知道模型究竟从数据中“学”到了哪些知识从而支撑该模型进行了最终的决策。

当准备将深度学习应用于某些特定成像领域时,除了获得最终的理想的成像结果,还希望能够了解到神经网络究竟提取了原始输入中的哪些信息,是基于什么形式的运算得到了这个结果。除此之外,不可解释同样意味着模型的“危险性”,图 18 所示为一个非常经典的关于对抗样本的例子,对于一个 CNN 模型,在图片中添加了轻微随机噪声之后熊猫却被判定为长臂猿<sup>[104]</sup>。因此,如何进一步推动深度学习的数据解释性,如何进一步提高深度学习神经网络结果的可溯性、稳定性,都是目前迫切需要解决的问题,这些问题的解决也将为深度学习在计算成像领域打开更广阔的应用空间。

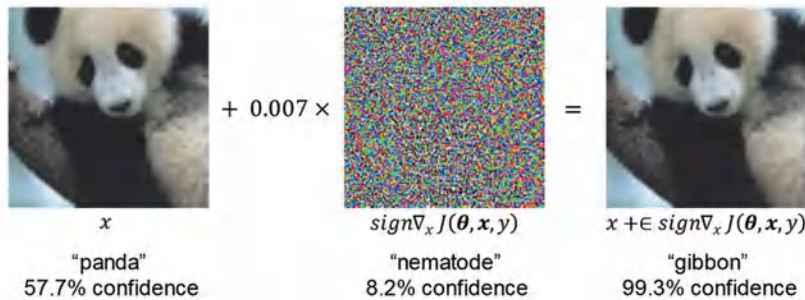


图 18 在熊猫图片中加入轻微随机噪声,CNN 模型将图片识别为长臂猿<sup>[104]</sup>

Fig. 18 After adding slight noise into Panda image, CNN model recognizes image as Gibbon<sup>[104]</sup>

### 4.4 脑神经科学启发的思路值得更多的重视

作为多层神经网络,深度学习是受脑神经科学启发而发展起来的。特别是卷积神经网络,其根源为 Fukushima 在 1980 年提出的认知机模型,而该模型的提出动机就是模拟哺乳动物视觉神经系统,通过逐层提取由简及繁的特征,实现语义逐级抽象的视觉神经通路。在诺贝尔奖获得者 Hubel 和 Wiesel 的共同努力下,该通路从 20 世纪 60 年代起逐渐清晰,为 CNN 的诞生提供了良好的参考。但值得注意的是,生物视觉神经通路极其复杂,神经科学家对初级视觉皮层中简单神经细胞的边缘提取功能是清晰的,对通路后面部分越来越复杂的神经细胞的功能也有一些探索,但对更高层级上的超复杂细胞的功能及其作用机制尚不清晰。这意味着 CNN 等深度模型是否真的能够模拟生物视觉通路还是不得而知的。但可以确定的是,生物神经系统的链接极为复杂,不仅仅有自下而上的前馈和同层递归,更有大量的自上而下的反馈,以及来自其他神

经子系统的外部链接,这些都是目前的深度模型尚未建模的。但无论如何,脑神经科学的进步可以为深度模型的发展提供更多的可能性,这是非常值得关注的。例如,最近越来越多的神经科学研究表明,曾一度被认为功能极为特异化的神经细胞其实具有良好的可塑性,例如,视觉皮层的大量神经细胞在失去视觉处理需求后不久即被重塑,转而处理触觉或其他模态的数据。神经系统的这种可塑性使其面向不同的智能处理任务时具有良好的通用性,这为通用人工智能的发展提供了参照。因此我们可以大胆展望,未来很有可能会出现更加智能的可塑化模型来代替现在的固定结构的深度学习模型。

### 4.5 既要“深度”又要“深入”

深度学习的出现似乎一度改变了计算机视觉与光学成像领域,有计算机视觉领域的研究者曾开玩笑地说道:“深度学习让我感觉到之前一切的所学似乎都白学了。”大家似乎再也不用绞尽脑汁自己去研究背后的数学物理机理、推导模型并进行求解预测。



只要统一地甩给计算机一张网络就可以让它自己去学习具体的模型了,这大大降低了知识学习的成本,减少了手动建模的工作量。理论研究的最终目的不就是建立真实物理世界的数学模型,然后利用模型再去造福世界的吗?现在不需要研究理论也可以建立一个近乎准确的模型了,为什么还要费尽力气深入研究理论?

这看似很有道理,但我们不妨换一个思路去思考一下?一项工作是选择由深度学习去完成还是经典理论算法去完成,归根到底还是在于二者谁能够完成得“更出色”。如果深度学习真的能够“保证”预测出一个近乎准确的模型,输出最为理想的重建结果,我们定能够安心地把这件事交给它而不必再劳

神费力。但谁能保证做到这一点?即使你是这么认为,你是否真的深入了解并实现过那些经典的物理模型驱动方法,并保证能对此结果做出一个不偏不倚的公正评价?你会为让经典算法得到一个最优的结果而通宵调参吗?你是否真的保证你所实现的深度学习算法在训练时没有因数据有意无意的泄露而得到一个过度美化的结果?你会客观而随机地选择测试数据并以此去评价最终的实验结果吗?因此,个人之见是“真正的”深度学习其实并没有(至少现阶段)把科学研究变得更简单,反而拉高了科学研究的门槛,因为他要求研究者不但既要公正且有效地利用“深度”学习这一工具,又要对此领域的研究足够“深入”,以保证真正得到客观而准确的结论(图 19)。

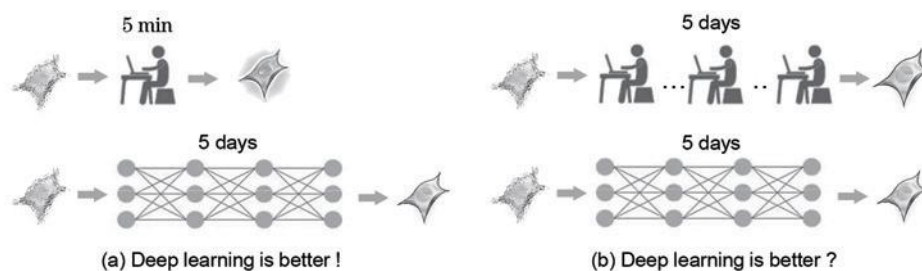


图 19 深度学习与经典理论算法之间的客观公证对比

Fig. 19 Comparison between deep learning and classical theoretical algorithm should be objective

#### 4.6 既要“有所为”又要“有所不为”

现阶段深度学习技术在很多领域(包括计算成像)已经步入了一个爆炸式增长的时代,它似乎成为了该领域研究的未来潮流。一个不能用经典模型解释的现象,只要套上“深度学习”似乎就变得“高大上”起来。这真的让人细思恐极!很多初出茅庐的研究者们似乎发现其实不用扎实的基础理论也可以发表一篇高档次的论文,他们当然会很快地成为该领域的拥护者,并幻想着似乎自己已经走在了该领域研究的最前沿。但熟不知,这大量各色各样的披着深度学习华丽外衣的研究论文背后只不过是一套“模板化”的菜谱,自己学着去做其实未必能做出真正漂亮且有营养的一道美食。

深度学习终究不过是一种基于大量样本数据的统计类方法,统计类方法在推理性的任务面前是不可靠的。在物理学中,想要建立一个模型通常需要三个步骤:1)学习数学物理理论;2)观测真实实验数据;3)基于数学物理理论对观测数据进行建模。看似最重要的建模只是三道工序中的最后一环,实际上大多数困难费事且重要的工作都花在前两步(这也就是为什么大部分人不喜欢花了那么多年时间才读到博士,即使读到了博士还要继续花大量时间读

文章做实验)。深度学习的优势是省去了第一步,简化了第二步,直接到了第三步,这听上去似乎有些难以置信。因为神经网络的参数选择没有理论基础,仅用数据驱动。显然只通过有限(局部的)数据得到的符合这些数据的函数可能是不唯一的,但真实的函数却只有一个,这就注定了仅依赖局部数据驱动预测的模型类似于一种赌博。诚然在计算成像领域某些很难公式化建模的任务上,神经网络取得了巨大的成功,但想要使用神经网络从大量衍射图样中学出菲涅耳衍射定律这样的通用的数学公式却不太现实。

最后值得一提的是,如果你真的对深度学习那么情有独钟,非深度学习不可,那请你在最后的时刻冷静下来思考一下这个问题:“这到底值不值得、不适合用深度学习去做?”深度学习切记不可乱用。在某些问题前面,传统基于物理模型的方法已经能够给出足够简单而精确的解决方案,并不需要深度学习。这就好比一个矩形的面积明明通过长乘宽就可以得到,却非要拿微积分算一遍一样。在有的时候,这种“不必要”有时候其实并不是那么容易被“意识到”。去年一篇来自哈佛和谷歌的用深度学习预测余震位置的 Nature 论文<sup>[105]</sup> 遭受了地震一般地猛

烈质疑,被封为“深度学习的错误用法”(图 20)。这其实并不是单单因为利用深度学习预测地震从常识上讲多么不靠谱,而是有人质疑如果采用文中(具有泄露嫌疑)的训练数据,传统任意一个简单模型,如支持向量机(SVM)、广义相加模型(GAM)等,只需要 1500 行数据都可以得出和原本 470 万行数据的深度神经网络相当的结果<sup>[106]</sup>。另一方面,你必须清楚地认清深度学习技术“它最好能做到多好”很大程度上取决于“你给它的数据有多可靠”。例如,在某些场合下,采集到的真实数据对应的真值无从得知。此时你若想偷个懒,直接用经典算法处理的结果去标注并让网络去建立关联,那么就算再理想再强大的神经网络也至多能和经典算法打个平手,那还为何去劳驾深度学习呢?恰恰是因为这个原因,很多问题并不适合直接用深度学习去解决。例如:干涉测量中低质量包裹相位图的空间解包裹(对应实测数据的正确绝对相位信息难以甚至无法获得),条纹投影中直接从相位到深度的端到端映射(真实世界物体的绝对三维坐标难以通过仪器量测)等。

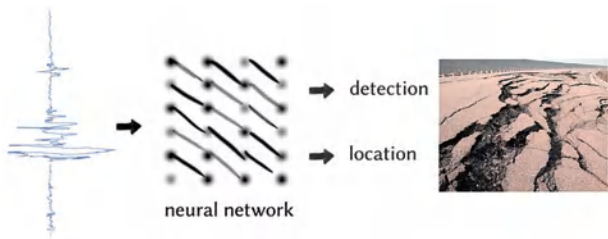


图 20 深度学习被用于预测地震遭到了质疑

Fig. 20 Forecasting earthquake using deep learning hit with rebuttals has been questioned

## 5 结束语

当下快速发展的深度学习技术为计算光学成像的发展打开了一扇新的窗户,它有效提升了传统计算成像技术的信息获取能力,降低了传统计算成像技术对“正向物理模型”或“逆向重构算法”的过度依赖,突破了传统计算成像技术所能够达到的功能/性能疆界,并为这个领域带来很多令人瞩目的开创性研究成果。但同样,深度学习技术在光学成像领域的应用还面临着巨大的挑战。这不仅需要依赖深度学习的专家们进一步去完善这一工具,还需要光学成像的专家们更加理性地去借鉴与使用。相信大家都期望看到今后越来越多的文章不再是各类网络结构与各类计算成像体制简单排列组合般的堆砌,而能够真正大胆地把深度学习下的计算成像所面临的这些挑战毫不避讳地拿到台面上去讨论,甚至有勇

气去挑战它们!只有这样,“深度学习下的计算成像”这一研究领域才不会“昙花一现”,而是能够真正地走得更远……

## 参 考 文 献

- [1] Wikipedia. Computational imaging [EB/OL]. (2019-10-15) [2019-11-20]. [https://en.wikipedia.org/wiki/Computational\\_imaging](https://en.wikipedia.org/wiki/Computational_imaging).
- [2] Wikipedia. AlphaGo versus Lee Sedol [EB/OL]. (2019-11-06) [2019-11-20]. [https://en.wikipedia.org/wiki/AlphaGo\\_versus\\_Lee\\_Sedol](https://en.wikipedia.org/wiki/AlphaGo_versus_Lee_Sedol).
- [3] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115(3): 211-252.
- [4] Rivenson Y, Zhang Y B, Günaydn H, et al. Phase recovery and holographic image reconstruction using deep learning in neural networks[J]. Light: Science & Applications, 2018, 7(2): 17141.
- [5] Wu Y C, Rivenson Y, Zhang Y B, et al. Extended depth-of-field in holographic imaging using deep-learning-based autofocusing and phase recovery[J]. Optica, 2018, 5(6): 704-710.
- [6] Ren Z B, Xu Z M, Lam E Y. Learning-based nonparametric autofocusing for digital holography [J]. Optica, 2018, 5(4): 337-344.
- [7] Wang K Q, Dou J Z, Qian K M, et al. Y-Net: a one-to-two deep learning framework for digital holographic reconstruction [J]. Optics Letters, 2019, 44(19): 4765-4768.
- [8] Nguyen T, Bui V, Lam V, et al. Automatic phase aberration compensation for digital holographic microscopy based on deep learning background detection [J]. Optics Express, 2017, 25 (13): 15043-15057.
- [9] Zhang G, Guan T, Shen Z Y, et al. Fast phase retrieval in off-axis digital holographic microscopy through deep learning [J]. Optics Express, 2018, 26(15): 19388-19405.
- [10] Nguyen T, Xue Y J, Li Y Z, et al. Deep learning approach for Fourier ptychography microscopy [J]. Optics Express, 2018, 26(20): 26470-26484.
- [11] Kappeler A, Ghosh S, Holloway J, et al. Ptychnet: CNN based Fourier ptychography [C] // 2017 IEEE International Conference on Image Processing (ICIP), September 17-20, 2017, Beijing, China. New York: IEEE, 2017: 1712-1716.
- [12] Jiang S W, Guo K K, Liao J, et al. Solving Fourier ptychographic imaging problems via neural network modeling and TensorFlow [J]. Biomedical Optics

- Express, 2018, 9(7): 3306-3319.
- [13] Cheng Y F, Strachan M, Weiss Z, et al. Illumination pattern design with deep learning for single-shot Fourier ptychographic microscopy [J]. Optics Express, 2019, 27(2): 644-656.
- [14] Lü M, Wang W, Wang H, et al. Deep-learning-based ghost imaging [J]. Scientific Reports, 2017, 7: 17865.
- [15] He Y C, Wang G, Dong G X, et al. Ghost imaging based on deep learning [J]. Scientific Reports, 2018, 8: 6469.
- [16] Shimobaba T, Endo Y, Nishitsuji T, et al. Computational ghost imaging using deep learning [J]. Optics Communications, 2018, 413: 147-151.
- [17] Wang H D, Rivenson Y, Jin Y Y, et al. Deep learning enables cross-modality super-resolution in fluorescence microscopy [J]. Nature Methods, 2019, 16(1): 103-110.
- [18] Nehme E, Weiss L E, Michaeli T, et al. Deep-STORM: super-resolution single-molecule microscopy by deep learning [J]. Optica, 2018, 5(4): 458-464.
- [19] Ouyang W, Aristov A, Lelek M, et al. Deep learning massively accelerates super-resolution localization microscopy [J]. Nature Biotechnology, 2018, 36(5): 460-468.
- [20] Rivenson Y, Göröcs Z, Günaydin H, et al. Deep learning microscopy [J]. Optica, 2017, 4(11): 1437-1443.
- [21] Heinrich L, Bogovic J A, Saalfeld S. Deep learning for isotropic super-resolution from non-isotropic 3D electron microscopy [M] // Descoteaux M, Maier-Hein L, Franz A, et al. Medical image computing and computer-assisted intervention-MICCAI 2017. Lecture notes in computer science. Cham: Springer, 2017, 10434: 135-143.
- [22] Wang H, Rivenson Y, Jin Y, et al. Deep learning achieves super-resolution in fluorescence microscopy [J]. Biorxiv, 2018: 309641.
- [23] Fang L Y, Cunefare D, Wang C, et al. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search [J]. Biomedical Optics Express, 2017, 8(5): 2732-2744.
- [24] Lee C S, Baughman D M, Lee A Y. Deep learning is effective for classifying normal versus age-related macular degeneration OCT images [J]. Ophthalmology Retina, 2017, 1(4): 322-327.
- [25] Schlegl T, Waldstein S M, Bogunovic H, et al. Fully automated detection and quantification of macular fluid in OCT using deep learning [J]. Ophthalmology, 2018, 125(4): 549-558.
- [26] Waller L, Tian L. Machine learning for 3D microscopy [J]. Nature, 2015, 523(7561): 416-417.
- [27] Nguyen T, Bui V, Nehmetallah G. Computational optical tomography using 3-D deep convolutional neural networks [J]. Optical Engineering, 2018, 57(4): 043111.
- [28] Lü M, Wang H, Li G, et al. Learning-based lensless imaging through optically thick scattering media [J]. Advanced Photonics, 2019, 1(3): 036002.
- [29] Li S, Deng M, Lee J, et al. Imaging through glass diffusers using densely connected convolutional networks [J]. Optica, 2018, 5(7): 803-813.
- [30] Horisaki R, Takagi R, Tanida J. Learning-based imaging through scattering media [J]. Optics Express, 2016, 24(13): 13738-13743.
- [31] Satat G, Tancik M, Gupta O, et al. Object classification through scattering media with deep learning on time resolved measurement [J]. Optics Express, 2017, 25(15): 17466-17479.
- [32] Cheng S F, Li H H, Luo Y Q, et al. Artificial intelligence-assisted light control and computational imaging through scattering media [J]. Journal of Innovative Optical Health Sciences, 2019, 12(4): 1930006.
- [33] Goy A, Arthur K, Li S, et al. Low photon count phase retrieval using deep learning [J]. Physical Review Letters, 2018, 121(24): 243902.
- [34] Chen C, Chen Q F, Xu J, et al. Learning to see in the dark [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 3291-3300.
- [35] Rivenson Y, Wang H D, Wei Z S, et al. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning [J]. Nature Biomedical Engineering, 2019, 3(6): 466-477.
- [36] Rivenson Y, Liu T R, Wei Z S, et al. PhaseStain: the digital staining of label-free quantitative phase microscopy images using deep learning [J]. Light: Science & Applications, 2019, 8: 23.
- [37] Feng S J, Chen Q, Gu G H, et al. Fringe pattern analysis using deep learning [J]. Advanced Photonics, 2019, 1(2): 025001.
- [38] Yan K T, Yu Y J, Huang C T, et al. Fringe pattern denoising based on deep learning [J]. Optics Communications, 2019, 437: 148-152.
- [39] Wang K Q, Li Y, Qian K M, et al. One-step



- robust deep learning phase unwrapping[J]. *Optics Express*, 2019, 27(10): 15100-15115.
- [40] Feng S J, Zuo C, Yin W, et al. Micro deep learning profilometry for high-speed 3D surface imaging[J]. *Optics and Lasers in Engineering*, 2019, 121: 416-427.
- [41] Luo W, Schwing A G, Urtasun R. Efficient deep learning for stereo matching[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 26-July 1, 2016, Las Vegas, Nevada. New York: IEEE, 2016: 5695-5703.
- [42] Kuznietsov Y, Stuckler J, Leibe B. Semi-supervised deep learning for monocular depth map prediction [C] // *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 2215-2223.
- [43] Gerchberg R W, Saxton W O. A practical algorithm for the determination of phase from image and diffraction plane pictures[J]. *Optik*, 1972, 35(2): 237-250.
- [44] Gerchberg R W. Phase determination from image and diffraction plane pictures in the electron microscope[J]. *Optik*, 1971, 34(3): 275-284.
- [45] Fienup J R, Wackerman C C. Phase-retrieval stagnation problems and solutions[J]. *Journal of the Optical Society of America A*, 1986, 3(11): 1897-1907.
- [46] Seldin J H, Fienup J R. Numerical investigation of the uniqueness of phase retrieval[J]. *Journal of the Optical Society of America A*, 1990, 7(3): 412-427.
- [47] Guizar-Sicairos M, Fienup J R. Understanding the twin-image problem in phase retrieval[J]. *Journal of the Optical Society of America A*, 2012, 29(11): 2367-2375.
- [48] Wackerman C C, Yagle A E. Use of Fourier domain real-plane zeros to overcome a phase retrieval stagnation[J]. *Journal of the Optical Society of America A*, 1991, 8(12): 1898-1904.
- [49] Lu G, Zhang Z, Yu F T S, et al. Pendulum iterative algorithm for phase retrieval from modulus data[J]. *Optical Engineering*, 1994, 33(2): 548-555.
- [50] Takajo H, Takahashi T, Kawanami H, et al. Numerical investigation of the iterative phase-retrieval stagnation problem: territories of convergence objects and holes in their boundaries [J]. *Journal of the Optical Society of America A*, 1997, 14(12): 3175-3187.
- [51] Misell D L. A method for the solution of the phase problem in electron microscopy [J]. *Journal of Physics D: Applied Physics*, 1973, 6(1): L6-L9.
- [52] Zheng G A, Horstmeyer R, Yang C. Wide-field, high-resolution Fourier ptychographic microscopy [J]. *Nature Photonics*, 2013, 7(9): 739-745.
- [53] Horstmeyer R, Chen R Y, Ou X Z, et al. Solving ptychography with a convex relaxation [J]. *New Journal of Physics*, 2015, 17(5): 053044.
- [54] Yeh L H, Dong J, Zhong J S, et al. Experimental robustness of Fourier ptychography phase retrieval algorithms [J]. *Optics Express*, 2015, 23(26): 33214-33240.
- [55] Zuo C, Sun J S, Chen Q. Adaptive step-size strategy for noise-robust Fourier ptychographic microscopy [J]. *Optics Express*, 2016, 24(18): 20724-20744.
- [56] Pittman T B, Shih Y H, Strekalov D V, et al. Optical imaging by means of two-photon quantum entanglement[J]. *Physical Review A*, 1995, 52(5): R3429-R3432.
- [57] Bennink R S, Bentley S J, Boyd R W. "Two-photon" coincidence imaging with a classical source [J]. *Physical Review Letters*, 2002, 89(11): 113601.
- [58] Gatti A, Brambilla E, Bache M, et al. Ghost imaging with thermal light: comparing entanglement and classical correlation[J]. *Physical Review Letters*, 2004, 93(9): 093602.
- [59] Sen P, Chen B, Garg G, et al. Dual photography [C]//*ACM SIGGRAPH 2005 Papers on SIGGRAPH'05*, July 31-August 4, 2005, Los Angeles, California. New York: ACM, 2005: 745-755.
- [60] Dharmpal T, Laska J N, Michael B, et al. A new compressive imaging camera architecture using optical-domain compression [J]. *Proceedings of SPIE*, 2006, 6065: 606509.
- [61] Duarte M F, Davenport M A, Takhar D, et al. Single-pixel imaging via compressive sampling[J]. *IEEE Signal Processing Magazine*, 2008, 25(2): 83-91.
- [62] Sun B, Edgar M P, Bowman R, et al. 3D computational imaging with single-pixel detectors [J]. *Science*, 2013, 340(6134): 844-847.
- [63] Vellekoop I M, Mosk A P. Focusing coherent light through opaque strongly scattering media [J]. *Optics Letters*, 2007, 32(16): 2309-2311.
- [64] Popoff S M, Lerosey G, Carminati R, et al. Measuring the transmission matrix in optics: an approach to the study and control of light propagation in disordered media [J]. *Physical*

- Review Letters, 2010, 104(10): 100601.
- [65] Leith E N, Upatnieks J. Holographic imagery through diffusing media[J]. Journal of the Optical Society of America, 1966, 56(4): 523.
- [66] Yaqoob Z, Psaltis D, Feld M S, et al. Optical phase conjugation for turbidity suppression in biological samples[J]. Nature Photonics, 2008, 2(2): 110-115.
- [67] Bertolotti J, van Putten E G, Blum C, et al. Non-invasive imaging through opaque scattering layers[J]. Nature, 2012, 491(7423): 232-234.
- [68] Yang W Q, Li G W, Situ G H. Imaging through scattering media with the auxiliary of a known reference object[J]. Scientific Reports, 2018, 8: 9614.
- [69] Lü M, Wang H, Li G, et al. Learning-based lensless imaging through optically thick scattering media[J]. Advanced Photonics, 2019, 1(3): 036002.
- [70] Wolf E. Three-dimensional structure determination of semi-transparent objects from holographic data[J]. Optics Communications, 1969, 1(4): 153-156.
- [71] Kak A C, Slaney M. Principles of computerized tomographic imaging[M]. New York: IEEE Press, 2001.
- [72] Haeberlé O, Belkebir K, Giovaninni H, et al. Tomographic diffractive microscopy: basics, techniques and perspectives[J]. Journal of Modern Optics, 2010, 57(9): 686-699.
- [73] Rappaz B, Marquet P, Cuche E, et al. Measurement of the integral refractive index and dynamic cell morphometry of living cells with digital holographic microscopy[J]. Optics Express, 2005, 13(23): 9361-9373.
- [74] Lauer V. New approach to optical diffraction tomography yielding a vector equation of diffraction tomography and a novel tomographic microscope[J]. Journal of Microscopy, 2002, 205(2): 165-176.
- [75] Choi W. Tomographic phase microscopy and its biological applications[J]. 3D Research, 2012, 3(4): 5.
- [76] Charrière F, Marian A, Montfort F, et al. Cell refractive index tomography by digital holographic microscopy[J]. Optics Letters, 2006, 31(2): 178-180.
- [77] Charrière F, Pavillon N, Colomb T, et al. Living specimen tomography by digital holographic microscopy: morphometry of testate amoeba[J]. Optics Express, 2006, 14(16): 7005-7013.
- [78] Choi W, Fang-Yen C, Badizadegan K, et al. Tomographic phase microscopy [J]. Nature Methods, 2007, 4(9): 717-719.
- [79] Sung Y, Choi W, Fang-Yen C, et al. Optical diffraction tomography for high resolution live cell imaging[J]. Optics Express, 2009, 17(1): 266-277.
- [80] Kim K, Yoon H, Diez-Silva M, et al. High-resolution three-dimensional imaging of red blood cells parasitized by *Plasmodium falciparum* and *in situ* hemozoin crystals using optical diffraction tomography[J]. Journal of Biomedical Optics, 2014, 19(1): 011005.
- [81] Cotte Y, Toy F, Jourdain P, et al. Marker-free phase nanoscopy[J]. Nature Photonics, 2013, 7(2): 113-117.
- [82] Devaney A. A filtered backpropagation algorithm for diffraction tomography[J]. Ultrasonic Imaging, 1982, 4(4): 336-350.
- [83] Barty A, Nugent K A, Roberts A, et al. Quantitative phase tomography [J]. Optics Communications, 2000, 175(4/5/6): 329-336.
- [84] Soto J M, Rodrigo J A, Alieva T. Label-free quantitative 3D tomographic imaging for partially coherent light microscopy[J]. Optics Express, 2017, 25(14): 15699-15712.
- [85] Li J, Chen Q, Sun J, et al. Three-dimensional tomographic microscopy technique with multi-frequency combination with partially coherent illuminations[J]. Biomedical Optics Express, 2018, 9(6): 2526-2542.
- [86] Soto J M, Rodrigo J A, Alieva T. Optical diffraction tomography with fully and partially coherent illumination in high numerical aperture label-free microscopy [Invited] [J]. Applied Optics, 2018, 57(1): A205-A214.
- [87] Horstmeyer R, Chung J, Ou X, et al. Diffraction tomography with Fourier ptychography[J]. Optica, 2016, 3(8): 827-835.
- [88] Tian L, Waller L. 3D intensity and phase imaging from light field measurements in an LED array microscope[J]. Optica, 2015, 2(2): 104-111.
- [89] Zuo C, Sun J, Li J, et al. Wide-field high-resolution 3D microscopy with Fourier ptychographic diffraction tomography[J/OL]. (2019-05-26) [2019-11-20]. <https://arxiv.org/abs/1904.09386>.
- [90] Kamilov U S, Papadopoulos I N, Shoreh M H, et al. Learning approach to optical tomography[J]. Optica, 2015, 2(6): 517-522.
- [91] Abbe E. Beiträge zur theorie des Mikroskops und der Mikroskopischen Wahrnehmung[J]. Archiv Für

- Mikroskopische Anatomie, 1873, 9(1): 413-468.
- [92] Hell S W, Wichmann J. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy [J]. Optics Letters, 1994, 19(11): 780-782.
- [93] Betzig E, Patterson G H, Sougrat R, et al. Imaging intracellular fluorescent proteins at nanometer resolution[J]. Science, 2006, 313(5793): 1642-1645.
- [94] Rust M J, Bates M, Zhuang X W. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)[J]. Nature Methods, 2006, 3(10): 793-796.
- [95] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks [M] // Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8689: 818-833.
- [96] Shwartz-Ziv R, Tishby N. Opening the black box of deep neural networks via information [J/OL]. (2017-04-29)[2019-11-20]. <https://arxiv.org/abs/1703.00810>.
- [97] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems, December 8-13, 2014, Montreal, Quebec, Canada. Canada: NIPS, 2014: 2672-2680.
- [98] Xu Z B, Sun J. Model-driven deep-learning [J]. National Science Review, 2018, 5(1): 22-24.
- [99] Lin H W, Tegmark M, Rolnick D. Why does deep and cheap learning work so well? [J]. Journal of Statistical Physics, 2017, 168(6): 1223-1247.
- [100] Behera B. Status of a deep learning based measurement of the inclusive muon neutrino charged-current cross section in the NOvA near detector [J/OL]. (2017-10-10) [2019-11-20]. <https://arxiv.org/abs/1710.03766>.
- [101] Nguyen T Q, Weitekamp D, Anderson D, et al. Topology classification with deep learning to improve real-time event selection at the LHC[J]. Computing and Software for Big Science, 2019, 3(1): 12.
- [102] Perera P, Patel V M. Deep transfer learning for multiple class novelty detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 16-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 11544-11552.
- [103] Perera P, Nallapati R, Xiang B. OCGAN: one-class novelty detection using GANs with constrained latent representations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 16-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 2898-2906.
- [104] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J/OL]. (2015-03-20)[2019-11-20]. <https://arxiv.org/abs/1412.6572>.
- [105] DeVries P M R, Viégas F, Wattenberg M, et al. Deep learning of aftershock patterns following large earthquakes[J]. Nature, 2018, 560(7720): 632-634.
- [106] Machine Learning. Deep learning of aftershock patterns following large earthquakes [R/OL]. [2019-11-20]. [https://www.reddit.com/r/MachineLearning/comments/9bo9i9/r\\_deep\\_learning\\_of\\_aftershock\\_patterns\\_following/](https://www.reddit.com/r/MachineLearning/comments/9bo9i9/r_deep_learning_of_aftershock_patterns_following/).



# Fringe pattern analysis using deep learning

Shijie Feng,<sup>a,b,c</sup> Qian Chen,<sup>a,b,\*</sup> Guohua Gu,<sup>a,b</sup> Tianyang Tao,<sup>a,b</sup> Liang Zhang,<sup>a,b,c</sup> Yan Hu,<sup>a,b,c</sup> Wei Yin,<sup>a,b,c</sup> and Chao Zuo<sup>a,b,c,\*</sup>

<sup>a</sup>Nanjing University of Science and Technology, School of Electronic and Optical Engineering, Nanjing, China

<sup>b</sup>Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense, Nanjing, China

<sup>c</sup>Nanjing University of Science and Technology, Smart Computational Imaging Laboratory (SCILab), Nanjing, China

**Abstract.** In many optical metrology techniques, fringe pattern analysis is the central algorithm for recovering the underlying phase distribution from the recorded fringe patterns. Despite extensive research efforts for decades, how to extract the desired phase information, with the highest possible accuracy, from the minimum number of fringe patterns remains one of the most challenging open problems. Inspired by recent successes of deep learning techniques for computer vision and other applications, we demonstrate for the first time, to our knowledge, that the deep neural networks can be trained to perform fringe analysis, which substantially enhances the accuracy of phase demodulation from a single fringe pattern. The effectiveness of the proposed method is experimentally verified using carrier fringe patterns under the scenario of fringe projection profilometry. Experimental results demonstrate its superior performance, in terms of high accuracy and edge-preserving, over two representative single-frame techniques: Fourier transform profilometry and windowed Fourier transform profilometry.

Keywords: fringe analysis; phase measurement; deep learning.

Received Aug. 22, 2018; accepted for publication Jan. 8, 2019; published online Feb. 28, 2019.

© The Authors. Published by SPIE and CLP under a Creative Commons Attribution 4.0 Unported License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI.

[DOI: [10.1117/1.AP.1.2.025001](https://doi.org/10.1117/1.AP.1.2.025001)]

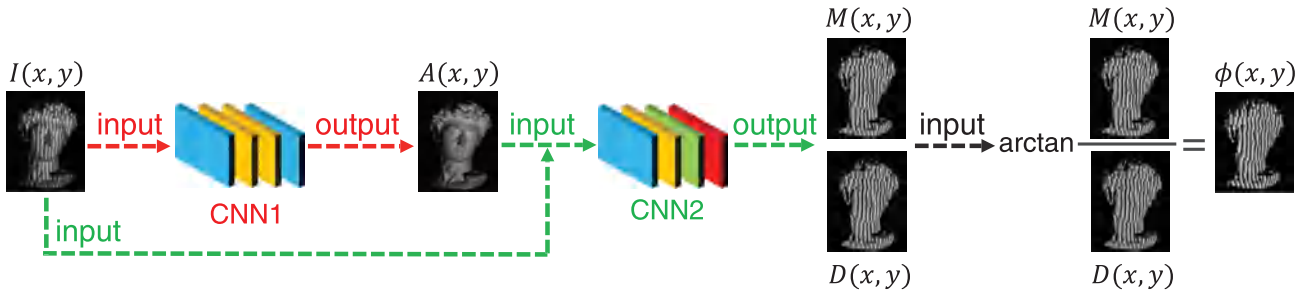
Optical measurement techniques such as holographic interferometry,<sup>1</sup> electronic speckle pattern interferometry,<sup>2</sup> and fringe projection profilometry<sup>3</sup> are quite popular for noncontact measurements in many areas of science and engineering, and have been extensively applied for measuring various physical quantities, such as displacement, strain, surface profile, and refractive index. In all these techniques, the information about the measured physical quantity is stored in the phase of a two-dimensional fringe pattern. The accuracy of measurements carried out by these optical techniques is thus fundamentally dependent on the accuracy with which the underlying phase distribution of the recorded fringe patterns is demodulated.

Over the past few decades, tremendous efforts have been devoted to developing various techniques for fringe analysis. The techniques can be broadly classified into two categories: (1) phase-shifting (PS) methods that require multiple fringe patterns to extract phase information,<sup>4</sup> and (2) spatial phase-demodulation methods that allow phase retrieval from a single

fringe pattern, such as the Fourier transform (FT),<sup>5</sup> windowed Fourier transform (WFT),<sup>6</sup> and wavelet transform (WT) methods.<sup>7</sup> Compared with spatial phase demodulation methods, multiple-shot PS techniques are generally more robust and can achieve pixel-wise phase measurement with higher resolution and accuracy. Furthermore, the PS measurements are quite insensitive to nonuniform background intensity and fringe modulation. Nevertheless, due to their multishot nature, these methods are difficult to apply to dynamic measurements and are more susceptible to external disturbance and vibration. Thus, for many applications, phase extraction from a single fringe pattern is desired, which falls under the purview of spatial fringe analysis. In contrast to PS techniques where the phase map is demodulated on a pixel-by-pixel basis, phase estimation at a pixel according to spatial methods is influenced by the pixel's neighborhood, or even all pixels in the fringe pattern, which provides better tolerance to noise, yet at the expense of poor performance around discontinuities and isolated regions in the phase map.<sup>8,9</sup>

Deep learning is a powerful machine learning technique that employs artificial neural networks with multiple layers of

\*Address all correspondence to Qian Chen, E-mail: [chenqian@njust.edu.cn](mailto:chenqian@njust.edu.cn); Chao Zuo, E-mail: [zuocho@njust.edu.cn](mailto:zuocho@njust.edu.cn)



**Fig. 1** Flowchart of the proposed method where two convolutional networks (CNN1 and CNN2) and the arctangent function are used together to determine the phase distribution. For CNN1 (in red), the input is the fringe image  $I(x, y)$ , and the output is the estimated background image  $A(x, y)$ . For CNN2 (in green), the inputs are the fringe image  $I(x, y)$  and the background image  $A(x, y)$  predicted by CNN1, and the outputs are the numerator  $M(x, y)$  and the denominator  $D(x, y)$ . The numerator and denominator are then fed into the arctangent function to calculate the phase  $\phi(x, y)$ .

increasingly richer functionality and has shown great success in numerous applications for which data are abundant.<sup>10,11</sup> In this letter, we demonstrate experimentally for the first time, to our knowledge, that the use of a deep neural network can substantially enhance the accuracy of phase demodulation from a single fringe pattern. To be concrete, the networks are trained to predict several intermediate results that are useful for the calculation of the phase of an input fringe pattern. During the training of the networks, we capture PS fringe images of various scenes to generate the training data. The training label (ground truth) of each training datum is a pair of intermediate results calculated from the PS algorithm. After appropriate training, the neural network can blindly take only one input fringe pattern and output the corresponding estimates of these intermediate results with high fidelity. Finally, a high-accuracy phase map can be retrieved through the arctangent function with the intermediate results estimated through deep learning. Experimental results on fringe projection profilometry confirm that this deep-learning-based method is able to substantially improve the quality of the retrieved phase from only a single fringe pattern, compared to state-of-the-art methods.

Here, the network configuration is inspired by the basic process of most phase demodulation techniques, which is briefly recalled as follows. The mathematical form of a typical fringe pattern can be represented as

$$I(x, y) = A(x, y) + B(x, y) \cos \phi(x, y), \quad (1)$$

where  $I(x, y)$  is the intensity of the fringe pattern,  $A(x, y)$  is the background intensity,  $B(x, y)$  is the fringe amplitude, and  $\phi(x, y)$  is the desired phase distribution. Here,  $x$  and  $y$  refer to the pixel coordinates. In most phase demodulation techniques, the background intensity  $A(x, y)$  is regarded as a disturbance term and should be removed from the total intensity. Then a wrapped phase map is recovered from an inverse trigonometric function whose argument is a ratio for which the numerator characterizes the phase sine [ $\sin \phi(x, y)$ ] and the denominator characterizes the phase cosine [ $\cos \phi(x, y)$ ]:

$$\phi(x, y) = \arctan \frac{M(x, y)}{D(x, y)} = \arctan \frac{cB(x, y) \sin \phi(x, y)}{cB(x, y) \cos \phi(x, y)}, \quad (2)$$

where  $c$  is a constant dependent on the phase demodulation algorithm (e.g., in FT  $c = 0.5$ , in  $N$ -step PS  $c = N/2$ ), and  $M(x, y)$  and  $D(x, y)$  represent the shorthand for the numerator and denominator terms, respectively. Note that the signs of  $M(x, y)$  and  $D(x, y)$  can be further used to uniquely define a quadrant for each calculation of  $\phi(x, y)$ . With the four-quadrant phasor space, the phase values at each point can be determined modulo  $2\pi$ .

In order to emulate the process above, two different convolutional neural networks (CNN) are constructed, which are connected cascadedly according to Fig. 1. The first convolutional neural network (CNN1) uses the raw fringe pattern  $I(x, y)$  as input and estimates the background intensity  $A(x, y)$  of the fringe pattern. With the estimated background image  $A(x, y)$  and the original fringe image  $I(x, y)$ , the second convolutional neural network (CNN2) is trained to predict the numerator  $M(x, y)$  and the denominator  $D(x, y)$  of the arctangent function, which are fed into the subsequent arctangent function [Eq. (2)] to obtain the final phase distribution  $\phi(x, y)$ .

To generate the ground truth data used as the label to train the two convolutional neural networks, the phase retrieval is achieved by using the  $N$ -step PS method. The corresponding  $N$  PS fringe patterns acquired can be represented as

$$I_n(x, y) = A(x, y) + B(x, y) \cos[\phi(x, y) - \delta_n], \quad (3)$$

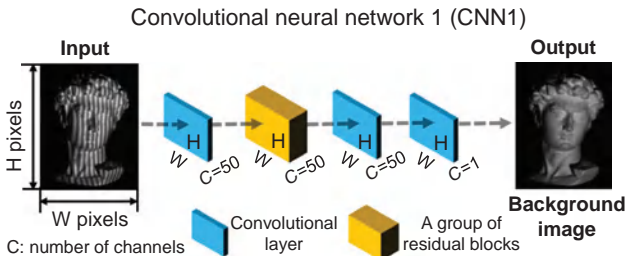
where the index  $n = 0, 1, \dots, N-1$ , and  $\delta_n$  is the phase shift that equals  $\frac{2\pi n}{N}$ . With the orthogonality of trigonometric functions, the background intensity can be obtained as

$$A(x, y) = \frac{1}{N} \sum_{n=0}^{N-1} I_n(x, y). \quad (4)$$

With the least square method, the phase can be calculated as

$$\phi(x, y) = \arctan \frac{\sum_{n=0}^{N-1} I_n(x, y) \sin \delta_n}{\sum_{n=0}^{N-1} I_n(x, y) \cos \delta_n}. \quad (5)$$

Thus, the numerator and the denominator of the arctangent function in Eq. (2) can be expressed as



**Fig. 2** Schematic of CNN1, which is composed of convolutional layers and several residual blocks.

$$M(x, y) = \sum_{n=1}^{N-1} I_n(x, y) \sin \delta_n = \frac{N}{2} B(x, y) \sin \phi(x, y), \quad (6)$$

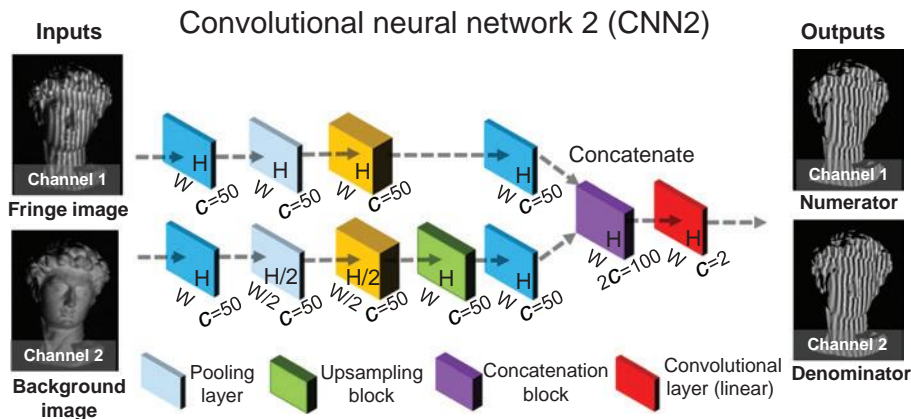
$$D(x, y) = \sum_{n=0}^{N-1} I_n(x, y) \cos \delta_n = \frac{N}{2} B(x, y) \cos \phi(x, y). \quad (7)$$

The expressions above show that the numerator  $M(x, y)$  and the denominator  $D(x, y)$  are closely related to the original fringe pattern in Eq. (1) through a quasilinear relationship with the background image  $A(x, y)$ . Thus,  $M(x, y)$  and  $D(x, y)$  can be learned by deep neural networks with ease given the knowledge of  $A(x, y)$ , which justifies our network. It should be noted that the simple input–output network structure [linking fringe pattern  $I(x, y)$  to phase  $\phi(x, y)$  directly] performs poorly in our case since it is difficult to follow the phase wraps ( $2\pi$  jumps) in the phase map precisely. Therefore, instead of estimating the phase directly, our deep neural networks are trained to predict the intermediate results, i.e., the numerator and the denominator of the arctangent function in Eq. (2), to obtain a better phase estimate. To further validate the superiority of the proposed method, an ablation analysis is presented in [Sec. 6 of the Supplementary Material](#), in which three methods that (1) estimate the phase  $\phi(x, y)$  directly; (2) predict  $D(x, y)$  and  $M(x, y)$  without  $A(x, y)$ ; and (3) calculate  $A(x, y)$ ,  $D(x, y)$ , and  $M(x, y)$  simultaneously are compared experimentally. The comparative

results indicate that our method is more advantageous in phase reconstruction accuracy than others.

To further reveal the internal structure of the two networks, the diagrams of the two convolutional neural networks are shown in Figs. 2 and 3. The labeled dimensions of the layers or the blocks show the size of their output data. The input of CNN1 is a raw fringe pattern with  $W \times H$  pixels. It is then successively processed by a convolutional layer, a group of residual blocks (containing four residual blocks) and two convolutional layers. The last layer estimates the gray values of the background image. With the predicted background intensity and the raw fringe pattern, as shown in Fig. 3, CNN2 calculates the numerator and denominator terms. In CNN2, the input data having two channels are downsampled by  $\times 1$  and  $\times 2$  in two different paths. In the second path, the data are first downsampled for a high-level perception and then upsampled to match the original dimensions. With the two-scale data flow paths, the network can perceive more surface details for both the numerator and the denominator. We provide additional details about the architectures of our networks in [Supplementary Sec. 3](#).

The performance of the proposed approach was demonstrated under the scenario of fringe projection profilometry. The experiment consisted of two steps: training and testing. In order to obtain the ground truth of training data, 12-step PS patterns with spatial frequency  $f = 160$  were created and projected by our projector (DLP 4100, Texas Instruments) onto various objects. The fringe images were captured simultaneously by a CMOS camera (V611, Vision Research Phantom) of 8-bit pixel depth and of resolution  $1280 \times 800$ . Training objects with different materials, colors, and reflectivity are preferable to enhance the generalization capability of the proposed method. Also, analogous to traditional approaches of fringe analysis that require fringes with enough signal-to-noise ratio or without saturated pixels, the proposed method prefers objects without very dark or shiny surfaces. Our training dataset is collected from 80 scenes. It consists of 960 fringe patterns and the corresponding ground truth data that are obtained by a 12-step PS method (see [Supplementary Secs. 1 and 2](#) for details about the optical setup and the collection of training data). Since one of the inputs of CNN2 is the output of CNN1, CNN1 was trained first and CNN2 was trained with the predicted background intensities and captured fringe patterns. These two



**Fig. 3** Schematic of CNN2, which is more sophisticated than CNN1 and further includes two pooling layers, an upsampling layer, a concatenation block, and a linearly activated convolutional layer.

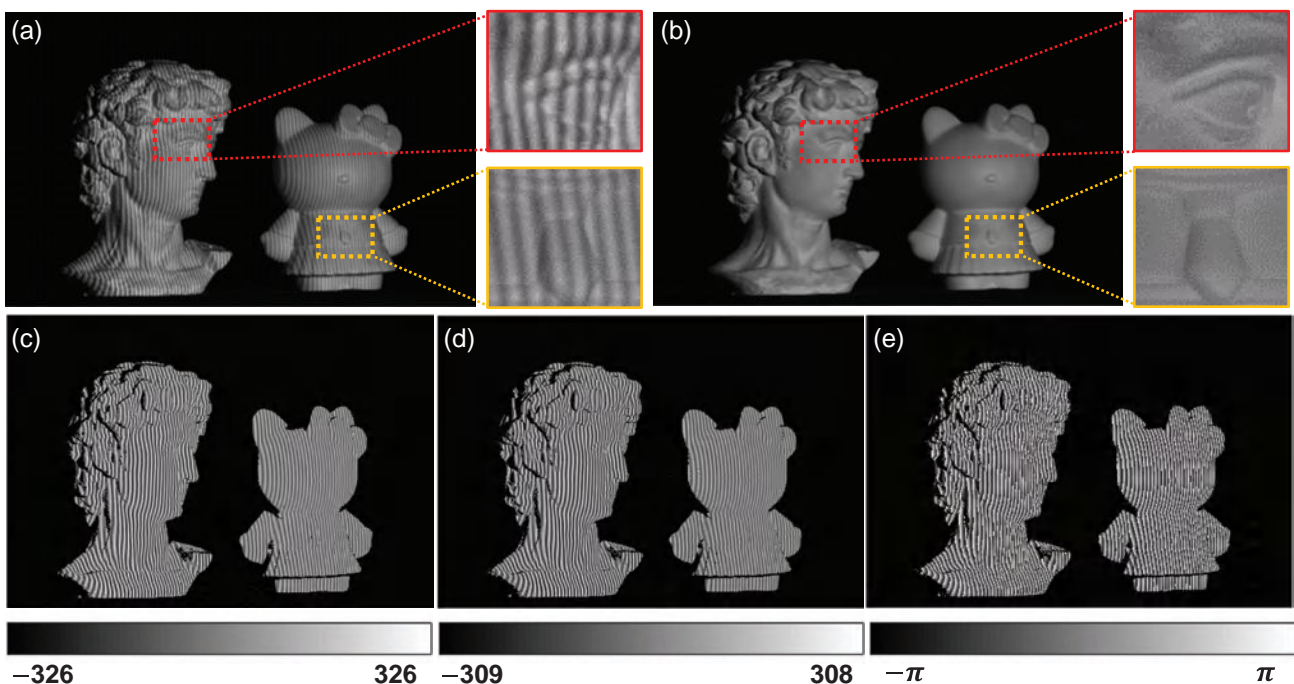


neural networks were implemented using the TensorFlow framework (Google) and were computed on a GTX Titan graphics card (NVIDIA). To monitor during training the accuracy of the neural networks on data that they have never seen before, we created a validation set including 144 fringe images from 12 scenes that are separate from the training scenarios. Additional details on the training of our networks are provided in [Supplementary Sec. 3](#).

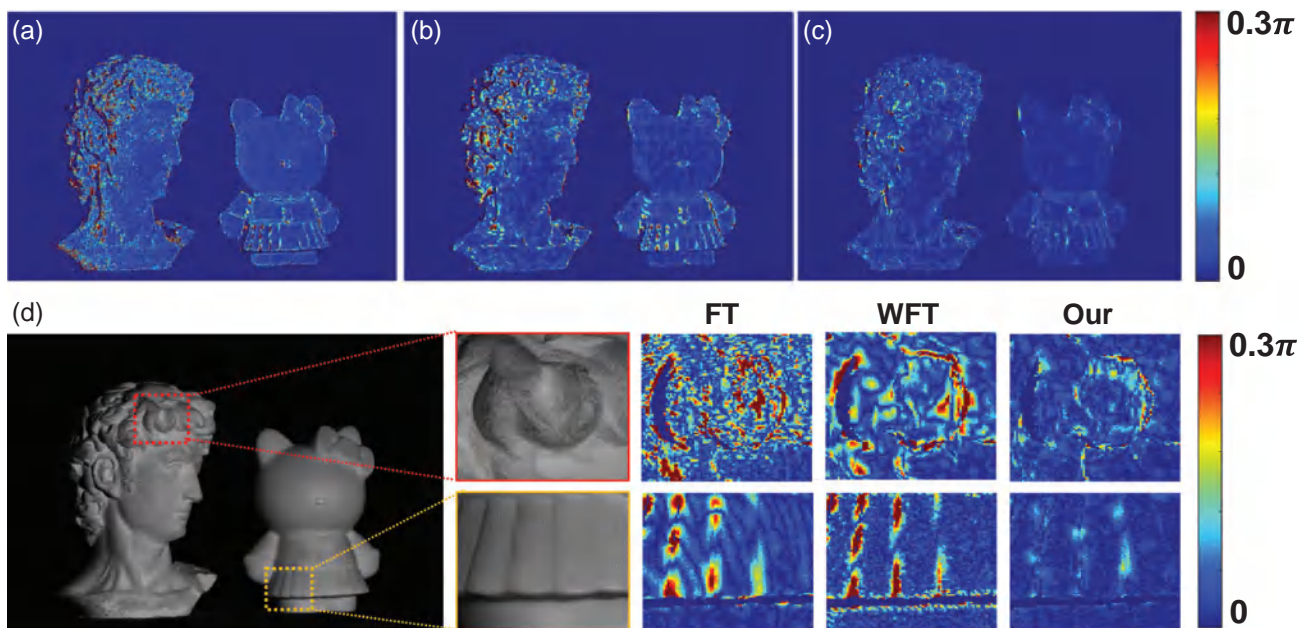
To test the trained neural networks versus classic single-frame approaches (i.e., FT<sup>5</sup> and WFT<sup>6</sup>), we measured a scene containing two isolated plaster models, as shown in Fig. 4(a). Compared with the right model, the left one has a more complex surface, e.g., the curly hair and the high-bridged nose. Note that this scenario was never seen by our neural networks during the training stage. The trained CNN1 using Fig. 4(a) as an input predicted a background intensity as shown in Fig. 4(b). From the enlarged views, we can see that the fringes have been removed completely through the deep neural network. Then, the trained CNN2 took the fringe pattern and the predicted background intensity as inputs and estimated the numerator  $M(x, y)$  and the denominator  $D(x, y)$ ; results are shown in Figs. 4(c) and 4(d), respectively. The phase was calculated by Eq. (2) and is shown in Fig. 4(e). In order to evaluate the quality of the estimated phase more easily, we unwrapped it by multifrequency temporal phase unwrapping,<sup>12</sup> in which additional phase maps of fringe patterns of different frequencies were computed with PS algorithm and were then used to unwrap the phase obtained through deep learning. To demonstrate the accuracy of the unwrapped phase, the phase error was calculated against a reference phase map, which was obtained by the 12-step PS method and was unwrapped with the same strategy.

Figures 5(a)–5(c) show the overall absolute phase errors of these approaches, and the calculated mean absolute error (MAE) of each method is listed in Table 1. Note that the adjustable parameters (e.g., the window size) in FT and WFT have been carefully tuned in order to get the best results possible. The result of FT shows the most prominent phase distortion as well as the largest MAE of 0.20 rad. WFT performed better than FT, with fewer errors for both models (MAE 0.19 rad). Among these approaches, the proposed deep-learning-based method demonstrates the least error, which is 0.087 rad. Furthermore, after the training stage, our method becomes fully automatic and does not require a manual parameter search to optimize its performance. To compare the error maps in detail, the phase errors of two complex areas are presented in Fig. 5(d): the hair of the left model and the skirt of the right one. From Fig. 5(d), obvious errors can be observed in the results of FT and WFT, which are mainly concentrated in the boundaries or abrupt depth-changing regions. By contrast, our approach greatly reduced the phase distortion, demonstrating its significantly improved performance in measuring objects with discontinuities and isolated complex surfaces. To further test and compare the performance of our technique with FT and WFT, [Sec. 7 of the Supplementary Material](#) details the measurements of more kinds of objects, which also shows that our method is superior to FT and WFT in terms of phase reconstruction accuracy.

For a more intuitive comparison, we converted the unwrapped phase into 3-D rendered geometries through stereo triangulation,<sup>13</sup> as shown in Fig. 6. Figure 6(a) shows that the reconstructed result from FT features many grainy distortions, which are mainly due to the inevitable spectral leakage and overlapping in the frequency domain. Compared with FT, the



**Fig. 4** Testing using the trained networks on a scene that is not present in the training phase. (a) Input fringe image  $I(x, y)$ , (b) background image  $A(x, y)$  predicted by CNN1, (c) and (d) numerator  $M(x, y)$  and denominator  $D(x, y)$  estimated by CNN2, (e) phase  $\phi(x, y)$  calculated with (c) and (d).

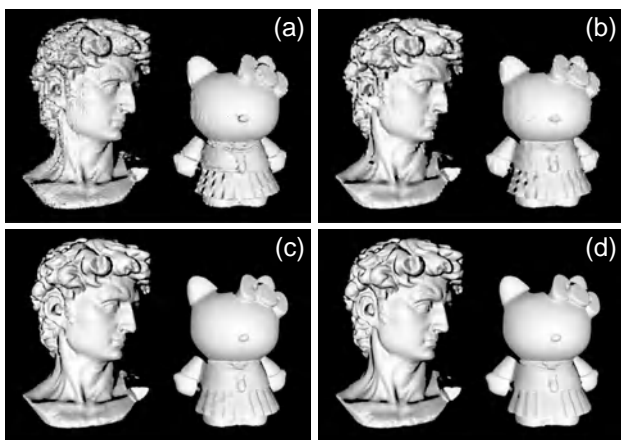


**Fig 5** Comparison of the phase error of different methods: (a) FT, (b) WFT, (c) our method, and (d) magnified views of the phase error for two selected complex regions.

**Table 1** Phase error of FT, WFT, and our method.

Method	FT	WFT	Our
MAE (rad)	0.20	0.19	0.087

WFT reconstructed the objects with more smooth surfaces but failed to preserve the surface details, e.g., the eyes of the left model and the wrinkles of the skirt of the right model, as can be seen in Fig. 6(b). Among these reconstructions, the deep-learning-based approach yielded the highest-quality 3-D reconstruction [Fig. 6(c)], which almost visually reproduced

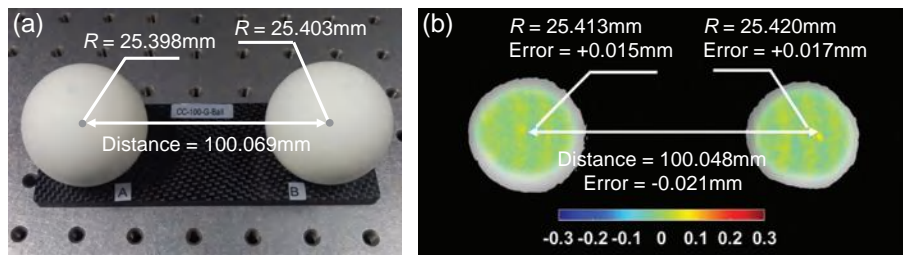


**Fig 6** Comparison of the 3-D reconstruction results for different methods: (a) FT, (b) WFT, (c) our method, and (d) ground truth obtained by the 12-step PS profilometry.

the ground truth data [Fig. 6(d)] where 12-step PS fringe patterns were used.

It should be further mentioned that, in the above experiment, the carrier frequency of the fringe pattern is an essential factor affecting the performance of FT and WFT, which was set sufficiently high ( $f = 160$ ) in order to yield results with reasonable accuracy and spatial resolution. However, it can be troublesome for them to analyze the fringe patterns where the carrier frequency is relatively low. As shown in Sec. 4 of the Supplementary Material, the reconstruction quality of FT and WFT degraded to 0.28 and 0.26 rad when the carrier frequency was reduced to 60. By contrast, our method produced a consistently more accurate phase reconstruction with the phase error of 0.10 rad. In addition, to find appropriate patterns, we suggest choosing a fringe with high frequency and adequate density, but which will not affect the contrast of captured patterns. Section 5 of the Supplementary Material provides detailed information on the selection of the optimal frequency for the network training.

Finally, to quantitatively determine the accuracy of the learned phase after converting to the desired physical quantity, i.e., 3-D shape of the object, we measured a pair of standard ceramic spheres whose shapes have been calibrated based on a coordinate measurement machine. Figure 7(a) shows the tested ceramic spheres. Their radii are 25.398 and 25.403 mm, respectively, and their center-to-center distance is 100.069 mm. We calculated the 3-D point cloud from the phase obtained by the proposed method and then fitted the 3-D points into the sphere model. The reconstructed result is shown in Fig. 7(b), where the “jet” colormap is used to represent the data values of reconstruction errors. The radii of reconstructed spheres are 25.413 and 25.420 mm, with deviations of 15 and 17  $\mu\text{m}$ , respectively. The measured center-to-center distance is 100.048 mm, with an error of  $-21 \mu\text{m}$ . As the measured dimensions are very close to the ground truth, this experiment



**Fig 7** Quantitative analysis of the reconstruction accuracy of the proposed method. (a) Measured objects: a pair of standard spheres and (b) 3-D reconstruction result showing the measurement accuracy.

demonstrates that our method not only provides reliable phase information using only a single fringe pattern but also facilitates high-accuracy single-shot 3-D measurements.

In this letter, we have demonstrated how deep learning significantly improves the accuracy of phase demodulation from a single fringe pattern. Compared with existing single-frame approaches, this deep-learning-based technique provides a framework in fringe analysis by rapidly predicting the background image and estimating the numerator and the denominator for the arctangent function, resulting in high-accuracy edge-preserving phase reconstruction without any human intervention. The effectiveness of the proposed method has been verified using carrier fringe patterns under the scenario of fringe projection profilometry. We believe that, after appropriate training with different types of data, the proposed network framework or its derivation should also be applicable to other forms of fringe patterns (e.g., exponential phase fringe patterns or closed fringe patterns) and other phase measurement techniques for immensely promising applications.

### Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (61722506, 61705105, and 11574152), the National Key R&D Program of China (2017YFF0106403), the Outstanding Youth Foundation of Jiangsu Province (BK20170034), the China Postdoctoral Science Foundation (2017M621747), and the Jiangsu Planned Projects for Postdoctoral Research Funds (1701038A).

### References

1. T. Kreis, *Handbook of Holographic Interferometry: Optical and Digital Methods*, John Wiley & Sons, Hoboken, New Jersey (2006).
2. P. K. Rastogi, *Digital Speckle Pattern Interferometry & Related Techniques*, John Wiley & Sons, Hoboken, New Jersey (2000).
3. S. S. Gorthi and P. Rastogi, "Fringe projection techniques: whither we are?," *Opt. Lasers Eng.* **48**(2), 133–140 (2010).
4. C. Zuo et al., "Phase shifting algorithms for fringe projection profilometry: a review," *Opt. Lasers Eng.* **109**, 23–59 (2018).
5. X. Su and Q. Zhang, "Dynamic 3-D shape measurement method: a review," *Opt. Lasers Eng.* **48**(2), 191–204 (2010).
6. Q. Kema, "Two-dimensional windowed Fourier transform for fringe pattern analysis: principles, applications and implementations," *Opt. Lasers Eng.* **45**(2), 304–317 (2007).
7. J. Zhong and J. Weng, "Spatial carrier-fringe pattern analysis by means of wavelet transform: wavelet transform profilometry," *Appl. Opt.* **43**(26), 4993–4998 (2004).
8. L. Huang et al., "Comparison of Fourier transform, windowed Fourier transform, and wavelet transform methods for phase extraction from a single fringe pattern in fringe projection profilometry," *Opt. Lasers Eng.* **48**(2), 141–148 (2010).
9. Z. Zhang et al., "Comparison of Fourier transform, windowed Fourier transform, and wavelet transform methods for phase calculation at discontinuities in fringe projection profilometry," *Opt. Lasers Eng.* **50**(8), 1152–1160 (2012).
10. A. Sinha et al., "Lensless computational imaging through deep learning," *Optica* **4**(9), 1117–1125 (2017).
11. Y. Rivenson et al., "Phase recovery and holographic image reconstruction using deep learning in neural networks," *Light: Sci. Appl.* **7**, 17141 (2018).
12. C. Zuo et al., "Temporal phase unwrapping algorithms for fringe projection profilometry: a comparative review," *Opt. Lasers Eng.* **85**, 84–103 (2016).
13. C. Zuo et al., "High-speed three-dimensional profilometry for multiple objects with complex shapes," *Opt. Express* **20**(17), 19493–19510 (2012).

**Shijie Feng** received his PhD in optical engineering at Nanjing University of Science and Technology. He is an associate professor at Nanjing University of Science and Technology. His research interests include phase measurement, high-speed 3D imaging, fringe projection, machine learning, and computer vision.

**Qian Chen** received his BS, MS, and PhD degrees from the School of Electronic and Optical Engineering, Nanjing University of Science and Technology. He is currently a professor and a vice principal of Nanjing University of Science and Technology. He has been selected as Changjiang Scholar Distinguished Professor. He has broad research interests around photoelectric imaging and information processing, and has authored more than 200 journal papers. His research team develops novel technologies and systems for mid-/far-wavelength infrared thermal imaging, ultrahigh sensitivity low-light-level imaging, noninterferometric quantitative phase imaging, and high-speed 3D sensing and imaging, with particular applications in national defense, industry, and bio-medicine. He is a member of SPIE and OSA.

**Guohua Gu** received his BS, MS, and PhD degrees at Nanjing University of Science and Technology. He is a professor at Nanjing University of Science and Technology. His research interests include optical 3D measurement, fringe projection, infrared imaging, and ghost imaging.

**Tianyang Tao** received his BS degree at Nanjing University of Science and Technology. He is a fourth-year PhD student at Nanjing University of Science and Technology. His research interests include multiview optical 3D imaging, computer vision, and real-time 3D measurement.

**Liang Zhang** received his BS and MS degrees at Nanjing University of Science and Technology. He is a fourth-year PhD student at Nanjing



University of Science and Technology. His research interests include high-dynamic-range 3D imaging and computer vision.

**Yan Hu** received his BS degree at Wuhan University of Technology. He is a fourth-year PhD student at Nanjing University of Science and Technology. His research interests include microscopic imaging, 3D imaging, and system calibration.

**Wei Yin** is a second-year PhD student at Nanjing University of Science and Technology. His research interests include deep learning, high-speed 3D imaging, fringe projection, and computational imaging.

**Chao Zuo** received his BS and PhD degrees from Nanjing University of Science and Technology (NJUST) in 2009 and 2014, respectively. He was a research assistant at Centre for Optical and Laser Engineering, Nanyang Technological University from 2012 to 2013. He is now a professor at the Department of Electronic and Optical Engineering and the principal investigator of the Smart Computational Imaging Laboratory ([www.scilaboratory.com](http://www.scilaboratory.com)), NJUST. He has broad research interests around computational imaging and high-speed 3D sensing, and has authored over 100 peer-reviewed journal publications. He has been selected into the Natural Science Foundation of China (NSFC) for Excellent Young Scholars and the Outstanding Youth Foundation of Jiangsu Province, China. He is a member of SPIE, OSA, and IEEE.



# Deep-learning-based fringe-pattern analysis with uncertainty estimation

SHIJIE FENG,<sup>1,2,3</sup> CHAO ZUO,<sup>1,2,\*</sup> YAN HU,<sup>1,2</sup> YIXUAN LI,<sup>1,2</sup> AND QIAN CHEN<sup>2,4</sup>

<sup>1</sup>Smart Computational Imaging (SCI) Laboratory, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China

<sup>2</sup>Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China

<sup>3</sup>e-mail: shijiefeng@njust.edu.cn

<sup>4</sup>e-mail: chenqian@njust.edu.cn

\*Corresponding author: zuochao@njust.edu.cn

Received 16 June 2021; revised 1 November 2021; accepted 1 November 2021; published 23 November 2021

Deep learning has gained increasing attention in the field of optical metrology and demonstrated great potential in solving a variety of optical metrology tasks, such as fringe analysis and phase unwrapping. However, deep neural networks cannot always produce a provably correct solution, and the prediction error cannot be easily detected and evaluated unless the ground-truth is available. This issue is critical for optical metrology, as the reliability and repeatability of the measurement are of major importance for high-stakes scenarios. In this paper, for the first time to our knowledge, we demonstrate that a Bayesian convolutional neural network (BNN) can be trained to not only retrieve the phase from a single fringe pattern but also produce uncertainty maps depicting the pixel-wise confidence measure of the estimated phase. Experimental results show that the proposed BNN can quantify the reliability of phase predictions under conditions of various training dataset sizes and never-before-experienced inputs. Our work allows for making better decisions in deep learning solutions, paving a new way to reliable and practical learning-based optical metrology. © 2021 Optical Society of America under the terms of the OSA Open Access Publishing Agreement

<https://doi.org/10.1364/OPTICA.434311>

Fringe-pattern analysis is key to many optical metrology applications [1], such as optical interferometry, fringe projection profilometry, digital holography, moiré interferometry, shearography, and corneal topography. The purpose of the fringe-pattern analysis is to extract the underlying phase information of test objects from one or several fringe pattern(s). Normally, a fringe pattern  $I$  can be expressed as

$$I(x, y) = A(x, y) + B(x, y) \cos \varphi(x, y), \quad (1)$$

where  $(x, y)$  is the pixel coordinate,  $A$  is the background signal,  $B$  is the modulation, and  $\varphi$  is the phase of test objects. As  $A$  and  $B$  are unknown, it is an ill-posed problem to extract  $\varphi$  if only one fringe image is at hand. Single-shot phase demodulation approaches, e.g., Fourier transform profilometry (FTP) [2], resort to the assistance of a spatial carrier to handle the ill-posed issue. Although they are of high efficiency, they are susceptible to complex surfaces that

can easily cause spectral aliasing during the phase demodulation. On the contrary, multi-shot phase demodulation approaches, such as phase-shifting (PS) algorithms [3], can carry out pixel-wise phase measurements with high accuracy. However, they are fragile for disturbances and vibrations due to the limited efficiency resulting from the multi-frame nature.

Recently, the deep learning technique has been introduced to the fringe-pattern analysis [4]. It is reported that the phase information can be extracted from a single fringe pattern with substantially enhanced phase accuracy for complex objects by a trained deep neural network (DNN). Therefore, the learning-based fringe analysis has great potential in realizing high-efficiency and high-accuracy phase demodulation. However, as most DNNs are driven by data completely, the reasoning process is quite different from that of a traditional physical model. Actually, when the training data are insufficient or the testing data are rare, the output of DNN may not be reliable enough. A recent example in computer vision has shown a disastrous prediction where an image classification network mistakenly identified two African Americans as gorillas, giving rise to concerns of racial discrimination [5]. Therefore, *how to trust the prediction of a DNN is still a big challenge*.

For the task of single-shot fringe-pattern analysis, the uncertainty estimation of the predicted phase is indispensable as it is an ill-posed problem to retrieve the phase from Eq. (1) with a single image. Inspired by recent successful applications of Bayesian deep learning approaches [6], we demonstrate for the first time, to the best of our knowledge, that a Bayesian convolutional neural network (BNN) can be trained to not only demodulate the phase from a single fringe pattern, but also evaluate two uncertainties of the prediction. They are the data uncertainty and the model uncertainty. The data uncertainty is also referred to as the aleatoric uncertainty that can quantify the randomness of the prediction due to the noise and data imperfection. The model uncertainty can be referred to as the epistemic uncertainty, which captures the robustness and the uncertainty of the model. The proposed BNN is easy to construct and can be extended to traditional DNNs readily. Experimental results on fringe projection profilometry show that the uncertainty maps predicted by BNN can indicate the actual error distribution faithfully in the absence of standard reference data.

According to Eq. (1), the phase can be retrieved by

$$\varphi(x, y) = \arctan \frac{M(x, y)}{D(x, y)} = \arctan \frac{c B(x, y) \sin \varphi(x, y)}{c B(x, y) \cos \varphi(x, y)}, \quad (2)$$

where the numerator  $M(x, y)$  characterizes the phase sine  $[\sin \varphi(x, y)]$  and the denominator  $D(x, y)$  characterizes the phase cosine  $[\cos \varphi(x, y)]$ .  $c$  is a constant parameter that depends on the used phase demodulation approach [2,3]. To emulate this process, a DNN can be trained to learn  $M(x, y)$  and  $D(x, y)$ , which are then fed into the arctangent function for retrieving the phase.

Here, we present a BNN that uses the Concrete dropout [7] to approximate Bayesian inference in deep Gaussian processes for learning the numerator  $M(x, y)$  and the denominator  $D(x, y)$  statistically. We assume that  $\mathbf{X}$  is a set of input fringe images, which can be represented as  $\mathbf{X} = \{\mathbf{x}^k\}_{k=1}^K$ , where  $\mathbf{x}^k$  is the  $k$ th input fringe pattern and  $K$  is the size of the set.  $\mathbf{Y}$  is a set of ground-truth labels corresponding to the training data, which can be written as  $\mathbf{Y} = \{\mathbf{y}^k\}_{k=1}^K$ , where  $\mathbf{y}^k$  consists of the ground-truth numerator and denominator ( $M^k, D^k$ ).  $\mathbf{w}$  represents the weight matrix of the BNN. To investigate the distribution of the output of BNN, we model the predictive distribution  $p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{X}, \mathbf{Y})$  as

$$p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{X}, \mathbf{Y}) = \int p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{w}) p(\mathbf{w} | \mathbf{X}, \mathbf{Y}) d\mathbf{w}, \quad (3)$$

where  $p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{w})$  is the probability of the output  $\mathbf{y}^*$  given the input  $\mathbf{x}^*$ , the weights  $\mathbf{w}$ , and  $p(\mathbf{w} | \mathbf{X}, \mathbf{Y})$  the probability of the weights  $\mathbf{w}$  given the training data  $(\mathbf{X}, \mathbf{Y})$ . The distribution  $p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{w})$  describes the data uncertainty, and the distribution  $p(\mathbf{w} | \mathbf{X}, \mathbf{Y})$  characterizes the model uncertainty.

To measure the data uncertainty, we assume that  $\mathbf{y}^k$  has  $N$  pixels, and  $p(\mathbf{y}^k | \mathbf{x}^k, \mathbf{w})$  can then be written as

$$p(\mathbf{y}^k | \mathbf{x}^k, \mathbf{w}) = \prod_{i=1}^N p(y_i^k | \mathbf{x}^k, \mathbf{w}). \quad (4)$$

Assuming that the distribution of  $\mathbf{y}^k$  is Gaussian for each pixel, the data uncertainty can be captured by minimizing the negative log-likelihood function at the training stage,

$$-\frac{1}{K} \sum_k \log p(\mathbf{y}^k | \mathbf{x}^k, \mathbf{w}) = \frac{1}{K} \sum_{k=1}^K \left[ \frac{1}{2(\sigma^k)^2} \|\mathbf{y}^k - \hat{\mathbf{y}}^k\|^2 + \frac{1}{2} \log (\sigma^k)^2 \right], \quad (5)$$

where  $\mathbf{y}$  is the ground-truth label,  $\hat{\mathbf{y}}$  is the result predicted by BNN, and  $\sigma^2$  is the predicted variance.

To measure the model uncertainty, the Concrete dropout network is applied. By placing the Concrete dropout before every weight layer, we can use a simple variational distribution  $q(\mathbf{w})$  to approximate  $p(\mathbf{w} | \mathbf{X}, \mathbf{Y})$ , which is usually hard to calculate analytically. By using the Monte Carlo (MC) integration over  $T$  samples satisfying  $\mathbf{w}^{(t)} \sim q(\mathbf{w})$ , Eq. (3) can be approximated as

$$p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{X}, \mathbf{Y}) \approx \int p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{w}) q(\mathbf{w}) d\mathbf{w} \approx \frac{1}{T} \sum_{t=1}^T p(\mathbf{y}^* | \mathbf{x}^*, \mathbf{w}^{(t)}). \quad (6)$$

At the prediction stage, the dropout layers in our BNN randomly set input neurons to zero with a learned dropout rate.

By collecting the results of stochastic forward propagation through the trained model, the predictive mean can be computed and be used as the prediction of the BNN,

$$\hat{\mu} = E(\mathbf{y}^* | \mathbf{x}^*, \mathbf{X}, \mathbf{Y}) \approx \frac{1}{T} \sum_{t=1}^T E(\mathbf{y}^* | \mathbf{x}^*, \mathbf{w}^{(t)}) = \frac{1}{T} \sum_{t=1}^T \mathbf{y}^{*(t)}, \quad (7)$$

where  $E$  is the expectation. The model uncertainty is measured by the variance of the predicted results:

$$\hat{\sigma}^{\text{model}} = \sqrt{E[\text{Var}(\mathbf{y}^* | \mathbf{w}, \mathbf{x}^*, \mathbf{X}, \mathbf{Y})]} \approx \sqrt{\frac{1}{T} \sum_{t=1}^T (\mathbf{y}^{*(t)} - \hat{\mu})^2}. \quad (8)$$

Then, the data uncertainty is quantified by the average of the estimated variance:

$$\hat{\sigma}^{\text{data}} = \sqrt{\text{Var}(E[\mathbf{y}^* | \mathbf{w}, \mathbf{x}^*, \mathbf{X}, \mathbf{Y}])} \approx \sqrt{\frac{1}{T} \sum_{t=1}^T (\sigma^2)^{(t)}}. \quad (9)$$

Our BNN follows the architecture of the U-Net. In the training stage, the dropout rate of each layer is not fixed and can be learned automatically by BNN. More details about the theory, the structure, and the learned dropout rates of BNN are provided in [Supplement 1](#).

The diagram of the testing process of our method is shown in [Fig. 1](#). With an input fringe pattern, the trained BNN outputs  $T$  different sets of data including the numerator, the denominator, and their variance maps. The mean numerator and the mean denominator are obtained for calculating the final wrapped phase  $\hat{\mu}_\varphi$  by Eq. (2). To obtain the data/model uncertainty of the phase, we calculate the data/model uncertainty of the numerator and the denominator using Eqs. (9) and (8) first, and then apply the propagation of uncertainty:

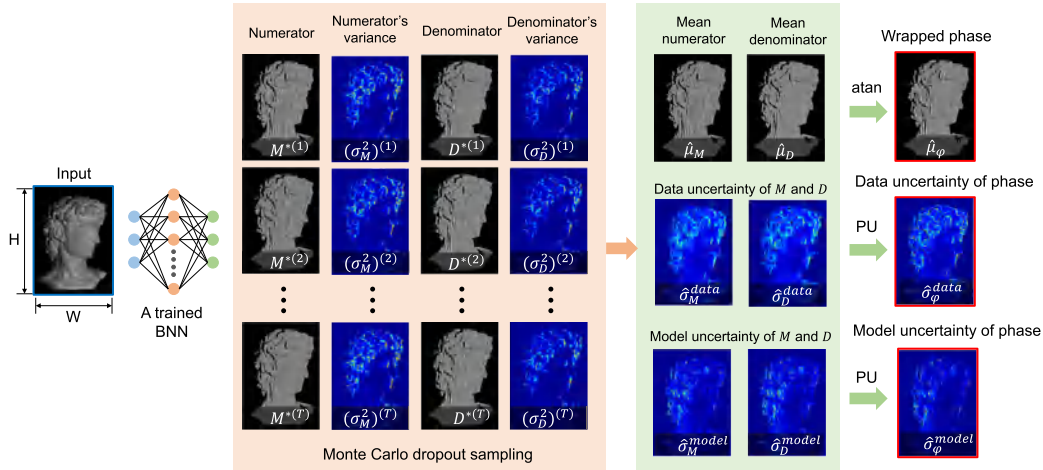
$$\hat{\sigma}_\varphi^{\text{model/data}} = \sqrt{\left( \frac{\partial \varphi}{\partial M} \hat{\sigma}_M^{\text{model/data}} \right)^2 + \left( \frac{\partial \varphi}{\partial D} \hat{\sigma}_D^{\text{model/data}} \right)^2}. \quad (10)$$

More details on the calculation of the phase and its uncertainties are provided in [Supplement 1](#).

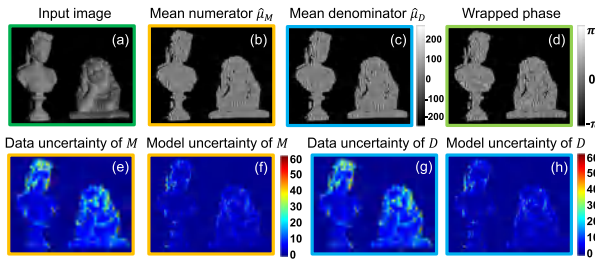
We tested the proposed method under the scenario of fringe projection profilometry. Our system consisted of a projector (DLP 4100, Texas Instruments) and a camera (V611, Vision Research Phantom). The projector illuminated test objects with pre-designed fringe patterns and the camera captured 8-bit grayscale images simultaneously from a different perspective. The spatial frequency of the projected fringes was  $f = 160$ . To collect training data, we captured many fringe images of different kinds of objects and generated the ground-truth labels by a 12-step PS algorithm. The BNN was implemented by using the Keras and computing on a graphic card (GTX Titan, NVIDIA). Further details about the optical setup, implementation of BNN, and tests with fringe patterns of different spatial frequencies are provided in [Supplement 1](#).

The test scene shown in [Fig. 2\(a\)](#) contains two plaster statues that are not present in the training stage. The trained BNN used the fringe image as an input and made  $T = 50$  predictions. The mean of the numerator and the denominator, and the wrapped phase, are shown in [Figs. 2\(b\)–2\(d\)](#), respectively. The corresponding uncertainties are demonstrated in [Figs. 2\(e\)–2\(h\)](#), respectively.





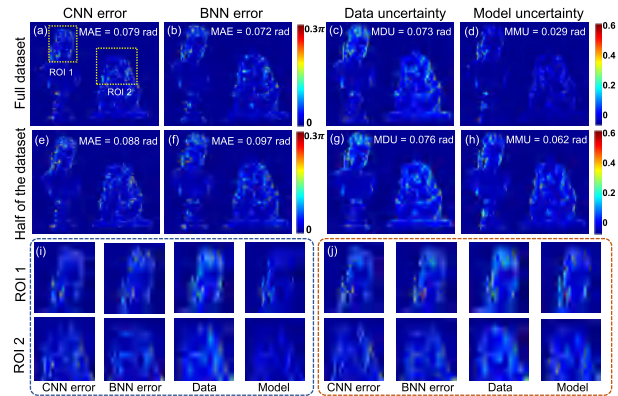
**Fig. 1.** Schematic of the proposed method. With the Monte Carlo dropout sampling,  $T$  samples of the BNN's prediction are obtained for an input fringe pattern. Each prediction outputs a set of data including  $M^{*(l)}$ ,  $(\sigma_M^2)^{(l)}$ ,  $D^{*(l)}$ , and  $(\sigma_D^2)^{(l)}$ . The wrapped phase  $\hat{\mu}_\phi$  is obtained by feeding the mean  $\hat{\mu}_M$  and the mean  $\hat{\mu}_D$  into the arctangent function. To obtain the phase uncertainties, we first calculate the uncertainties of the numerator and the denominator and then apply the propagation of uncertainty (PU).



**Fig. 2.** Test of the trained BNN. (a) The input fringe pattern. (b)–(d) Mean numerator, mean denominator, and wrapped phase, respectively. (e) and (f) Data uncertainty and the model uncertainty of the estimated numerator, respectively. (g) and (h) Corresponding uncertainties of the denominator.

Our BNN is well-calibrated, and the evaluation of the predicted uncertainties is provided in Supplement 1. To investigate the phase accuracy, we unwrapped the phase by using the temporal phase unwrapping approach [8] and calculated the phase error against a ground-truth phase map, which was obtained by the 12-PS algorithm. In Supplement 1, the unwrapped phase has been converted into the 3D reconstruction for better investigation of recovered surface details.

To demonstrate the efficacy of the uncertainties, we also trained the BNN with only half of the training data. For comparison, a convolutional U-Net (termed as “CNN”) that had no dropout layers was trained as well. Figures 3(a) and 3(b) show the absolute phase error when both models were trained with all of the data. The two networks demonstrated similar performance on the phase measurement as the BNN followed the main structure of the U-Net. Two regions of interest (ROIs) were selected, and their error distributions are shown in Figs. 3(i) and 3(j). For both the CNN and BNN, the phase errors are small for smooth areas, such as the statues' faces. But, the error begins to increase rapidly for the sharp regions, e.g., the hairs of the statues. From Figs. 3(c) and 3(d), we can see that the distribution of uncertainties faithfully indicate the error distribution, where the areas with large errors have been labeled with large uncertainties. We find the model uncertainty is



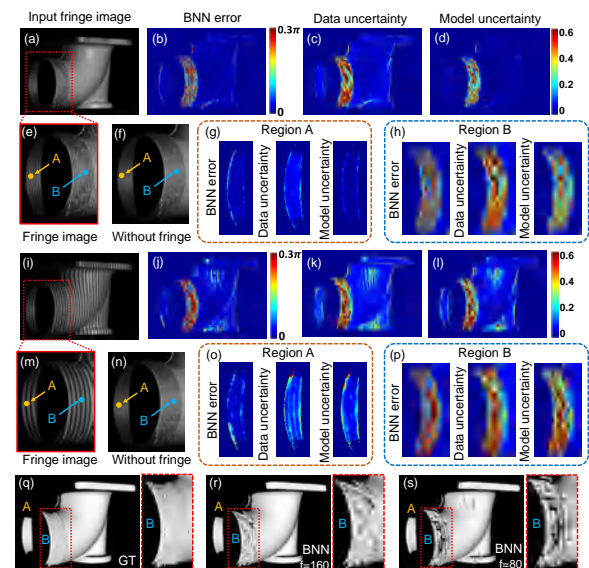
**Fig. 3.** Analysis of the phase error and uncertainties of the BNN in two different cases. For the first case where the full training dataset was used for training: (a) absolute phase error of CNN; (b) absolute phase error of BNN; (c) and (d) BNN's data uncertainty and model uncertainty of the phase. (e)–(h) Corresponding results for the second case where only half of the dataset was used for training. (i) and (j) show the errors and uncertainties of the two ROIs in the first case and the second case, respectively.

small, implying that the phase prediction can be performed consistently by the BNN. The data uncertainty is more significant, which is the result of the image noise in the captured images. In fringe projection, dense fringe patterns (e.g.,  $f = 160$ ) are usually captured with compromised fringe contrast. Next, the errors of both CNN and BNN increase when only half of the data were used, as can be seen in Figs. 3(e) and 3(f). We can see the data uncertainty almost does not change as the data reduction did not affect the data noise. However, the model uncertainty rises significantly. Its mean value surges from 0.029 rad to 0.062 rad, as can be seen from Figs. 3(d) and 3(h). The reduction of training data has an adverse effect on the robustness of the model, thus increasing its doubt about the prediction.

Further, we tested the BNN by using a tough sample that is a complex industrial part with screw thread shown in Fig. 4(a). The absolute phase error of the BNN is shown in Fig. 4(b). It can be

seen that the error of the smooth cylindrical area is small but that of the screw thread region is quite large. The data uncertainty and the model uncertainty are demonstrated in Figs. 4(c) and 4(d). We can see the BNN has faithfully indicated the overall error distribution. For detailed investigations, we have a magnified view of the screw thread region as shown in Fig. 4(e), where A represents the internal area and B represents the screw thread. A background image without fringes was also captured, and the selected area is shown in Fig. 4(f), which demonstrates that the internal area A is smooth without any screw structure. As the smooth surface is common and has been seen by the BNN during training, the uncertainty maps indicated high credibility, and the error is small, as can be seen in Fig. 4(g). For region B, however, the error shown in Fig. 4(h) is very serious. By comparing Figs. 4(e) and 4(f), we can see the projected fringe patterns happened to couple with the structure of the screw thread at region B, forming an approximate low-frequency moiré pattern. As a result, it is difficult for the neural network to handle this rare case, thus resulting in the significant model uncertainty. We also find that the moiré pattern has also been captured by the data uncertainty, which implies that it may also be treated as a kind of image noise by BNN. Moreover, an out-of-distribution (OOD) fringe image that has a different spatial frequency ( $f = 80$ ) was also tested. The corresponding results are shown in Figs. 4(i)–4(p), where the phase error and the predictive uncertainties are more severe for the whole scene. For region A, the mean data uncertainty and model uncertainty rise to 0.14 rad and 0.12 rad from 0.074 rad and 0.025 rad, respectively. For region B, they increase to 0.55 rad and 0.48 rad from 0.45 rad and 0.31 rad, respectively. We can see that the model is very suspicious of its prediction for the OOD data. Further, if considered in a quality control setting, this experiment would provide a typical example of how the BNN allows for making better decisions. When using deep learning methods for detecting surface defects, one may face the risk of incorrectly classifying an industrial part as a defective product due to a failure of the DNN. By converting the phase results into 3D reconstructions [Figs. 4(q)–4(s)], we can see that the 12-step PS method successfully measured the profile of the complex threaded region B, while the network produced inconsistent and distorted reconstructions. In this case, the “defect” is caused by the network rather than the object itself. It is worth noting that the estimated uncertainty maps have captured this problem by showing high uncertainties for this region. Consequently, instead of blindly believing that the product is defective, we should resort to alternative (preferably more reliable) methods to further check this dubious result. More experimental results of the BNN’s performance in handling never-experienced input data are provided in Supplement 1.

In this work, we have presented a fringe-pattern analysis framework using a BNN that can not only demodulate the phase information from a single fringe image but also output pixel-wise uncertainty maps describing the confidence of the neural network on its prediction. The BNN is developed by using the MC Concrete dropout approximation. This strategy is easy to implement and can be extended to other existing neural networks by simply adding extra Concrete dropout layers. To validate the proposed method, we tested the performance of the BNN in the conditions of varying training dataset size, rare test inputs, and OOD data, respectively. Experimental results have shown that the predicted uncertainty maps can successfully indicate the distribution of real phase errors without using any ground-truth data. In the future, error-reduction methods based on the estimated



**Fig. 4.** Uncertainty analysis of a measured complex industrial part with screw thread. (a) A captured fringe image ( $f = 160$ ). (b) Absolute phase error of BNN. (c) Data uncertainty of BNN. (d) Model uncertainty of BNN. (e) Magnified view of the fringe image for the selected area, where A indicates the internal area and B the area with screw thread. (f) Magnified view of a background image (without projected fringes) for the same selected area. (g) and (h) phase errors and uncertainty maps of region A and region B, respectively. (i)–(p) Corresponding results when the frequency of the projected fringes is  $f = 80$ . (q)–(s) 3D reconstructions obtained by the 12-PS method (ground-truth method, GT) and the BNNs.

uncertainty maps will be further investigated. We believe that a DNN that can provide confidence measure of the estimated phase is crucial to fringe-pattern analysis and that it has great potential for inspiring novel and reliable learning-based optical metrology approaches.

**Funding.** National Natural Science Foundation of China (62075096); Leading Technology of Jiangsu Basic Research Plan (BK20192003); Jiangsu Provincial “One belt and one road” innovation cooperation project (BZ2020007); Fundamental Research Funds for the Central Universities (30921011208).

**Disclosures.** The authors declare no conflicts of interest.

**Data availability.** Data underlying the results presented in this Letter may be obtained from the authors upon reasonable request.

**Supplemental document.** See Supplement 1 for supporting content.

## REFERENCES

1. M. Servin, J. A. Quiroga, and M. Padilla, *Fringe Pattern Analysis for Optical Metrology: Theory, Algorithms, and Applications* (Wiley, 2014).
2. M. Takeda and K. Mutoh, *Appl. Opt.* **22**, 3977 (1983).
3. C. Zuo, S. Feng, L. Huang, T. Tao, W. Yin, and Q. Chen, *Opt. Laser Eng.* **109**, 23 (2018).
4. S. Feng, Q. Chen, G. Gu, T. Tao, L. Zhang, Y. Hu, W. Yin, and C. Zuo, *Adv. Photon.* **1**, 025001 (2019).
5. A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?” arXiv:1703.04977 (2017).
6. Y. Xue, S. Cheng, Y. Li, and L. Tian, *Optica* **6**, 618 (2019).
7. Y. Gal, J. Hron, and A. Kendall, “Concrete dropout,” arXiv:1705.07832 (2017).
8. C. Zuo, L. Huang, M. Zhang, Q. Chen, and A. Asundi, *Opt. Laser Eng.* **85**, 84 (2016).

# Opto-Electronic Advances

ISSN 2096-4579

CN 51-1781/TN

## Physics-informed deep learning for fringe pattern analysis

Wei Yin, Yuxuan Che, Xinsheng Li, Mingyu Li, Yan Hu, Shijie Feng, Edmund Y. Lam, Qian Chen and Chao Zuo

**Citation:** Yin W, Che YX, Li XS, et al. Physics-informed deep learning for fringe pattern analysis. *Opto-Electron Adv* 7, 230034(2024).

<https://doi.org/10.29026/oea.2024.230034>

Received: 7 March 2023; Accepted: 12 May 2023; Published online: 31 August 2023

## Related articles

### Deep-learning-enabled dual-frequency composite fringe projection profilometry for single-shot absolute 3D shape measurement

Yixuan Li, Jiaming Qian, Shijie Feng, Qian Chen, Chao Zuo

*Opto-Electronic Advances* 2022 5, 210021 doi: [10.29026/oea.2022.210021](https://doi.org/10.29026/oea.2022.210021)

### Deep learning assisted variational Hilbert quantitative phase imaging

Zhuoshi Li, Jiasong Sun, Yao Fan, Yanbo Jin, Qian Shen, Maciej Trusiak, Maria Cywińska, Peng Gao, Qian Chen, Chao Zuo

*Opto-Electronic Science* 2023 2, 220023 doi: [10.29026/oes.2023.220023](https://doi.org/10.29026/oes.2023.220023)

### Deep learning enabled single-shot absolute phase recovery in high-speed composite fringe pattern profilometry of separated objects

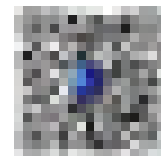
Maciej Trusiak, Malgorzata Kujawinska

*Opto-Electronic Advances* 2023 6, 230172 doi: [10.29026/oea.2023.230172](https://doi.org/10.29026/oea.2023.230172)

More related article in Opto-Electronic Journals Group website 



<http://www.ojournal.org/oea>



 OE\_Journal



 @OptoElectronAdv



DOI: [10.29026/oea.2024.230034](https://doi.org/10.29026/oea.2024.230034)

# Physics-informed deep learning for fringe pattern analysis

Wei Yin<sup>1,2,3†</sup>, Yuxuan Che<sup>1,2,3†</sup>, Xinsheng Li<sup>1,2,3</sup>, Mingyu Li<sup>1,2,3</sup>, Yan Hu<sup>1,2,3</sup>, Shijie Feng<sup>1,2,3\*</sup>, Edmund Y. Lam<sup>4\*</sup>, Qian Chen<sup>3\*</sup> and Chao Zuo<sup>1,2,3\*</sup>

Recently, deep learning has yielded transformative success across optics and photonics, especially in optical metrology. Deep neural networks (DNNs) with a fully convolutional architecture (e.g., U-Net and its derivatives) have been widely implemented in an end-to-end manner to accomplish various optical metrology tasks, such as fringe denoising, phase unwrapping, and fringe analysis. However, the task of training a DNN to accurately identify an image-to-image transform from massive input and output data pairs seems at best naïve, as the physical laws governing the image formation or other domain expertise pertaining to the measurement have not yet been fully exploited in current deep learning practice. To this end, we introduce a physics-informed deep learning method for fringe pattern analysis (PI-FPA) to overcome this limit by integrating a lightweight DNN with a learning-enhanced Fourier transform profilometry (LeFTP) module. By parameterizing conventional phase retrieval methods, the LeFTP module embeds the prior knowledge in the network structure and the loss function to directly provide reliable phase results for new types of samples, while circumventing the requirement of collecting a large amount of high-quality data in supervised learning methods. Guided by the initial phase from LeFTP, the phase recovery ability of the lightweight DNN is enhanced to further improve the phase accuracy at a low computational cost compared with existing end-to-end networks. Experimental results demonstrate that PI-FPA enables more accurate and computationally efficient single-shot phase retrieval, exhibiting its excellent generalization to various unseen objects during training. The proposed PI-FPA presents that challenging issues in optical metrology can be potentially overcome through the synergy of physics-priors-based traditional tools and data-driven learning approaches, opening new avenues to achieve fast and accurate single-shot 3D imaging.

**Keywords:** optical metrology; deep learning; physics-informed neural networks; fringe analysis; phase retrieval

Yin W, Che YX, Li XS et al. Physics-informed deep learning for fringe pattern analysis. *Opto-Electron Adv* 7, 230034 (2024).

## Introduction

Optical metrology, as a general-purpose metrology technique that uses light as information carriers for non-contact and non-destructive measurement<sup>1</sup>, is fundamental to manufacturing, basic research, and engineering ap-

plications. With the invention of the laser<sup>2</sup> and charge-coupled device (CCD)<sup>3</sup>, many optical metrology methods and instruments are employed in state-of-the-art manufacturing processes, precision positioning, and quality assessment because of their advantages in terms

<sup>1</sup>Smart Computational Imaging Laboratory (SCILab), School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; <sup>2</sup>Smart Computational Imaging Research Institute (SCIRI) of Nanjing University of Science and Technology, Nanjing 210019, China; <sup>3</sup>Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing 210094, China; <sup>4</sup>Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam, Hong Kong SAR 999077, China.

<sup>†</sup>These authors contributed equally to this work.

\*Correspondence: SJ Feng, E-mail: [shijiefeng@njust.edu.cn](mailto:shijiefeng@njust.edu.cn); EY Lam, E-mail: [elam@eee.hku.hk](mailto:elam@eee.hku.hk); Q Chen, E-mail: [chenqian@njust.edu.cn](mailto:chenqian@njust.edu.cn); C Zuo, E-mail: [zuochoao@njust.edu.cn](mailto:zuochoao@njust.edu.cn)

Received: 7 March 2023; Accepted: 12 May 2023; Published online: 31 August 2023



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License.

To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024. Published by Institute of Optics and Electronics, Chinese Academy of Sciences.

of accuracy, sensitivity, repeatability, and speed. In optical metrology, based on physical models of the image formation, the observed measurements (e.g., deformed fringe/speckle images) can be transformed into the desired physical properties of the objects (the profile, distance, strain, etc.). For many optical measurement techniques such as interferometry<sup>4</sup>, digital holography<sup>5</sup>, and fringe projection profilometry (FPP)<sup>6,7</sup>, the accuracy and efficiency of phase retrieval from the recorded fringe images are essential to reconstruct various underlying quantities dynamically. The most efficient method for phase measurement is recovering the phase distribution from a single fringe image, but as a typical case in optical metrology, it is an ill-posed inverse problem. The spatial phase-demodulation (SPD) methods can achieve single-frame fringe analysis by imposing some prior assumptions on the recovered phase (spatially smooth, limited spectral extension, piecewise constant, etc.)<sup>8-10</sup>, but at the cost of accuracy and resolution. Since optical metrology experiments are generally carried out in highly customized systems and stringent environments, phase-shifting (PS) methods can provide a deterministic and straightforward solution to the phase retrieval problem by additionally capturing multiple fringe patterns<sup>11</sup>. PS methods have obvious advantages in terms of speed, accuracy, and repeatability, which have brought up many high-end optical metrology instruments. However, when the optical system is under harsh measurement conditions or the state of the object changes dynamically, PS methods will be severely limited and cannot provide accurate phase recovery results for dynamic measurements. Despite extensive research efforts for decades, how to achieve phase measurement with the highest possible accuracy from the minimum number (preferably single shot) of fringe patterns remains one of the most challenging problems in optical metrology.

With the explosive growth of available data and computing resources, deep learning, as a “data-driven” machine learning technique, has achieved impressive success in numerous fields, such as computer vision and computational imaging<sup>12</sup>. Deep learning pervades almost all aspects of optical metrology<sup>13</sup>, and provides solutions to many challenging problems, such as fringe denoising<sup>14,15</sup>, fringe analysis<sup>16</sup>, and digital holographic reconstruction<sup>17-19</sup>. Feng et al.<sup>16</sup> proposed a deep learning method for fringe pattern analysis that establishes an inverse mapping between single-frame fringe and the label phase obtained using 12-step PS method. The trained

network can directly estimate the sine and cosine components of fringes, enabling single-shot phase reconstruction with higher accuracy than SPD methods. Recently, phase retrieval methods based on deep learning have been applied to ultrafast 3D imaging (speed up to 20 kHz)<sup>20</sup>, phase measuring deflectometry<sup>21</sup>, and single-frame absolute 3D measurement<sup>22</sup> by adopting diverse deep neural networks (DNNs) with a fully convolutional architecture<sup>23,24</sup> or combining the predictions of multiple networks with ensemble learning<sup>25</sup>. However, these deep learning approaches focus mainly on training a DNN to accurately identify an image-to-image transform from massive input and output data pairs of training datasets without considering the physical laws governing the image formation or other domain expertise pertaining to the measurement. Consequently, the performance of deep learning approaches in solving complex physical problems relies heavily on the underlying statistical characteristics within the dataset. To improve the performance of the network under real experimental conditions, it is necessary to pay a high price for collecting a large amount of high-quality data. In addition, due to the highly customized nature of optical metrology systems, networks trained on one system may not be directly transferable to another system of the same type. Once the new input is different even slightly from the training data, data-driven DNNs may exhibit a poor generalization under diverse measurement conditions, and cannot ensure the interpretability and traceability of their output results. On the contrary, based on accurate physical models of the image formation and its inverse solutions, traditional SPD methods can achieve reliable phase measurements for different types of samples<sup>26</sup>, but their measurement precision is limited. If the forward physical models of the image formation or traditional solvers of the inverse problem are incorporated into the DNN, it is expected to enhance the performance of deep learning methods while utilizing fewer network parameters. Goy et al.<sup>27</sup> proposed a physics-informed deep learning method for phase retrieval at low photon counts that leverages physical priors to convert the raw intensity measurement with noise into an initial estimate of the object, thereby significantly improving the phase reconstruction accuracy by using deep learning. Wang et al.<sup>28</sup> demonstrated an unsupervised single-beam phase imaging network to reconstruct the phase of the measured diffraction pattern by integrating a numerically propagated diffraction model. Saba et al.<sup>29</sup> proposed a physics-

informed neural network for tomographic reconstructions of biological samples, which minimizes the physical loss based on the Helmholtz equation, accurately and quickly retrieving the refractive index distribution from the scattered fields of the sample collected by different illumination directions.

For the limited ability of fringe analysis networks without physics priors, we present a physics-informed deep learning method for fringe pattern analysis (PI-FPA). A learning-enhanced Fourier transform profilometry (LeFTP) module with the prior knowledge of SPD methods is embedded in the DNN to directly provide accurate and reliable phase recovery results for new types of samples, while circumventing the requirement of collecting a large amount of high-quality data in supervised learning methods. The phase results are then refined using a lightweight DNN to further improve the accuracy and computational efficiency of single-shot phase retrieval. Experimental results show that the proposed PI-FPA exhibits superior single-shot fringe analysis performance in speed, accuracy, repeatability, and generalization to various unseen objects during training.

## Principle

### Phase retrieval from fringe images

Phase retrieval from fringe images is a fundamental task and a representative case among many applications of deep learning in optical metrology. The fringe image  $I(x, y)$  is expressed as<sup>30,31</sup>

$$I(x, y) = A(x, y) + B(x, y)\cos[\phi(x, y)], \quad (1)$$

where  $A(x, y)$  and  $B(x, y)$  are the background intensity and the fringe amplitude, and  $\phi(x, y)$  is the phase of the tested object. Retrieving the desired  $\phi(x, y)$  from only one fringe image  $I(x, y)$  is an ill-posed inverse problem due to two unknown parts  $A(x, y)$  and  $B(x, y)$ . In FPP, PS methods<sup>11</sup> transform the original ill-posed problem into a well-posed and solvable one by projecting a set of PS patterns to obtain additional observations of the target object:

$$I_n(x, y) = A(x, y) + B(x, y)\cos[\phi(x, y) - 2\pi n/N], \quad (2)$$

$$\phi(x, y) = \arctan \frac{\sum_{n=0}^{N-1} I_n(x, y)\sin(2\pi n/N)}{\sum_{n=0}^{N-1} I_n(x, y)\cos(2\pi n/N)}, \quad (3)$$

where  $I_n(x, y)$  represents  $N$ -step PS images,  $\phi(x, y)$  can be obtained by the least-squares algorithm. However, when the measured object is under harsh measurement conditions, the relative motion between the object and PS pat-

terns will introduce non-negligible errors into phase retrieval results<sup>32,33</sup>. Unlike PS methods, SPD methods can realize single-shot phase retrieval using different spatial transform techniques (such as the Fourier transform (FT)<sup>9</sup> and the windowed Fourier transform<sup>10</sup>) under the local smoothness assumption. In Fourier transform profilometry (FTP), the Fourier transform of  $I(x, y)$  in Eq. (1) gives

$$\mathcal{F}_I(f_x, f_y) = \mathcal{F}_A(f_x, f_y) + \mathcal{F}_C(f_x, f_y) + \mathcal{F}_{C^*}(f_x, f_y), \quad (4)$$

$$C(x, y) = \frac{1}{2}B(x, y)\exp\{i2\pi f_0 x\}\exp\{\phi_0(x, y)\}, \quad (5)$$

where  $\mathcal{F}_A$  and  $\mathcal{F}_C$  are the Fourier transform of  $A(x, y)$  and  $C(x, y)$ .  $\phi(x, y)$  is taken as the sum of two independent parts: the object component  $\phi_0(x, y)$  and the carrier frequency  $2\pi f_0 x$ . Based on the Fourier shift theorem, the zero order  $\mathcal{F}_A$  is separated with  $\pm 1$  orders  $\mathcal{F}_C$  and  $\mathcal{F}_{C^*}$ , so  $\mathcal{F}_C$  can be extracted by a band-pass filter and converted inversely to the retrieved phase,

$$\phi(x, y) = \arctan \frac{\text{Im}\{C(x, y)\}}{\text{Re}\{C(x, y)\}}. \quad (6)$$

However, when the measured surface contains sharp edges or discontinuities, the support of the zero order and  $\pm 1$  orders will be extended to cause the spectrum overlapping, precluding high-accuracy phase measurement of complex objects.

Unlike traditional methods that focus on understanding the image formation and solving inverse problems, Feng et al.<sup>16</sup> utilized DNNs to directly estimate the sine and cosine components of  $I(x, y)$  for single-shot fringe analysis:

$$\phi(x, y) = \arctan \frac{M(x, y)}{D(x, y)} = \arctan \frac{\rho B(x, y)\sin\phi(x, y)}{\rho B(x, y)\cos\phi(x, y)}, \quad (7)$$

where  $\rho$  is a constant that depends on phase retrieval methods, e.g.,  $\rho = 0.5$  for FT methods and  $\rho = N/2$  for  $N$ -step PS methods. However, the performance of phase retrieval networks relies heavily on a large amount of high-quality data. Once the new input is different from the training data, the reliability of phase reconstruction results output by data-driven DNNs cannot be guaranteed.

### Physics-informed deep learning method for fringe pattern analysis (PI-FPA)

As shown in Fig. 1, different from traditional physics-driven methods (FT methods) and data-driven deep learning approaches (e.g., U-Net and its derivatives) for



fringe pattern analysis, the proposed PI-FPA mainly contains a LeFTP module with physics priors and a lightweight network. The LeFTP module, which parameterizes the phase retrieval process of FT methods, utilizes the learnable filters operating in the Fourier transform domain to directly output initial phases in the manner of FTP in Fig. 1(a, b). Physics-driven LeFTP is highly generalizable to provide reliable phase results for various unseen objects during training. The lightweight network refines the initial phase to further improve the phase accuracy at a low computational cost, compared with universal end-to-end image transform networks (U-Net and its derivatives).

The schematic diagram of the proposed PI-FPA is shown in Fig. 2. First, a net head with a simple convolutional structure is adopted to extract rich low-level features of the input single-frame fringe, which can reduce the effect of the zero order  $\mathcal{F}_A$  of the fringe after training. For the LeFTP module in Fig. 2(c), similar to traditional FT methods, the input tensor is transformed into the Fourier domain through Fourier transform and spectrum centering. Instead of the simple filtering operation of FTP, two learnable filters with multiple channels are utilized to adaptively extract the +1 order  $\mathcal{F}_C$  closely related to the desired phase. Specifically, a learnable filter  $w_1$  with  $K_1 \times H_1 \times W_1$  size is applied to weaken the zero order located at the center C1 by weighting each feature of the input spectrum  $\mathcal{F}_{in}$  pixel by pixel:

$$\mathcal{F}_1^{K_1 \times H_1 \times W_1} = w_1^{K_1 \times H_1 \times W_1} \circ \mathcal{F}_{in}^{K_1 \times H_1 \times W_1}, \quad (8)$$

where  $\circ$  is the Hadamard product. Note that the un-

filtered high-frequency component is kept to avoid missing details, and the redundant negative Fourier spectrum is removed. Then, a series of filtering operations are implemented to extract delicately the +1 order in various ways using another learnable filter  $w_2$  with  $K_2 \times H_2 \times W_2$  size:

$$\mathcal{F}_2^{K_2 \times H_2 \times W_2} = w_2^{K_2 \times H_2 \times W_2} \cdot \sum_{k=0}^{K_1-1} \mathcal{F}_1^{k \times H \times W}, \quad (9)$$

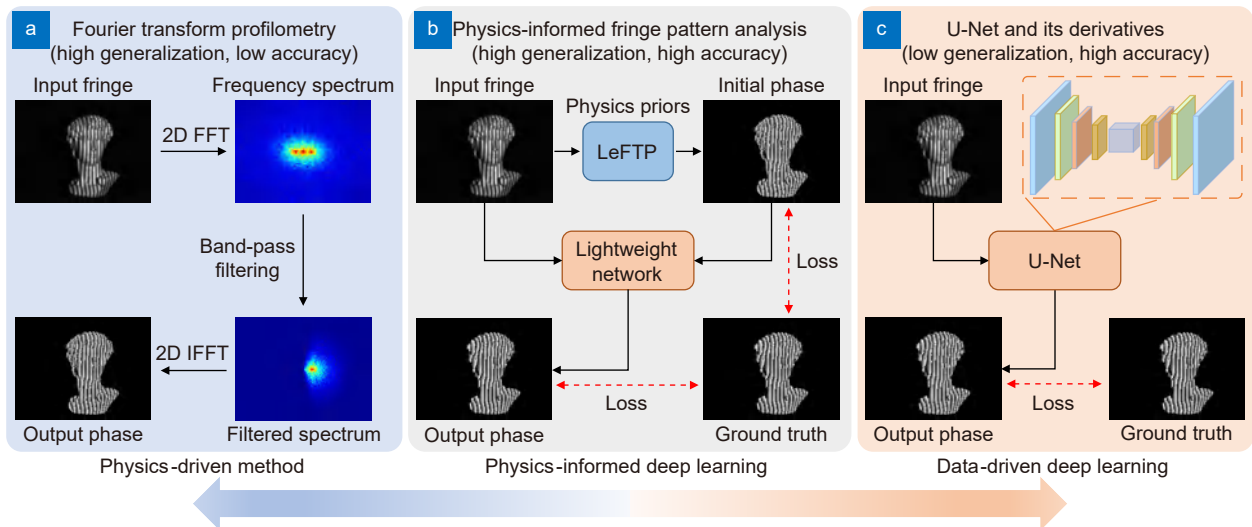
where the center of  $w_2$  is set as C2 estimated by  $N$ -step PS. Due to the asymmetry of the spectrum, a large number of reliable and initial phases can be recovered inversely from the filtered spectrum  $\mathcal{F}_2$  according to Eq. (6). Further, to optimize the phase retrieval performance of LeFTP, a priors-based initialization strategy for the filter weights is adopted to facilitate its efficient learning and avoid anchoring in local minima during the training phase by following background-normalized Fourier transform profilometry (BNFTP)<sup>34</sup>. The filter  $w_1$  is initialized as an inverse Hanning window for filtering the zero-order component of the input spectrum centered on C1:

$$w_1^{\text{init}}(k, f_x, f_y) = 1 - \cos \frac{2\pi f_x}{W_1}. \quad (10)$$

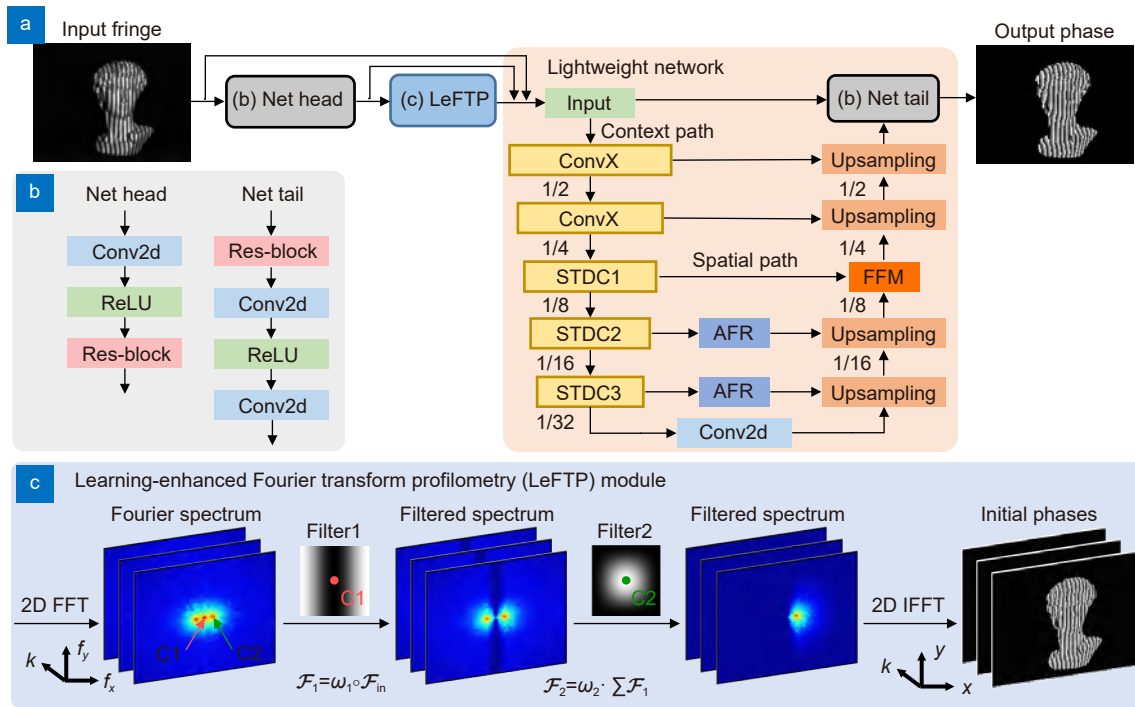
In addition, the +1 order of the spectrum centered on C2 is strengthened using another Hanning filter  $w_2$ :

$$w_2^{\text{init}}(k, f_x, f_y) = \cos \frac{2\pi f_x}{H_2} \cos \frac{2\pi f_y}{W_2}. \quad (11)$$

At present, mainstream fringe analysis approaches using deep learning exploit end-to-end fully convolutional networks in a naïve manner to build an image-to-image



**Fig. 1 | Diagrams of the physics-driven method, physics-informed deep learning approach, and data-driven deep learning approach for fringe pattern analysis.**



**Fig. 2 | Overview of the proposed PI-FPA. (a)** PI-FPA including a LeFTP module and a lightweight network. **(b)** Net head and Net tail. **(c)** The phase retrieval process of the LeFTP module.

inverse mapping between single-frame fringe and the label phase using massive network parameters. Thanks to robust phase estimation of LeFTP, it not only helps PI-FPA to circumvent the requirement of collecting a large amount of high-quality data in supervised learning methods, but also relieves the burden of phase refinement for lightweight DNNs. The lightweight network, consisting of the context path and the spatial path inspired by BiSeNet<sup>35,36</sup> in Fig. 2(a) (see Supplementary information for detailed analysis), is utilized to further improve the phase accuracy at a low computational cost compared with universal end-to-end image transform networks (U-Net and its derivatives). Instead of configuring more channels for higher-level layers as U-Net, the context path aims at collecting the fringe and initial phase features with a large receptive field through fast down-sampling and encoding global context information to guide the refined high-level features for learning, while the spatial path captures spatial information encoding rich detail information and outputs low-level features. In the encoder part of the context path, a fast down-sampling strategy with several ConvX blocks and the Short-Term Dense Concatenate (STDC) module is used to extract the feature information with scalable receptive field and multi-scale information. In the decoder phase, the attention-based feature refinement (AFR) module

and the fast upsampling operation based on bilinear interpolation are utilized to improve the feature resolution progressively. In the spatial path, its encoder part shares the same parameters with the context path, and captures the spatial information encoding rich detailed information and outputs low-level features. The features from the context path and the spatial path are concatenated by Feature Fusion module (FFM), and upsampled to output final phases using the predicted  $M(x, y)$  and  $D(x, y)$  in Eq. (7). The objective of PI-FPA is to minimize the joint loss of the phase and its Fourier domain:

$$Loss = Loss_{\text{phase}} + Loss_{\text{Fourier}}, \quad (12)$$

$$Loss_{\text{phase}} = \frac{\alpha_1 |Y - Y_{\text{GT}}|^2 + \alpha_2 |N \cdot Y_{\text{LeFTP}} - Y_{\text{GT}}|^2}{HW}, \quad (13)$$

$$Loss_{\text{Fourier}} = \frac{\beta_1 |\mathcal{F}_Y - \mathcal{F}_{Y_{\text{GT}}}| + \beta_2 |N \cdot \mathcal{F}_{Y_{\text{LeFTP}}} - \mathcal{F}_{Y_{\text{GT}}}|}{HW}, \quad (14)$$

where  $Y_{\text{GT}} = (M_{\text{GT}}, D_{\text{GT}})$  is the ground truth obtained using  $N$ -step PS,  $Y = (M, D)$  is the network's output,  $Y_{\text{LeFTP}} = (M_{\text{LeFTP}}, D_{\text{LeFTP}})$  is the LeFTP module's output, and  $\mathcal{F}_Y$ ,  $\mathcal{F}_{Y_{\text{GT}}}$ , and  $\mathcal{F}_{Y_{\text{LeFTP}}}$  are the 2D Discrete Fourier Transform of  $Y$ ,  $Y_{\text{GT}}$ , and  $Y_{\text{LeFTP}}$ .

## Experiments

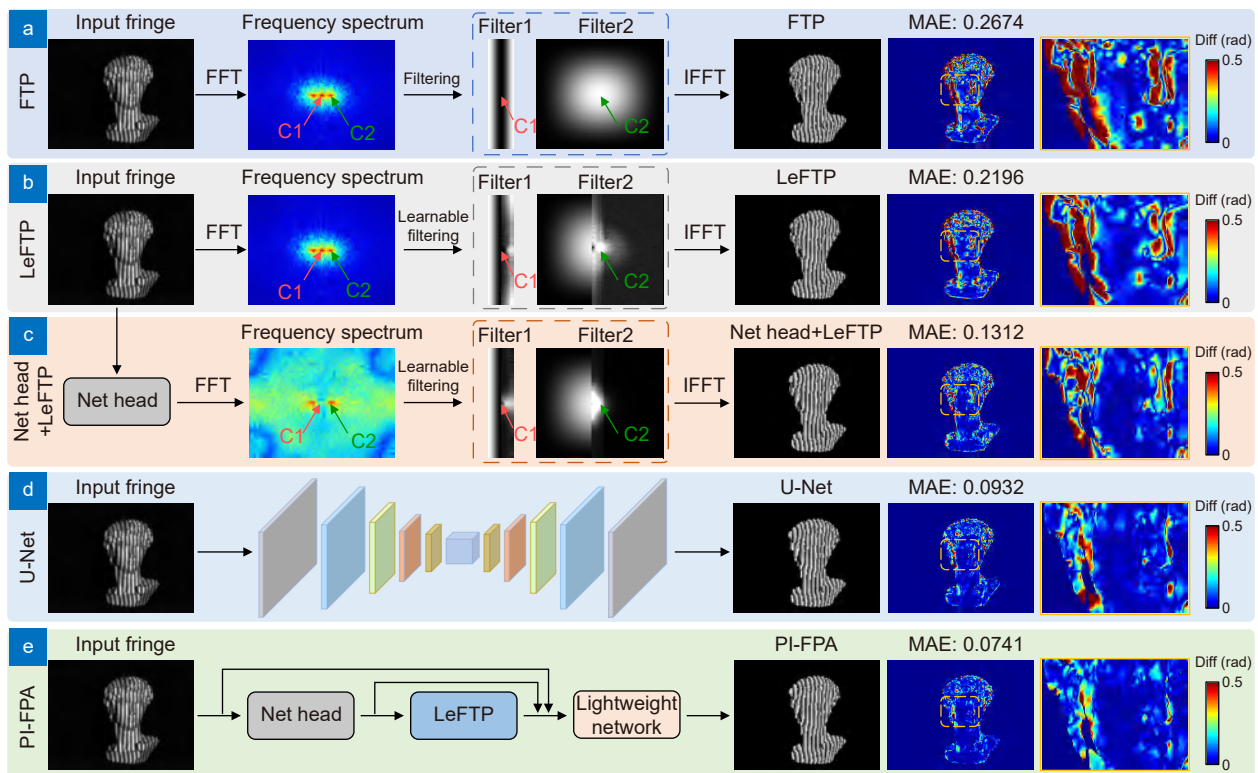
In order to verify the proposed PI-FPA under the

scenario of FPP, we built a multi-view structured light system that consisted of a projector (LightCrafter 4500Pro, Texas Instruments) and three cameras (acA640-750um, Basler) (see Supplementary information for detailed analysis). To collect fringe data for training, the projector projects three sets of PS patterns with different periods (including 1, 8, and 64) onto the test objects. The captured 64-period fringe image is the input of PI-FPA, and the label phase is obtained by 12-step PS. In the experiment, we collected the dataset including 1200 image pairs, which are divided into 800 image pairs for training, 200 image pairs for validation, and 200 image pairs for testing. The proposed PI-FPA is implemented using Pytorch framework (Facebook) and is computed on an NVIDIA GeForce RTX2080Ti graphics card. The composite loss function consists of mean square error (MSE) and mean absolute error (MAE) in Eq. (12). The optimizer is Adam, and the training epoch is set as 300.

First, a David plaster was measured to reveal single-shot phase retrieval process of PI-FPA, and FTP, LeFTP, Net head + LeFTP, and U-Net were implemented for comparison. In Fig. 3(a, b), LeFTP makes use of two learnable filters operating in the Fourier domain and re-

duces the MAE of phase errors by about 18% compared with FTP. By visualizing the filter weights, it demonstrates that LeFTP facilitates adaptive spectrum extraction through learning-enhanced filtering, which provides an interpretable guide for parameter optimization of FTP to improve the phase accuracy. In addition, due to the removal of redundant negative Fourier spectra in LeFTP, the left half of the filter weights is the same as its initial state, which is not updated during network training. To further speed up the LeFTP module, it is optional to cut down the size of two learnable filters in half to reduce the total parameters of the network and improve the inference speed of the network. Further, the Net head in Fig. 3(c), taken as the filtering operation in image pre-processing, is embedded in the front of LeFTP to extract rich low-level fringe features for removing the zero order, further reducing the phase errors by about 40%. It proves that LeFTP is plug-and-play to significantly boost the performance of single-frame fringe pattern analysis.

Different from these methods above, U-Net automatically exploits massive low-level and high-level features to optimize the phase accuracy as shown in Fig. 3(d), but at the cost of computational overhead. Specifically, U-Net needs 3.5 GB of GPU memory to process single-



**Fig. 3 | Comparative results for single-shot fringe pattern analysis of the David model. (a–e)** The phase retrieval process, wrapped phases, phase errors, and magnified views of the phase errors using FTP, LeFTP, Net head + LeFTP, U-Net, and PI-FPA.

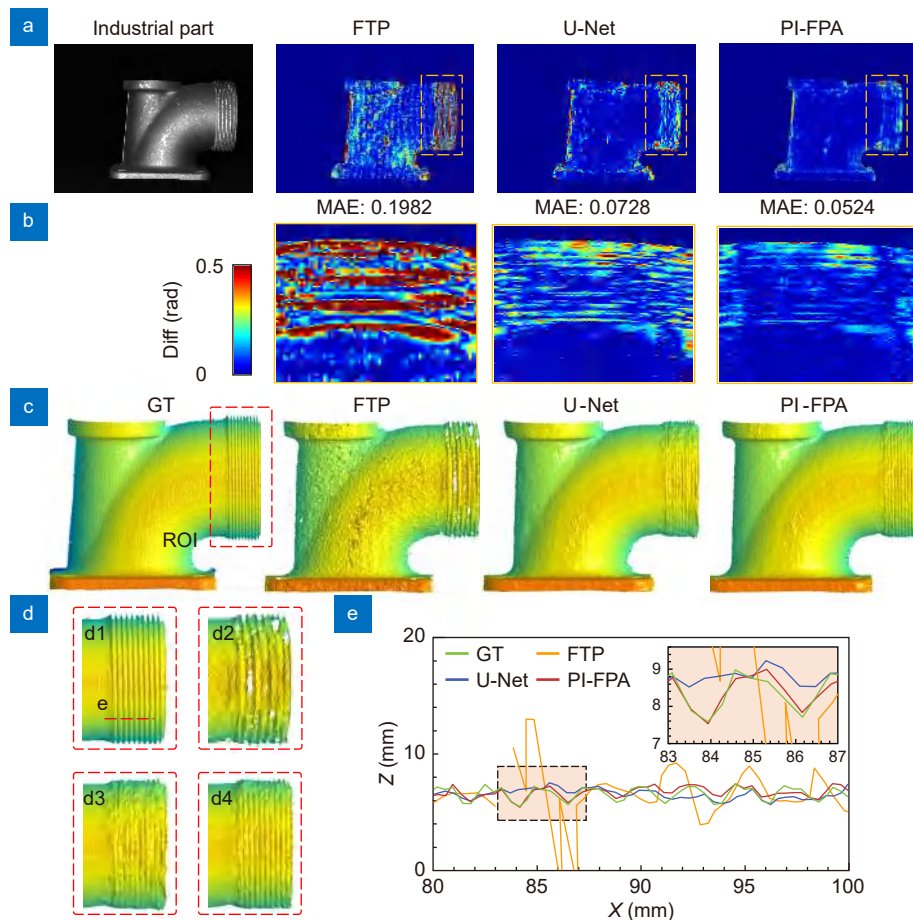


frame fringe and takes a runtime of 65.02 ms on Nvidia RTX 2080Ti. Guided by reliable phase results provided by Net head + LeFTP, in Fig. 3(e), PI-FPA refines the phases through a lightweight DNN, which reduces the GPU memory to 1.5 GB and improves the speed to 53.23 FPS while decreasing the MAE by about 20%. The magnified maps of phase errors in Fig. 3 indicate that the trained PI-FPA is able to reconstruct high-quality phase information for local fine details of objects with complex surfaces. In addition, the results of U-Net and PI-FPA using different amounts of training images are presented in Fig. S5 (see Supplementary information for detailed analysis). Compared with U-Net with 800 training image pairs, PI-FPA reduces the MAE of the phase errors by about 13% while requiring only 400 training image pairs, which demonstrates its good generalization.

To verify the generalization of PI-FPA for complex surfaces, we tested an industrial part, and fringe analysis results using different methods show that the phase er-

rors are smaller in smooth cylindrical regions but larger in sharp edges, while PI-FPA brings better phase quality among these methods as shown in Fig. 4(a, b). Further, we adopted stereo phase unwrapping<sup>37</sup> to achieve single-shot 3D imaging in Fig. 4(c) (see Supplementary information for detailed analysis). As the magnified regions in Fig. 4(d), the screw thread of the workpiece, which is relatively rare in the training dataset, causes significant degradation in the performance of U-Net, precluding high-precision reconstruction of complex surfaces. The line profiles in Fig. 4(e) prove that the proposed PI-FPA with physics-driven LeFTP can successfully recover the fine profiles of the threads and provide accurate and physically consistent 3D imaging results to approach the ground truth (GT), even though the network has not seen such experimental data during the training phase.

To quantitatively analyze the 3D imaging accuracy of PI-FPA, our system was applied to measuring a dynamic scene at the camera speed of 100 Hz: a ceramic plane and



**Fig. 4 | Comparative fringe analysis results of the industrial part.** (a) The industrial part and the phase errors using FTP, U-Net, and PI-FPA. (b) The magnified views of the phase errors. (c) Single-shot 3D imaging results using different methods. (d) The magnified views of (c). (e) The line profiles in (d).

a standard sphere moving along the  $Z$  axis, and 3D reconstruction results at different time points are shown in Fig. 5(a). The error distributions of the moving sphere are obtained by sphere fitting at  $T = 0$  s, 0.81 s, and 1.62 s, where major measured errors are less than 100  $\mu\text{m}$  with the RMS of 52.198  $\mu\text{m}$ , 42.112  $\mu\text{m}$ , and 53.295  $\mu\text{m}$  as shown in Fig. 5(b). Similarly, Fig. 5(c) shows the measured RMS of the moving plane is 51.425  $\mu\text{m}$ , 38.922  $\mu\text{m}$ , and 37.183  $\mu\text{m}$ . In Fig. 5(d, e), we further perform temporal precision analysis by collecting long-term data over a 1.62 s period using 3-step PS, FTP, U-Net, and PI-FPA. In Table 1, quantitative analysis results of the moving plane and sphere for different methods show that the measured results obtained by PI-FPA exhibited higher 3D reconstruction accuracy with a lower temporal standard deviation (STD) of  $43 \pm 4.1$   $\mu\text{m}$  and  $47 \pm 5.1$   $\mu\text{m}$ .

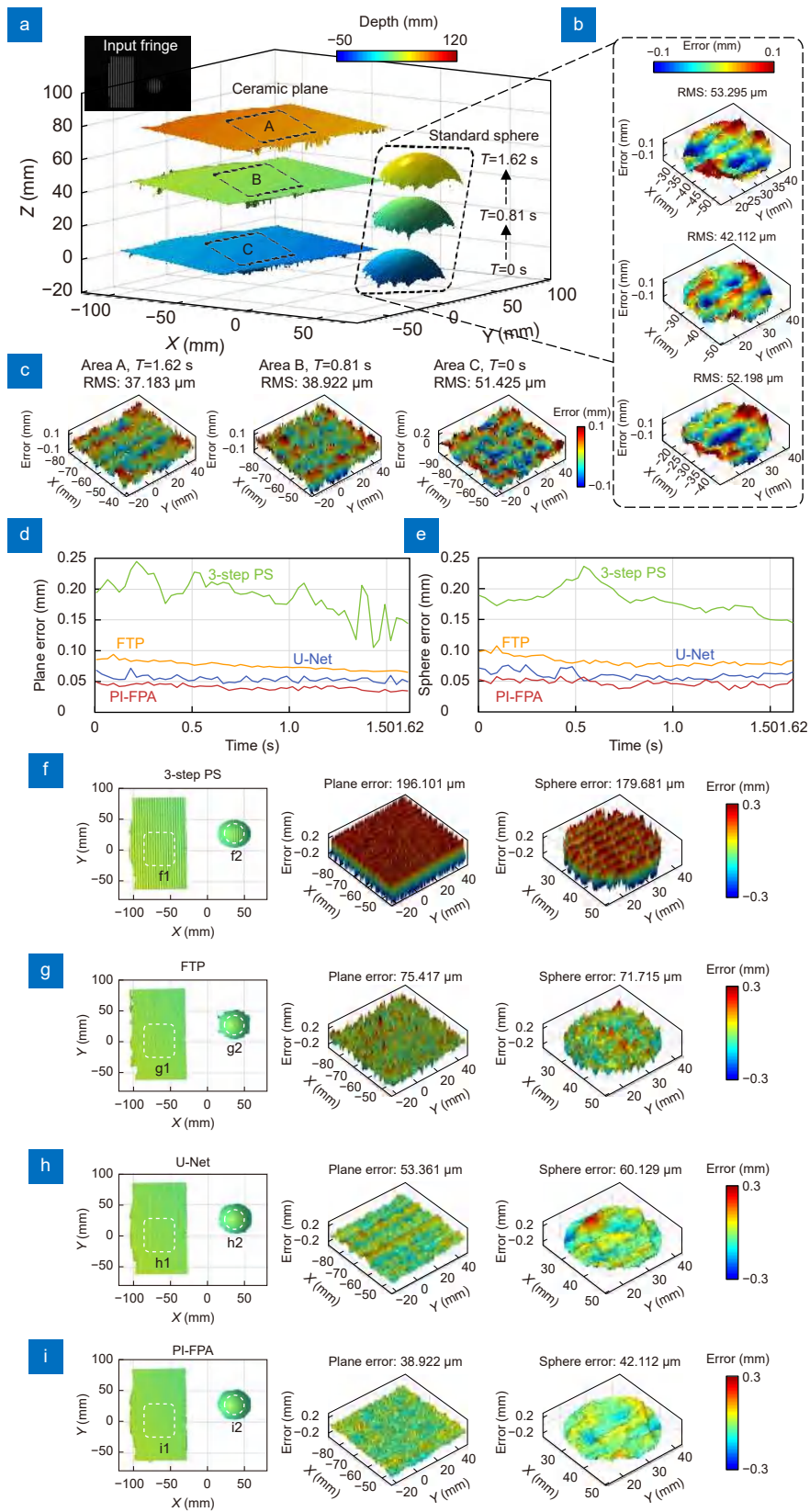
In Fig. 5(f–i), we additionally provide the measurement results of the moving plane and sphere at  $T = 0.81$  s using different methods. Different from FTP for single-shot phase retrieval, PS methods can realize pixel-by-pixel phase measurements with higher accuracy for complex shapes, but it needs to project at least three fringe patterns to obtain a phase map theoretically. As the most common and efficient case in  $N$ -step PS methods, 3-step PS is implemented for comparison. When dynamic scenes are measured, the relative motion between the object and the phase-shifting fringe patterns sequentially projected will cause motion artifacts and thus introduce non-negligible phase errors into the phase map. As a consequence, there are severe measurement errors with the RMS of 196.101  $\mu\text{m}$  and 179.681  $\mu\text{m}$  in the measurement results of 3-step PS in Fig. 5(f). In addition, for real-time 3D measurement based on 3-step PS, the whole procedure of 3D reconstruction is composed of phase retrieval, stereo phase unwrapping, and phase-to-height mapping, which is implemented with a graphics processing unit (GPU)<sup>38</sup> and several look-up tables<sup>39</sup> to speed up the 3D reconstruction. The 3D imaging speed is determined by the maximum between the image acquisition time and the runtime of 3D reconstruction. The

runtime of stereo phase unwrapping<sup>37</sup> and phase-to-height mapping for processing the images with the resolution of  $640 \times 480$  pixels is less than 5 ms on RTX2080Ti. Since 3-step PS needs to capture three fringe images and its runtime of the phase retrieval is negligible ( $5.22 \times 10^{-3}$  ms) in Table 1, its 3D imaging speed is limited to 33.33 FPS. On the contrary, the single-frame fringe analysis capability of FTP can significantly improve the accuracy and repeatability of fast 3D measurement to reduce the RMS to 75.417  $\mu\text{m}$  and 71.715  $\mu\text{m}$ , while its runtime ( $2.06 \times 10^{-2}$  ms) promotes the speed of 3D measurement to 100 FPS in Fig. 5(g). This result proves that single-frame fringe analysis methods are more suitable for dynamic scene measurement when the target's movement speed is in the same order of magnitude as the 3D imaging speed. Then, in Fig. 5(h), the RMS of the measurement error can be further decreased to 53.361  $\mu\text{m}$  and 60.129  $\mu\text{m}$  thanks to the powerful feature extraction capability of U-Net, but at the cost of lower inference speed (65.02 ms), precluding real-time 3D measurement. Finally, benefiting from the proposed LeFTP module and the lightweight DNN, PI-FPA takes a runtime of 18.78 ms to achieve fast single-shot phase reconstruction with higher accuracy in Fig. 5(i). However, PI-FPA only retrieves the phase of the first in the three-step PS images and reduces the 3D imaging speed to 33.33 FPS. 3D measurement results in Fig. 5 confirm that PI-FPA, whether measuring the moving plane or sphere, achieves successfully single-shot 3D shape measurement with higher accuracy and good repeatability for multiple moving objects simultaneously. The whole 3D measurement results can refer to Supplementary Video S1.

Last, to further demonstrate the advantages of PI-FPA, we applied our single-shot 3D imaging system to 360-degree reconstruction of a workpiece model and non-rigid dynamic face measurement as shown in Fig. 6 and Supplementary Video S2–S3. Fig. 6(a, b) show the captured fringe images of the rotated workpiece and non-rigid dynamic face at different time points and the corresponding color-coded 3D reconstruction results using different

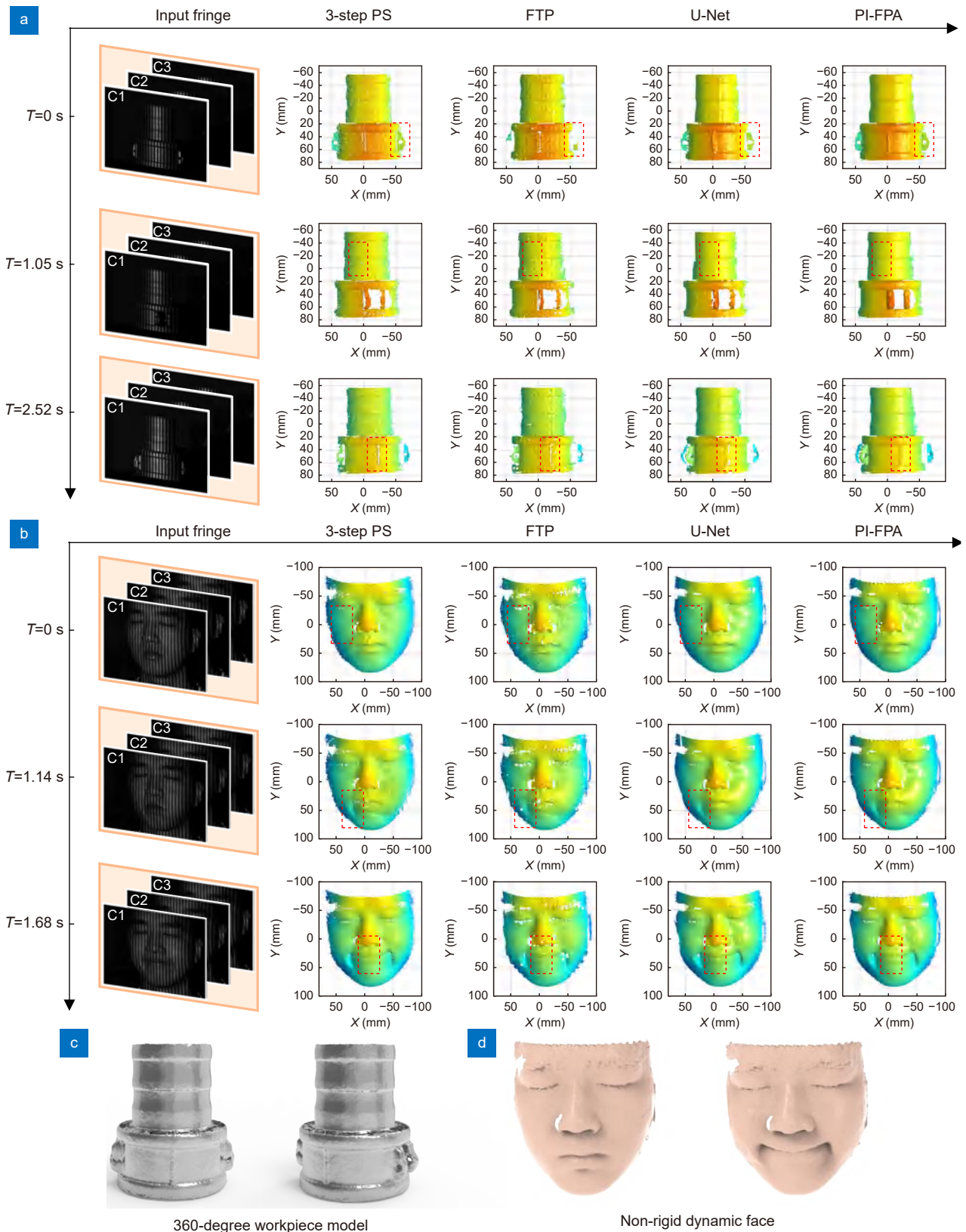
**Table 1 | Quantitative analysis results of the moving plane and sphere for different methods.**

Method	Time (ms)	RMS ( $\mu\text{m}$ )	
		Plane	Sphere
3-step PS	$5.22 \times 10^{-3}$	$188 \pm 29.8$	$179 \pm 19.9$
FTP	$2.06 \times 10^{-2}$	$77 \pm 6.8$	$81 \pm 7.4$
U-Net	65.02	$56 \pm 4.9$	$59 \pm 6.6$
PI-FPA	18.78	<b><math>43 \pm 4.1</math></b>	<b><math>47 \pm 5.1</math></b>



**Fig. 5 | Precision analysis for a ceramic plane and a standard sphere moving along the Z axis. (a)** 3D reconstruction results using PI-FPA at different time points. **(b–c)** the error distributions of the sphere and plane. **(d–e)** temporal precision analysis results of the plane and sphere over a 1.62 s period using 3-step PS, FTP, U-Net, and PI-FPA. **(f–i)** the color-coded 3D reconstruction and the corresponding error distributions of the plane and the standard sphere using different methods at  $T = 0.81$  s.





**Fig. 6 | Fast 3D measurement results using different fringe pattern analysis methods.** (a) The representative fringe images at different time points and the corresponding color-coded 3D reconstructions results for the rotated workpiece model using 3-step PS, FTP, U-Net, and PI-FPA. (b) The representative fringe images at different time points and the corresponding color-coded 3D reconstructions results for non-rigid dynamic face using 3-step PS, FTP, U-Net, and PI-FPA. (c) 360-degree 3D reconstruction of the workpiece model using PI-FPA. (d) 3D measurement results of non-rigid dynamic face using PI-FPA.

methods. For the rotated workpiece, the highlighted regions in Fig. 6(a) show that 3-step PS cannot recover the fine shapes of smooth surfaces due to the phase errors introduced by motion artifacts. For single-frame fringe analysis, FTP is suitable for dynamic 3D measurement, but yields coarse 3D results with low quality in terms of accuracy and resolution due to the spectrum overlapping. U-Net can further improve the quality of 3D reconstruction, but it cannot reliably retrieve the phase of the object with metal materials which is relatively rare in the training dataset, precluding the recovery of fine surfaces. This experiment demonstrates that the proposed PI-FPA can be applied for high-quality and efficient 3D modeling of complex structure parts as shown in Fig. 6(c). Similarly, for non-rigid dynamic face, there are inevitably a large amount of ripple-like measurement errors in 3D results of 3-step PS due to motion artifacts. And then, FTP is performed to significantly reduce measurement errors, but is unable to recover high-quality local details of the face. Due to the smooth and diffuse properties of faces, both PI-FPA and U-Net provide acceptable 3D face measurement results. Because of the lack of 3D label data for the tested face, it cannot identify precisely which of these two results is better, but there are slight differences in some local details, such as the left cheek and the tip of the nose in Fig. 6(b). In the whole measuring procedure, the reconstructed dynamic face at different time points verified the reliability of PI-FPA to perform fast 3D shape measurement with high completeness as well as see Supplementary Video S3. These results suggest that PI-FPA is a promising tool for fast 3D measurement and reverse modeling with high quality for objects with complex shapes.

## Conclusions and discussion

In summary, we have demonstrated a physics-informed deep learning method for fringe pattern analysis (PI-FPA) that is able to achieve accurate and computationally efficient single-shot phase reconstruction and exhibits strong generalization capability to new types of samples. By introducing the LeFTP module with the prior knowledge of traditional phase demodulation methods, PI-FPA circumvents the requirement of collecting a large amount of high-quality data, while overcoming the degradation of reconstruction quality for rare samples or structures in supervised learning methods. Utilizing reliable phase results from LeFTP as the network input, PI-FPA strengthens the ability of the lightweight DNN to

further improve the phase recovery accuracy at a low computational cost compared with existing end-to-end networks. The effectiveness of PI-FPA has been verified by several experiments for measuring various types of static and dynamic scenes. The single-shot phase retrieval results of the David plaster confirmed that PI-FPA can reconstruct high-quality phase information for objects with complex surfaces, while also achieving an improvement of 3.46× in its network inference speed compared with U-Net. By adopting stereo phase unwrapping, PI-FPA has the capability of single-frame 3D imaging to successfully recover the fine profiles of the industrial part with the threads, exhibiting its good generalization to rare samples never seen by the network. Temporal precision analysis results verified the high accuracy and excellent repeatability of PI-FPA for measuring multiple moving objects simultaneously. Finally, 360-degree reconstruction of a workpiece model and non-rigid dynamic face measurement revealed the applicability of PI-FPA for fast 3D measurement with high quality for objects with complex shapes and different materials. In the future, the performance of PI-FPA for phase retrieval from various types of fringe images will be investigated. We wish that PI-FPA can be applicable to other fringe analysis applications in optical metrology, further pushing the limits of fringe pattern analysis in speed, accuracy, repeatability, and generalization.

## References

1. Gäsvisk KJ. *Optical Metrology* 3rd ed (John Wiley & Sons, West Sussex, 2002).
2. Wyant JC, Creath K. Recent advances in interferometric optical testing. *Laser Focus* **21**, 118–132 (1985).
3. Kulkarni R, Rastogi P. Optical measurement techniques-A push for digitization. *Opt Lasers Eng* **87**, 1–17 (2016).
4. Hariharan P. *Basics of Interferometry* 2nd ed (Elsevier, Amsterdam, 2010).
5. Schnars U, Falldorf C, Watson J et al. *Digital Holography and Wavefront Sensing* 2nd ed (Springer, Berlin Heidelberg, 2015).
6. Gorthi SS, Rastogi P. Fringe projection techniques: whither we are? *Opt Lasers Eng* **48**, 133–140 (2010).
7. Geng J. Structured-light 3D surface imaging: a tutorial. *Adv Opt Photon* **3**, 128–160 (2011).
8. Servin M, Quiroga JA, Padilla JM. *Fringe Pattern Analysis for Optical Metrology: Theory, Algorithms, and Applications* (John Wiley & Sons, Weinheim, 2014).
9. Su XY, Chen WJ. Fourier transform profilometry: a review. *Opt Lasers Eng* **35**, 263–284 (2001).
10. Kemao Q. Windowed fourier transform for fringe pattern analysis. *Appl Opt* **43**, 2695–2702 (2004).
11. Zuo C, Feng SJ, Hiang L et al. Phase shifting algorithms for fringe projection profilometry: a review. *Opt Lasers Eng* **109**,

- 23–59 (2018).
12. Barbastathis G, Ozcan A, Situ G. On the use of deep learning for computational imaging. *Optica* **6**, 921–943 (2019).
  13. Zuo C, Qian JM, Feng SJ et al. Deep learning in optical metrology: a review. *Light Sci Appl* **11**, 39 (2022).
  14. Yan KT, Yi YJ, Huang CT et al. Fringe pattern denoising based on deep learning. *Opt Commun* **437**, 148–152 (2019).
  15. Kulkarni R, Rastogi P. Fringe denoising algorithms: a review. *Opt Lasers Eng* **135**, 106190 (2020).
  16. Feng SJ, Chen Q, Gu GH et al. Fringe pattern analysis using deep learning. *Adv Photon* **1**, 025001 (2019).
  17. Ren ZB, Xu ZM, Lam EYM. End-to-end deep learning framework for digital holographic reconstruction. *Adv Photon* **1**, 016004 (2019).
  18. Rivenson Y, Zhang YB, Günaydin H et al. Phase recovery and holographic image reconstruction using deep learning in neural networks. *Light Sci Appl* **7**, 17141 (2018).
  19. Liu KX, Wu JC, He ZH et al. 4K-DMDNet: diffraction model-driven network for 4K computer-generated holography. *Opto-Electron Adv* **6**, 220135 (2023).
  20. Feng SJ, Zuo C, Yin W et al. Micro deep learning profilometry for high-speed 3D surface imaging. *Opt Lasers Eng* **121**, 416–427 (2019).
  21. Qiao G, Huang YY, Song YP et al. A single-shot phase retrieval method for phase measuring deflectometry based on deep learning. *Opt Commun* **476**, 126303 (2020).
  22. Li YX, Qian JM, Feng SJ et al. Deep-learning-enabled dual-frequency composite fringe projection profilometry for single-shot absolute 3D shape measurement. *Opto-Electron Adv* **5**, 210021 (2022).
  23. Yang T, Zhang ZZ, Li HH et al. Single-shot phase extraction for fringe projection profilometry using deep convolutional generative adversarial network. *Meas Sci Technol* **32**, 015007 (2020).
  24. Yin W, Zhong JX, Feng SJ et al. Composite deep learning framework for absolute 3D shape measurement based on single fringe phase retrieval and speckle correlation. *J Phys Photon* **2**, 045009 (2020).
  25. Feng SJ, Xiao YL, Yin W et al. Fringe-pattern analysis with ensemble deep learning. *Adv Photon Nexus* **2**, 036010 (2023).
  26. Osten W. What optical metrology can do for experimental mechanics. *Appl Mech Mater* **70**, 1–20 (2011).
  27. Goy A, Arthur K, Li S et al. Low photon count phase retrieval using deep learning. *Phys Rev Lett* **121**, 243902 (2018).
  28. Wang F, Bian YM, Wang HC et al. Phase imaging with an untrained neural network. *Light Sci Appl* **9**, 77 (2020).
  29. Saba A, Gigli C, Ayoub AB et al. Physics-informed neural networks for diffraction tomography. *Adv Photon* **4**, 066001 (2022).
  30. Reid GT. Automatic fringe pattern analysis: a review. *Opt Lasers Eng* **7**, 37–68 (1986–1987).
  31. Rajshekhar G, Rastogi P. Fringe analysis: premise and perspectives. *Opt Lasers Eng* **50**, iii–x (2012).
  32. Weise T, Leibe B, Van Gool L. Fast 3D scanning with automatic motion compensation. In *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition* 1–8 (IEEE, 2007); <http://doi.org/10.1109/CVPR.2007.383291>.
  33. Feng SJ, Zuo C, Tao TY et al. Robust dynamic 3-D measurements with motion-compensated phase-shifting profilometry. *Opt Lasers Eng* **103**, 127–138 (2018).
  34. Zuo C, Tao TY, Feng SJ et al. Micro Fourier Transform Profilometry ( $\mu$ FTP): 3D shape measurement at 10, 000 frames per second. *Opt Lasers Eng* **102**, 70–91 (2018).
  35. Yu CQ, Wang JB, Peng C et al. BiSeNet: bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the 15th European Conference on Computer Vision* 325–341 (Springer, 2018); [http://doi.org/10.1007/978-3-030-01261-8\\_20](http://doi.org/10.1007/978-3-030-01261-8_20).
  36. Fan MY, Lai SQ, Huang JS et al. Rethinking BiSeNet for real-time semantic segmentation. In *Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9716–9725 (IEEE, 2021); <http://doi.org/10.1109/CVPR46437.2021.00959>.
  37. Tao TY, Chen Q, Feng SJ et al. High-precision real-time 3D shape measurement based on a quad-camera system. *J Opt* **20**, 014009 (2018).
  38. Feng SJ, Chen Q, Zuo C. Graphics processing unit-assisted real-time three-dimensional measurement using speckle-embedded fringe. *Appl Opt* **54**, 6865–6873 (2015).
  39. Liu K, Wang YC, Lau DL et al. Dual-frequency pattern scheme for high-speed 3-D shape measurement. *Opt Express* **18**, 5229–5244 (2010).

## Acknowledgements

This project was funded by National Key Research and Development Program of China (2022YFB2804603, 2022YFB2804604), National Natural Science Foundation of China (62075096, 62205147, U21B2033), China Postdoctoral Science Foundation (2023T160318, 2022M711630, 2022M721619), Jiangsu Funding Program for Excellent Postdoctoral Talent (2022ZB254), The Leading Technology of Jiangsu Basic Research Plan (BK20192003), The “333 Engineering” Research Project of Jiangsu Province (BRA2016407), The Jiangsu Provincial “One belt and one road” innovation cooperation project (BZ2020007), Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense (JSGP202105), Fundamental Research Funds for the Central Universities (30922010405, 30921011208, 30920032101, 30919011222), and National Major Scientific Instrument Development Project (62227818).

## Competing interests

The authors declare no competing financial interests.

## Supplementary information

Supplementary information for this paper is available at <https://doi.org/10.29026/oea.2024.230034>



Scan for Article PDF





# PHOTONICS Research

## Generalized framework for non-sinusoidal fringe analysis using deep learning

SHIJIE FENG,<sup>1,2,3</sup> CHAO ZUO,<sup>1,2,4</sup> LIANG ZHANG,<sup>1,2</sup> WEI YIN,<sup>1,2</sup> AND QIAN CHEN<sup>2,5</sup>

<sup>1</sup>Smart Computational Imaging (SCI) Laboratory, Nanjing University of Science and Technology, Nanjing 210094, China

<sup>2</sup>Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing University of Science and Technology, Nanjing 210094, China

<sup>3</sup>e-mail: shijiefeng@njjust.edu.cn

<sup>4</sup>e-mail: zuochao@njjust.edu.cn

<sup>5</sup>e-mail: chenqian@njjust.edu.cn

Received 2 February 2021; revised 26 March 2021; accepted 13 April 2021; posted 13 April 2021 (Doc. ID 420944); published 27 May 2021

Phase retrieval from fringe images is essential to many optical metrology applications. In the field of fringe projection profilometry, the phase is often obtained with systematic errors if the fringe pattern is not a perfect sinusoid. Several factors can account for non-sinusoidal fringe patterns, such as the non-linear input–output response (e.g., the gamma effect) of digital projectors, the residual harmonics in binary defocusing projection, and the image saturation due to intense reflection. Traditionally, these problems are handled separately with different well-designed methods, which can be seen as “one-to-one” strategies. Inspired by recent successful artificial intelligence-based optical imaging applications, we propose a “one-to-many” deep learning technique that can analyze non-sinusoidal fringe images resulting from different non-sinusoidal factors and even the coupling of these factors. We show for the first time, to the best of our knowledge, a trained deep neural network can effectively suppress the phase errors due to various kinds of non-sinusoidal patterns. Our work paves the way to robust and powerful learning-based fringe analysis approaches. © 2021 Chinese Laser Press

<https://doi.org/10.1364/PRJ.420944>

### 1. INTRODUCTION

Three-dimensional (3D) measurement plays an essential role in many fields, e.g., industrial manufacturing [1], medical treatment [2], entertainment [3], and identity recognition [4]. In convention, coordinate measuring machines provide users with accurate 3D data by way of point-by-point measurements [5]. However, its measuring speed is limited due to the point-wise and contact inspection. By contrast, optical 3D measurement techniques can obtain full-field geometric measurements within a single or several shots [6,7]. Among current optical 3D measurement techniques, structured light illumination profilometry has received extensive attention and is becoming one of the most promising 3D shape measurement techniques [8,9].

In structured light illumination profilometry, one illuminates test objects with patterns of various structures, such as sinusoidal fringes [10], de Bruijn patterns [11], speckle patterns [12], and aperiodic fringes [13]. For high-accuracy 3D measurements, sinusoidal fringe patterns are often preferred. Many fringe analysis methods have been proposed for extracting the object's phase from sinusoidal fringes. They can be broadly classified into two categories: spatial-demodulation methods [14–18] and temporal-demodulation methods [19–24]. For spatial-demodulation approaches, one can compute the phase

by using a single fringe image, demonstrating the advantage of high efficiency. Nevertheless, they tend to compromise for complex surfaces since high-frequency details are difficult to retrieve with only a single image. For temporal-demodulation methods, pixel-wise measurements with higher resolution and accuracy can be achieved. By representative phase-shifting (PS) algorithms [10], one captures several sinusoidal fringe images with a given phase shift and calculates the phase using a least-square method. As multiple images can provide more information about the same measured point, the phase of complex structures can be recovered with high accuracy. However, the main limitation of temporal-demodulation approaches is the reduced efficiency as several images have to be recorded. It is noteworthy that we need to ensure that the sinusoidal fringe patterns are captured with high quality for either spatial-demodulation or temporal-demodulation techniques.

Several inherent factors in structured light illumination can account for the collection of non-sinusoidal patterns. The first one is the gamma distortion of digital projectors. For visual quality, digital projectors or displays are often manufactured with specific gamma distortion, leading to a non-linear relationship between the output intensity and the input intensity that is  $I_{\text{out}} = I_{\text{in}}^\gamma$ . Researchers have proposed many approaches that can be roughly classified into system-based methods [8,25–27]

and algorithm-based methods [28–36] to relieve the gamma distortion. The system-based approaches suggest replacing commercial projectors with illumination units free from gamma effect, e.g., coherent light illumination setups [25] and programmable digital light processing (DLP) modules [8]. Although effective, they may increase the cost or the complexity of the whole system. To eliminate the gamma distortion without changing the system hardware, one can record the input and the output light intensity and predict the gamma value using the non-linear model [28–30]. Then, to counteract the gamma effect, one can pre-distort the input intensity using  $(I_{in})^{\frac{1}{\gamma}}$ , which can recover the true output intensity  $I_{out} = [(I_{in})^{\frac{1}{\gamma}}]^{\gamma} = I_{in}$ . Also, gamma-induced phase errors can be compensated by lookup tables that depict the relationship between the phase difference and the actual phase [31,32]. In addition, the weights of harmonic errors duo to the gamma effect can be predicted through some iteration algorithms, which can then be used for error compensation [33].

The second cause for captured non-sinusoidal fringes is the residual high-order harmonics in binary defocusing projection. In high-speed fringe projection, binary defocusing techniques have the advantages of fast image projection [37]. For projectors using digital micromirror devices, 8-bit fringe images are usually projected at the speed limit of 120 Hz as a relatively long integration time is required. For 1-bit binary fringes, however, the integration time of projection can be reduced to the minimum, allowing the projector to operate at kilohertz to tens of kilohertz. By defocusing the projector, we can have the binary stripe patterns transformed into gray-scale sinusoidal patterns. In practice, users should carefully adjust the defocusing degree of the projector. When the projector is defocused excessively, the fringe images are captured with a low contrast. On the opposite, if the defocusing degree is not enough systematic errors would occur, since harmonics in the binary fringes have not been filtered completely by the defocusing process. In practice, people prefer to defocus the projector slightly and then try to remove the systematic errors with well-designed algorithms, such as pulse width modulation [38,39], sinusoidal pulse width modulation (SPWM) [40], tripolar SPWM [41], optimal pulse width modulation [42], and dithering methods [43,44]. The main idea of these methods is to shift harmonics in the binary fringe from low-frequency areas to high-frequency sections of its spectrum, facilitating the low-pass filtering effect induced by the defocusing projection.

The third cause of non-sinusoidal fringes is the image saturation in high dynamic range (HDR) 3D shape measurements. For fringe projection profilometry, it is challenging to measure objects with a considerable variation in surface reflectivity, e.g., a scenario contains both dark and bright objects. The fringe patterns reflected from the dark regions are often captured with a low signal-to-noise ratio, whereas the pixels are usually saturated for the reflective surfaces. When dark objects are captured with proper fringe patterns, bright areas in the same scene are often measured with saturated (pure white) fringes. As object details have been covered up with the saturated fringes, it is hard to retrieve the phase. Various approaches to HDR fringe projection techniques have been proposed [45]. In general, these techniques can be classified into two groups:

equipment-based techniques [46–56] and algorithm-based techniques [57–62]. In the group of equipment-based methods, researchers try to acquire ideal fringe images by adjusting the imaging system, such as the exposure time [47], the intensity of projected light [50], the polarization states of illumination [46], and the number of camera views [46]. As to the algorithm-based methods, researchers concentrate on the design of phase retrieval algorithms instead of changing the imaging system's hardware, allowing the phase to be measured directly from saturated fringe images.

Further, the case will be more complicated if some of the non-sinusoidal factors are coupled together, which is seldom discussed in the current literature. For example, fringe images are captured with both the gamma effect and the image saturation, or with both the insufficient defocusing projection and the image saturation. This paper shows that the causes of these kinds of individual/coupling non-sinusoidal problems are similar, which can boil down to a superposition of the original sine wave of the fundamental frequency with several unknown sine waves at high frequencies (high-order harmonics). In practice, stochastic factors, e.g., the random noise, may also affect the captured fringe pattern but they are not discussed here as they will not change the main profile of a sinusoid.

Deep learning is a powerful machine learning technique that uses artificial neural networks with deep layers to fit complex mathematical functions. Compared with traditional algorithms that rely on physical models completely, deep learning approaches handle problems by searching and establishing sophisticated mapping between the input and the target data owing to the powerful computation capability. In many applications, learning-based methods have shown superiority to classic physical-model-based methods. In the field of image denoising, denoising autoencoders have been trained to obtain high level features for robust reconstruction of clean images [63,64]. In the field of nanophotonics, artificial intelligence has been applied to knowledge discovery, which shows great potential in understanding of the physics of electromagnetic nanostructures [65]. In the field of optical imaging, recent years have witnessed great successes of deep learning assisted applications. First, the deep neural network can significantly improve optical microscopy and increase its spatial resolution over a large field of view and depth of field [66]. Then, the deep learning techniques can be used for phase recovery and holographic image reconstruction in digital holography [67]. With only one hologram image, the twin-image and self-interference-related artifacts can be removed. Also, deep-learning-based ghost imaging techniques have shown much better performance than conventional ghost imaging in terms of different noise and measurement ratio conditions [68]. Furthermore, researchers have utilized deep learning strategies to build powerful models that can fit all scattering media within the same class, which improves the scalability of imaging through scattering [69]. Lastly, in optical coherence tomography (OCT), deep neural networks can be used to identify clinical features similar to how clinicians interpret an OCT image, allowing successful automated segmentations of clinically relevant image features [70].

In recent years, researchers have demonstrated that deep neural networks can be used to improve the performance of

fringe projection profilometry effectively. In fringe analysis, deep convolutional neural networks can be trained to retrieve the phase information from a single fringe image with favorable accuracy [71–74]. In phase unwrapping, learning-based temporal phase unwrapping [75] and stereo phase unwrapping methods [76] were developed to suppress noise effects and unwrap dense fringe patterns robustly. To handle complex surfaces, our previous work has shown that the deep learning technique can recover the phase from saturated fringe images [57]. Here, we show that more non-sinusoidal issues can benefit from deep learning. We demonstrate for the first time, to our knowledge, a generalized neural network can cope with various kinds of non-sinusoidal fringes that are caused by either single or multiple non-sinusoidal factors. Experimental results show that compared with traditional three-step phase-shifting algorithms, the proposed method can substantially improve the reconstruction accuracy by more than 50% without reducing the measurement efficiency.

## 2. PRINCIPLE

### A. Phase-Shifting Algorithm

In fringe projection profilometry, a projector illuminates test objects with pre-designed fringe images and a camera captures the images simultaneously from a different angle. The fringe patterns are distorted due to the varying height of measured areas. The phase retrieved from captured patterns serves as temporary textures of test objects and can be converted into the object's height. The  $N$ -step PS algorithm is widely applied to phase retrieval as it has the advantages of high accuracy, insensitivity to ambient light, and pixel-wise phase measurement. The captured  $N$ -step PS fringe image can be expressed as

$$I_n(x, y) = A(x, y) + B(x, y) \cos[\phi(x, y) - \delta_n], \quad (1)$$

where  $\phi(x, y)$  is the phase,  $A(x, y)$  is the background intensity,  $B(x, y)$  is the modulation, and  $\delta_n$  is the phase shift that is equal to  $2\pi n/N$ , where  $n = 0, 1, 2, \dots, N-1$ . When there are at least three images ( $N \geq 3$ ), the phase can be solved by

$$\phi(x, y) = \arctan \frac{\sum_{n=0}^{N-1} I_n(x, y) \sin(\frac{2\pi n}{N})}{\sum_{n=0}^{N-1} I_n(x, y) \cos(\frac{2\pi n}{N})}. \quad (2)$$

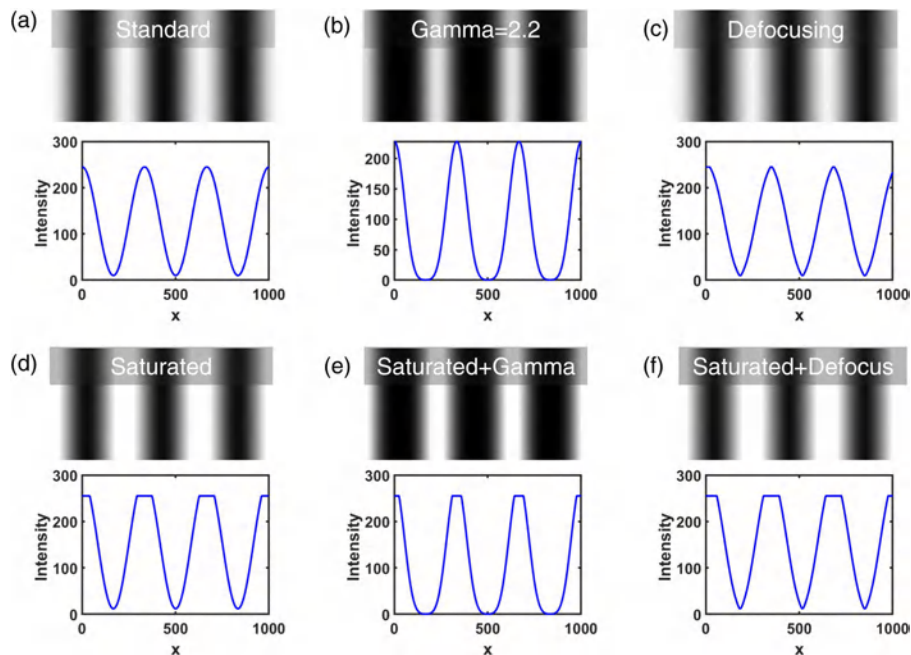
### B. Phase-Shifting Algorithm with Non-Sinusoidal Fringe Images

Non-sinusoidal PS images are often captured as a result of the projectors' gamma effect, the binary defocusing illumination, or the image saturation. A generalized model can be used to represent the captured images as

$$I_n^H(x, y) = A^H(x, y) + \sum_{j=1}^p B_j^H(x, y) \cos\{j[\phi(x, y) - \delta_n]\}, \quad (3)$$

where  $A^H(x, y)$  is the background intensity,  $B_j^H(x, y)$  is the modulation of the  $j$ th harmonic, and  $p$  is the number of harmonics. When  $p = 1$ ,  $I_n^H(x, y)$  is equal to  $I_n(x, y)$ , which is the case of perfect sinusoidal patterns. When  $p > 1$ , however, the fringe images become non-sinusoidal. This model has been proposed to characterize the fringe pattern affected by the gamma distortion [30]. We find that it is also applicable for more cases including the binary defocusing projection, fringe image saturation, and the coupling of these non-sinusoidal factors.

We illustrate different kinds of non-sinusoidal fringe images, as shown in Fig. 1. They are fringe patterns simulated under the gamma distortion, binary square wave with slight defocusing, image saturation, and the coupling of these non-sinusoidal factors, respectively. In the case of  $\gamma = 2.2$ , the sinusoidal wave's peaks



**Fig. 1.** Simulated sinusoidal fringe images and their cross sections. (a) An ideal sinusoidal pattern. (b) A gamma-distorted sinusoidal pattern. (c) A defocused binary stripe image. (d) A saturated sinusoidal fringe image. (e) A fringe image affected by both the gamma distortion and the image saturation. (f) A fringe image affected by both the defocusing and the image saturation.



become narrow while the valleys wide, giving rise to narrowed bright stripes compared with the ideal sinusoidal wave. For the defocused pattern as shown in Fig. 1(c), the intensity distribution looks like a triangular wave due to the presence of residual harmonics. For the case of image saturation, the intensity that exceeds the maximum dynamic range (i.e., 255 in this simulation) is truncated, while the rest keeps unchanged. Last, in the coupling cases as shown in Figs. 1(e) and 1(f), the image saturation further modifies the shape of the original non-sinusoidal waves by cutting off the intensity that exceeds the dynamic range, which further increases the non-sinusoidal characteristic of the wave.

Fourier analysis is then implemented to investigate the harmonics of these patterns. The results are shown in Figs. 2 and 3. The fundamental frequency  $f_0$  is three in our simulation. For the ideal sine wave, only the fundamental frequency  $f_0$  exists. For the case of  $\gamma = 2.2$ , the frequency components  $2f_0$  and  $3f_0$  begin to appear. For the case of the defocused binary pattern, as shown in Fig. 2(f), we can observe harmonics of  $3f_0$  and  $5f_0$  that survive the defocusing. Although their amplitudes are small, they can still destroy the phase retrieval. In Figs. 3(a) and 3(d), we can see that there are four additional frequency components that are from  $2f_0$  to  $5f_0$  in the saturated sine wave except for the fundamental frequency. Last, two coupling cases are discussed, which are the gamma effect coupled with the image saturation and the defocused pattern coupled with the image saturation, respectively. In Figs. 3(e) and 3(f), we can find that more harmonics have been introduced into the coupled gamma distorted pattern and the coupled defocused pattern due to the influence of image saturation, which has further destroyed the shape of the sinusoidal wave.

Next, we analyze the phase errors owing to the non-sinusoidal issues. Assume that the intensity difference for each PS image is

$$\Delta I_n(x, y) = I_n(x, y) - I_n^H(x, y). \quad (4)$$

The phase error caused by additional harmonics can be written as

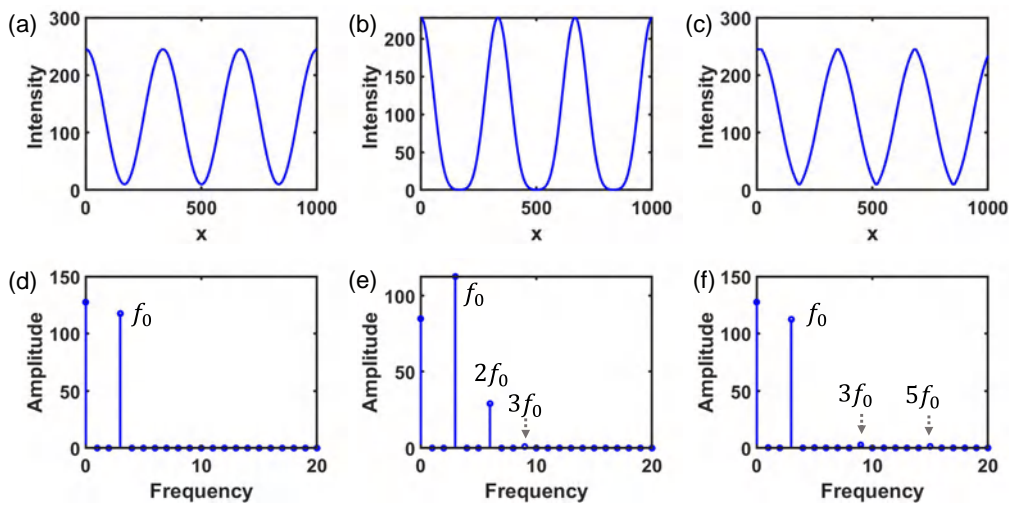
$$\Delta\phi(x, y) = \sum_{n=0}^{N-1} \frac{\partial\phi(x, y)}{\partial I_n(x, y)} \Delta I_n(x, y). \quad (5)$$

By substituting Eqs. (2) and (4) into Eq. (5), we have

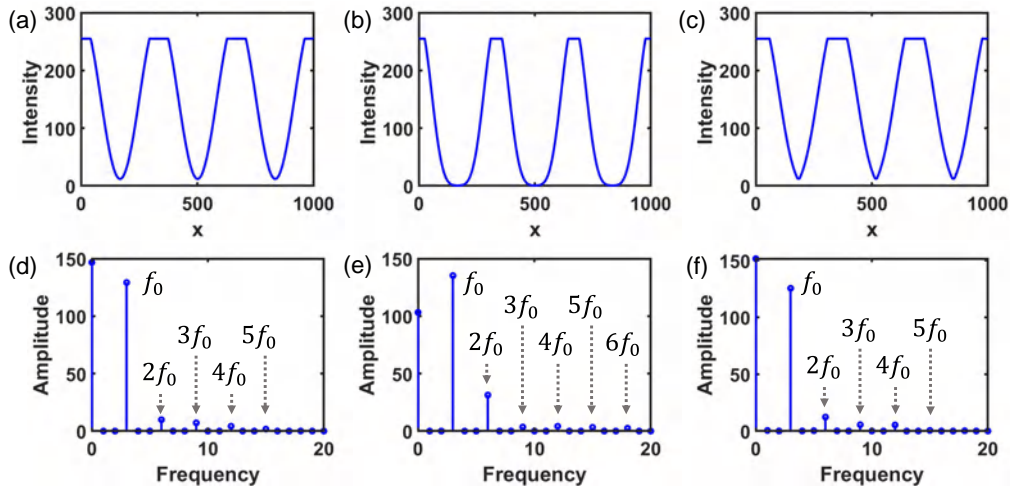
$$\Delta\phi = -\frac{4}{B^2 N^2} \sum_{n=0}^{N-1} \left\{ \left[ \sum_{m=0}^{N-1} I_m \sin \frac{2\pi(n-m)}{N} \right] \times \left[ \sum_{j=0}^p B_j \cos j \left( \phi - \frac{2\pi n}{N} \right) \right] \right\}. \quad (6)$$

Equation (6) shows the non-sinusoidal phase error of  $N$ -step PS algorithms. It can be found that the phase error  $\Delta\phi$  can be reduced by increasing the modulation of fundamental frequency and the number of phase shift  $N$ . As it is not easy to manipulate the former in practice, we study the influence of changing the number of phase shift.

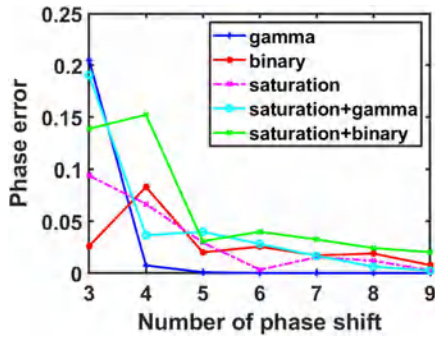
Figure 4 illustrates the performance of PS algorithms in analyzing different non-sinusoidal fringes. Here, the ground-truth phase is calculated with ideal sinusoidal fringes. The phase error is obtained by computing the standard deviation of the phase difference. In the simulation of gamma distortion, we set  $\gamma = 2.2$ . As can be seen, the phase error induced by the gamma effect decreases rapidly with the increase of the number of phase shift. For the case of defocused binary square pattern, the phase error reduces but with small fluctuations. The reason is that for an  $N$ -step PS algorithm, it is sensitive to  $(s+1)N \pm 1$ th harmonics (where  $s$  is an integer) [41]. For example, the four-step PS algorithm is sensitive to all of the odd harmonics present in the defocused pattern, showing the largest phase error among all of the PS algorithms. However, from the whole trend, the phase error still decreases with a large  $N$ . For the case of saturation, we truncated 20% of the maximum light intensity. Like the defocusing technique, its phase error also shows a trend that the error decreases with an increasing  $N$ . For the coupling case of image saturation and  $\gamma = 2.2$ , the phase error is larger than that of the case of pure gamma. For the coupling of defocusing projection and the image saturation, a more serious error is also observed than the one of the pure defocusing



**Fig. 2.** Intensity and spectrum of different sinusoidal patterns. (a)–(c) The intensity profile of the ideal sinusoidal fringe, the gamma-distorted fringe, and the defocused fringe, respectively. (d)–(f) The corresponding spectra of (a)–(c).



**Fig. 3.** Intensity and spectrum of different sinusoidal patterns. (a)–(c) The intensity profile of the saturated sinusoidal fringe, the saturated gamma-distorted fringe, and the saturated defocused fringe, respectively. (d)–(f) The corresponding spectra of (a)–(c).



**Fig. 4.** Performance of  $N$ -step phase-shifting algorithms for various kinds of non-sinusoidal fringes.

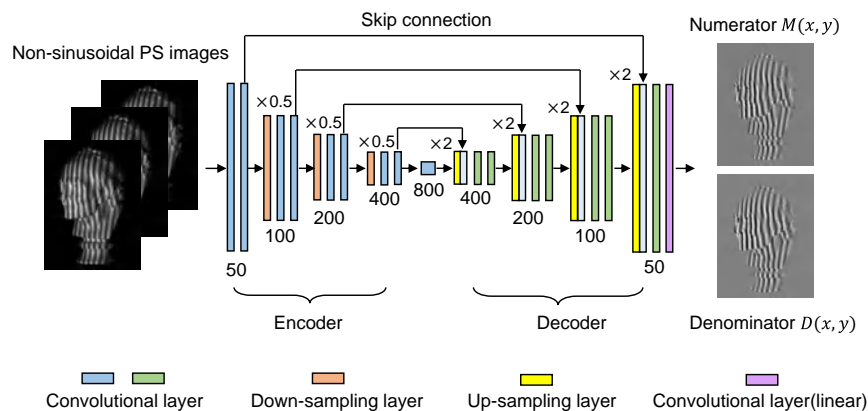
case. From the results, although different non-sinusoidal factors are superimposed in the coupling cases, the phase can still be robustly computed with a large step PS algorithm. In practice, however, the phase-shifting algorithm with a large number of steps requires many fringe images to be captured

for a single phase measurement, which limits the efficiency significantly.

### 3. ARCHITECTURE OF THE DEEP NEURAL NETWORK

From the previous section, the phase error due to the non-sinusoidal patterns can be reduced by increasing the number of phase steps. However, the efficiency of the 3D imaging will decrease obviously. To handle this issue, we resort to deep learning techniques to retrieve the phase accurately from the non-sinusoidal patterns without increasing PS images. In this work, our deep neural network is constructed following the architecture of U-net [77].

U-net is a fully convolutional network with an encoder-decoder architecture, which is widely used in image segmentation. As shown in Fig. 5, the input images are captured non-sinusoidal PS fringe images. We take the three-step PS algorithm as an example as it requires the minimum images. With the non-sinusoidal PS fringe images, the network learns to predict ideal numerator  $M(x, y)$  and denominator  $D(x, y)$ , which can be represented as



**Fig. 5.** Proposed deep neural network to process non-sinusoidal fringe images.

$$M(x, y) = \sum_{n=0}^{N-1} I_n(x, y) \sin\left(\frac{2\pi n}{N}\right), \quad (7)$$

$$D(x, y) = \sum_{n=0}^{N-1} I_n(x, y) \cos\left(\frac{2\pi n}{N}\right). \quad (8)$$

According to Eq. (2),  $M(x, y)$  and  $D(x, y)$  can be fed into the arctan function to calculate the final wrapped phase. At the beginning, the input fringe images are processed by the encoder to obtain 50-channel feature tensors with 1/2 resolution reduction along both the  $x$  and  $y$  directions. Then, these feature tensors successively go through three convolutional blocks to capture the multi-level feature information.

Contrary to the encoder subnetwork, the decoder subnetwork then performs up-sampling operations to restore results of the input image's original size. It is implemented by bilinear interpolation and is followed by two convolution layers. In the U-net, at every step of the decoder, a skip connection is used to concatenate the convolution layers' output with feature maps from the encoder at the same level. This structure helps obtain low-level and high-level information at the same time and weakens the typical gradient vanishing in deep convolutional networks, which is beneficial to achieve accurate results. The last layer of the network is a convolutional layer activated by a linear activation function and outputs two-channel data consisting of the numerator and the denominator. The objective of the neural network is to minimize the following loss function:

$$\text{Loss}(\theta) = \frac{1}{HW} [\|Y^M(\theta) - G^M\|^2 + \|Y^D(\theta) - G^D\|^2], \quad (9)$$

where  $\theta$  represents the set of parameters in the neural network that is adjusted automatically during the training.  $H$  and  $W$  are the image height and width, respectively.  $Y^M$  and  $Y^D$  are the predicted numerator and denominator.  $G^M$  and  $G^D$  are the ground-truth numerator and denominator. To obtain the ground-truth data, the PS algorithm with a large number of  $N$  is exploited as it is not sensitive to non-sinusoidal patterns. From Eq. (9), the deep neural network gradually learns to map non-sinusoidal fringe images to the numerator and the denominator that are close to the ideal ones during the training.

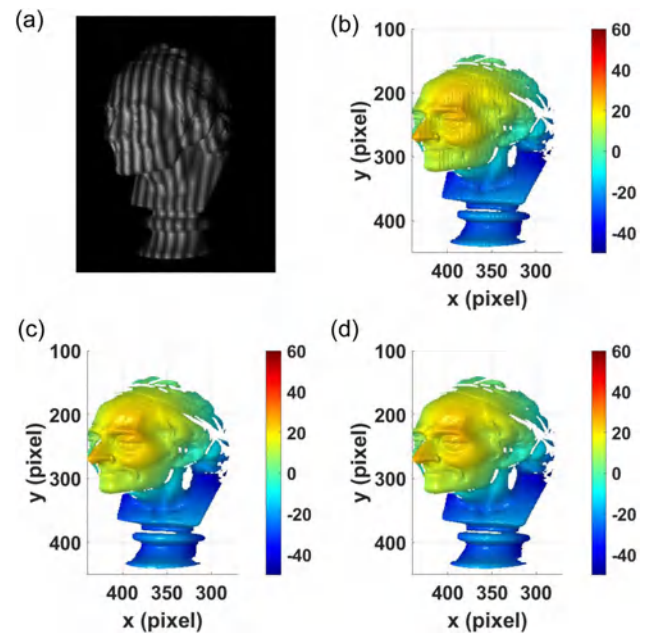
#### 4. EXPERIMENTS

To validate the proposed method, we built a structured light illumination system that consisted of a projector (DLP 4500, Texas Instruments) and an industrial camera (acA640-750  $\mu\text{m}$ , Basler). The camera was equipped with a lens of 8 mm focal length. The distance between the test object and the imaging system is about 1 m.

Non-sinusoidal fringe images due to five different causes were captured, respectively: (1) the pure gamma distortion (where  $\gamma$  was set as 2.2 during the pattern projection), (2) the pure binary defocusing projection, (3) the pure image saturation, (4) the coupling of gamma effect  $\gamma = 2.2$  with image saturation, and (5) the coupling of binary defocusing projection with image saturation. To collect the training data, we captured 750 sets of non-sinusoidal three-step PS fringe images from different objects. To obtain the ground-truth data, Eqs. (7) and (8) were applied, where  $N$  was selected as 12

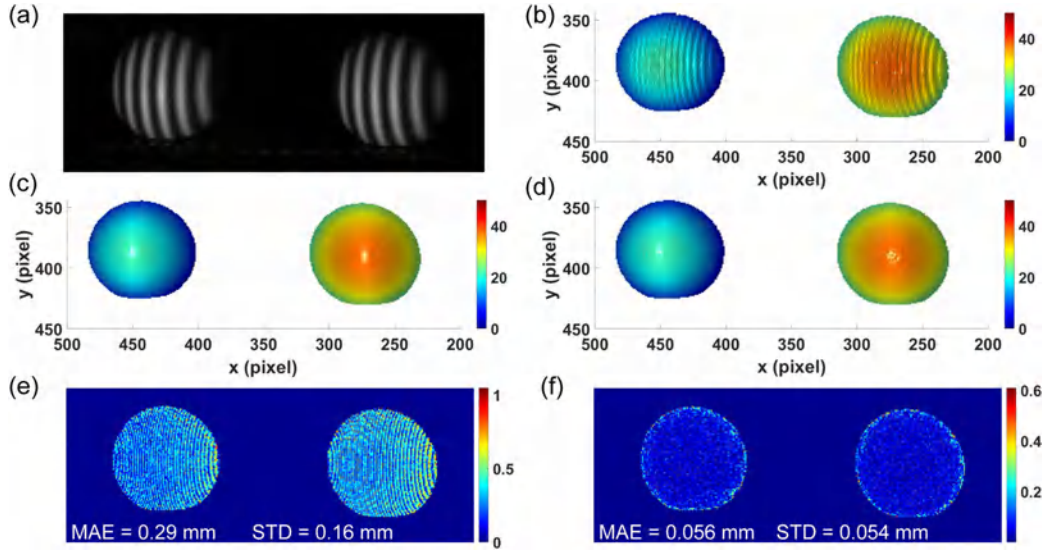
to remove the influence of the harmonics as much as possible. The pixel depth of captured three-step fringe images is 8-bit in our experiments. Before being fed into the neural network, they were divided by 255 for normalization, which can make the learning easier for the network. The neural network was implemented using the TensorFlow framework (Google) and was computed on a GTX Titan graphics card (NVIDIA). For each non-sinusoidal scenario, we trained and tested the neural network using only the data belonging to the same scenario. All of the objects used in the testing process were not present in the training stage.

First, we investigated the neural network's efficacy in the correction of gamma distortion. Figure 6(a) shows one of the captured three-step PS images. Figure 6(b) is the 3D reconstruction (depth map) obtained by the traditional three-step PS algorithm, in which obvious periodic ripple errors can be observed on the face of the retrieved model. Figures 6(c) and 6(d) demonstrate the 3D reconstructions by the proposed method and the 12-step PS algorithm, respectively. By comparison, these ripple errors have been suppressed effectively by the neural network. For quantitative evaluation, first we measured a pair of ceramic spheres. One of the captured gamma-distorted fringe images is as shown in Fig. 7(a). Figures 7(b)–7(d) demonstrate the 3D models obtained by the three-step PS method, the proposed method, and the 12-step method, respectively. The measurement error maps of the three-step PS method and the proposed method are shown in Figs. 7(e) and 7(f). The errors were calculated by referring to the high-accuracy profile obtained by the 12-step PS algorithm. With the trained neural network, the mean absolute error (MAE) and standard deviation error (STD) can be



**Fig. 6.** 3D reconstructions from fringe images that were distorted by the projector's gamma of 2.2. (a) One of the captured three-step PS images. (b) The 3D result obtained by the traditional three-step PS algorithm. (c) The 3D result obtained by the proposed method. (d) The 3D result obtained by the 12-step PS algorithm.



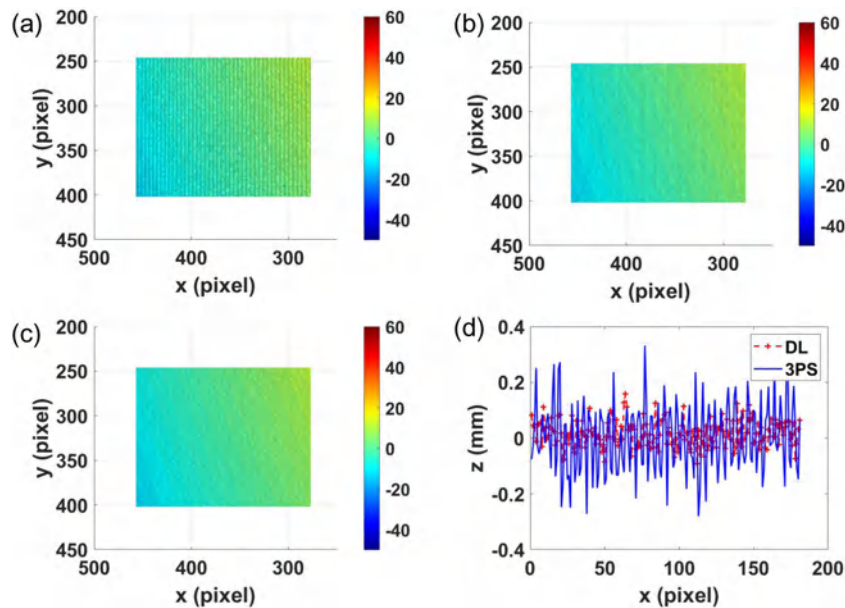


**Fig. 7.** 3D reconstructions of a pair of ceramic spheres when the projector's gamma was 2.2. (a) One of the captured three-step PS images. (b) The 3D result obtained by the traditional three-step PS algorithm. (c) The 3D result obtained by the proposed method. (d) The 3D result obtained by the 12-step PS algorithm. (e) The absolute error map of the three-step PS algorithm. (f) The absolute error map of the proposed method.

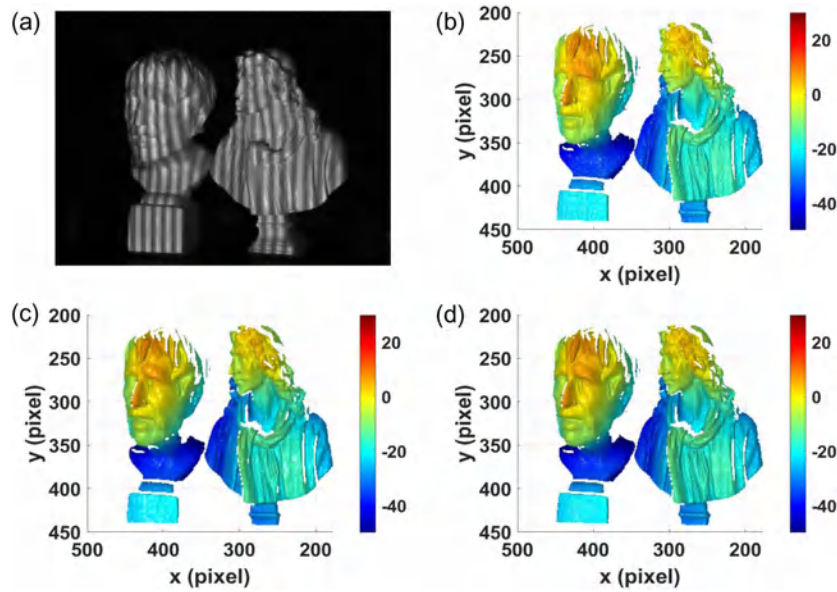
significantly reduced to 0.056 mm and 0.054 mm. Then, we measured a ceramic plate. Figures 8(a) and 8(b) show the 3D reconstruction of the traditional three-step PS algorithm and our method, respectively. The cross-section error of the plate is shown in Fig. 8(d). For the three-step PS algorithm, the MAE is 0.11 mm, and the STD is 0.075 mm. For our method, the MAE and the STD have been reduced to 0.045 mm and 0.034 mm, respectively, indicating the reduction of 59% for the MAE and 55% for the STD.

Then, the neural network was tested to obtain the phase from binary defocused fringe images. Here, we used the dith-

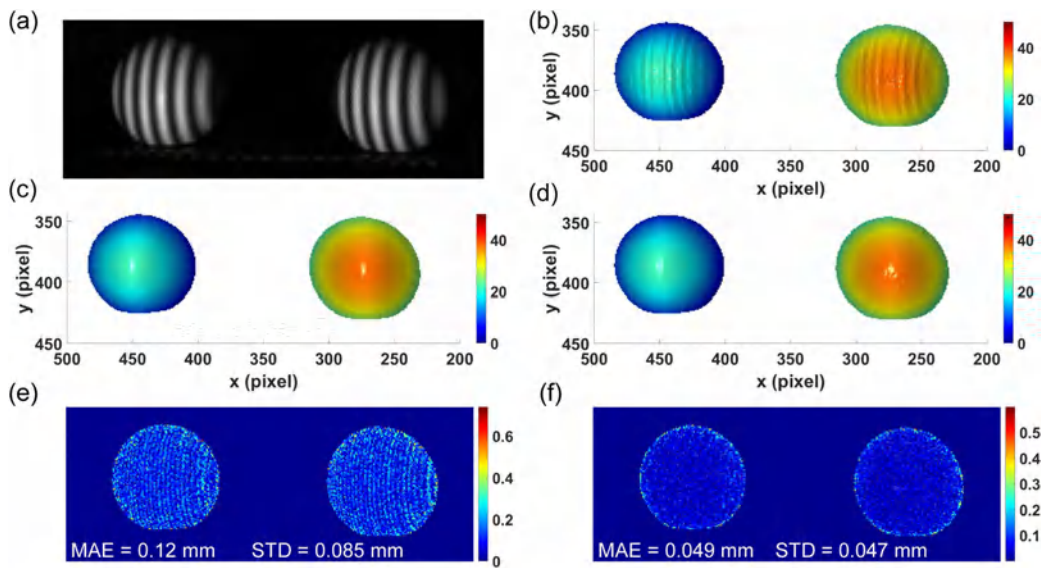
ering technique to generate the binary fringes projected with a slightly defocused projector [78]. Figure 9(a) shows one of the three-step PS patterns. The 3D reconstruction of the three-step PS method is shown in Fig. 9(b), where the surfaces have been measured with obvious stripe noise. Figures 9(c) and 9(d) demonstrate the 3D results of our method and the 12-step PS algorithm, respectively. We can see that these errors have been removed and smooth 3D reconstructions have been acquired. For the quantitative analysis, Fig. 10 shows the measurement results of a pair of ceramic spheres. The 3D result shown in Fig. 10(c) and the reconstruction error maps shown in



**Fig. 8.** 3D reconstructions of a ceramic plate when the gamma was 2.2. (a) The 3D result obtained by the traditional three-step PS algorithm (3PS). (b) The 3D result obtained by the deep-learning-based method (DL). (c) The 3D result obtained by the 12-step PS algorithm. (d) Comparison of the measurement errors of the three-step PS method and the proposed method.



**Fig. 9.** 3D reconstructions with slightly defocused binary fringe images. (a) One of the three-step PS images. (b) The 3D result obtained by the traditional three-step PS algorithm. (c) The 3D result obtained by the proposed method. (d) The 3D result obtained by the 12-step PS algorithm.

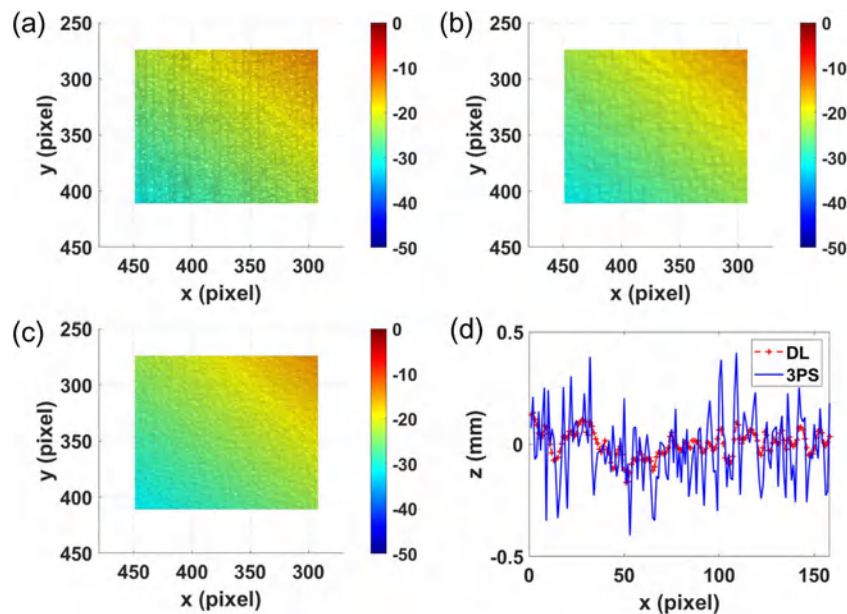


**Fig. 10.** 3D reconstructions of a pair of ceramic spheres with slightly defocused binary fringe images. (a) One of the captured three-step PS images. (b) The 3D result obtained by the traditional three-step PS algorithm. (c) The 3D result obtained by the proposed method. (d) The 3D result obtained by the 12-step PS algorithm. (e) The absolute error map of the three-step PS algorithm. (f) The absolute error map of the proposed method.

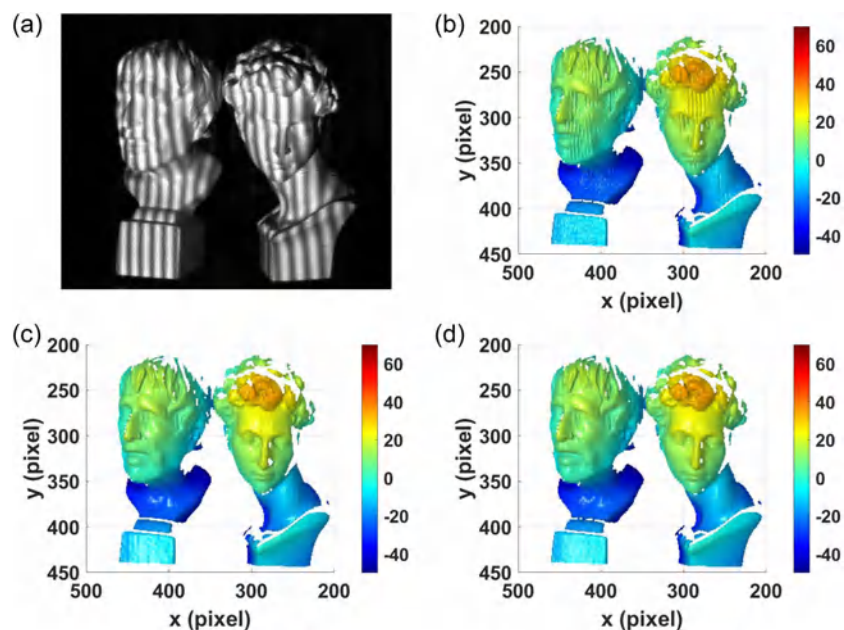
Fig. 10(f) demonstrate that the proposed method has effectively removed the periodic errors induced by non-sinusoidal components. Then, a ceramic plate was tested. Figures 11(a)–11(c) show the 3D images of the tested object. High-frequency ripple errors can be seen on the surface recovered by the three-step PS algorithm, indicating some harmonics of the projected pattern survived the defocused projection. The measurement errors are shown in Fig. 11(d). For the traditional three-step PS algorithm, the MAE and the STD are 0.12 mm and 0.096 mm, respectively. The MAE and the STD decreased to 0.046 mm and 0.034 mm, respectively,

when our method was applied, demonstrating the proposed method reduced the MAE and the STD by 62% and by 65%, respectively.

In the third experiment, the proposed neural network was used to analyze saturated fringe images. One of captured PS images is shown in Fig. 12(a), where some fringes have been captured with pure white on the two models' faces. Figure 12(b) demonstrates the 3D reconstruction by the traditional three-step PS method. Many ripple artifacts can be observed at the recovered faces of the two objects. In Fig. 12(c), with the assistance of deep learning, these errors were



**Fig. 11.** 3D reconstructions of a ceramic plate with slightly defocused binary fringe images. (a) The 3D result obtained by the traditional three-step PS algorithm (3PS). (b) The 3D result obtained by the proposed method (DL). (c) The 3D result obtained by the 12-step PS algorithm. (d) Comparison of the measurement errors of the three-step PS method and the proposed method.

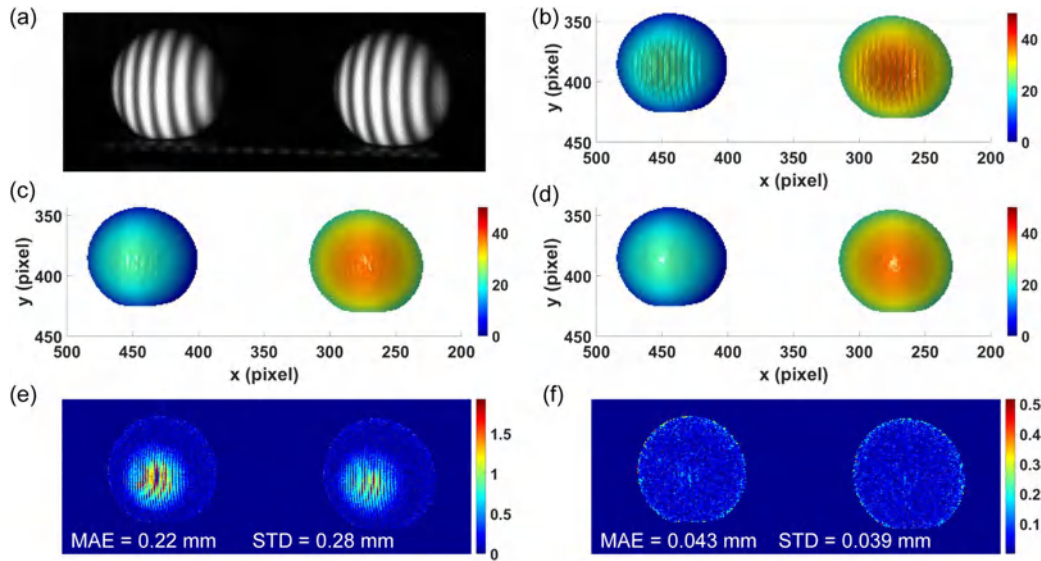


**Fig. 12.** 3D reconstructions with saturated PS images. (a) One of the captured three-step PS images. (b) The 3D result obtained by the traditional three-step PS algorithm. (c) The 3D result obtained by the proposed method. (d) The 3D result obtained by the 12-step PS algorithm.

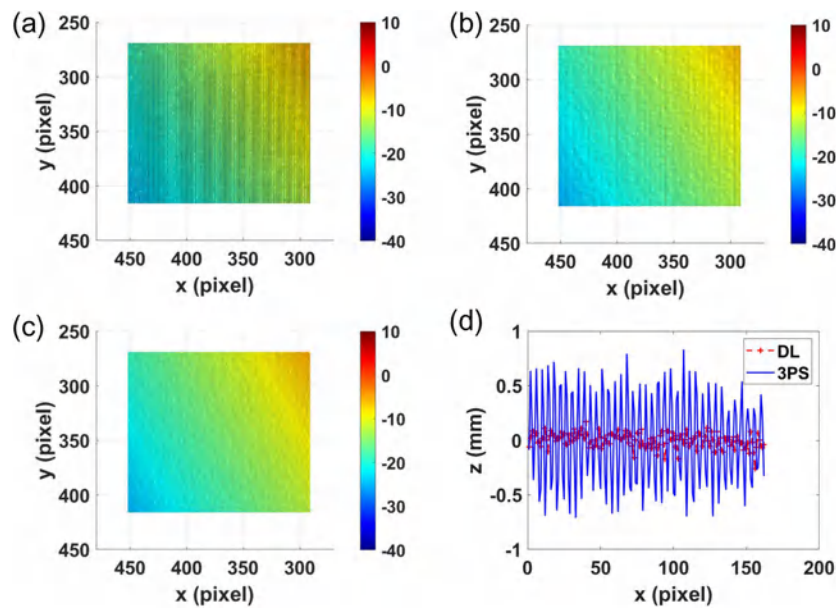
eliminated effectively by the proposed method. This reconstruction is very close to the one obtained by the 12-step PS method, as shown in Fig. 12(d). Figure 13 shows the measurement results of a pair of ceramic spheres. From the error maps demonstrated by Figs. 13(e) and 13(f), we can see that the MAE and STD have been reduced to 0.043 mm and 0.039 mm, respectively. In addition, Figs. 14(a)–14(c) show

the 3D reconstructions of a ceramic plate by the three-step PS method, the proposed method, and the 12-step PS algorithm, respectively. The measurement errors are demonstrated in Fig. 14(d). Due to the image saturation, the 3D reconstruction was distorted severely for the traditional method. Its MAE and STD are 0.34 mm and 0.19 mm, respectively. For our method, by contrast, these errors have been





**Fig. 13.** 3D reconstructions of a pair of ceramic spheres with saturated fringe images. (a) One of the captured three-step PS images. (b) The 3D result obtained by the traditional three-step PS algorithm. (c) The 3D result obtained by the proposed method. (d) The 3D result obtained by the 12-step PS algorithm. (e) The absolute error map of the three-step PS algorithm. (f) The absolute error map of the proposed method.

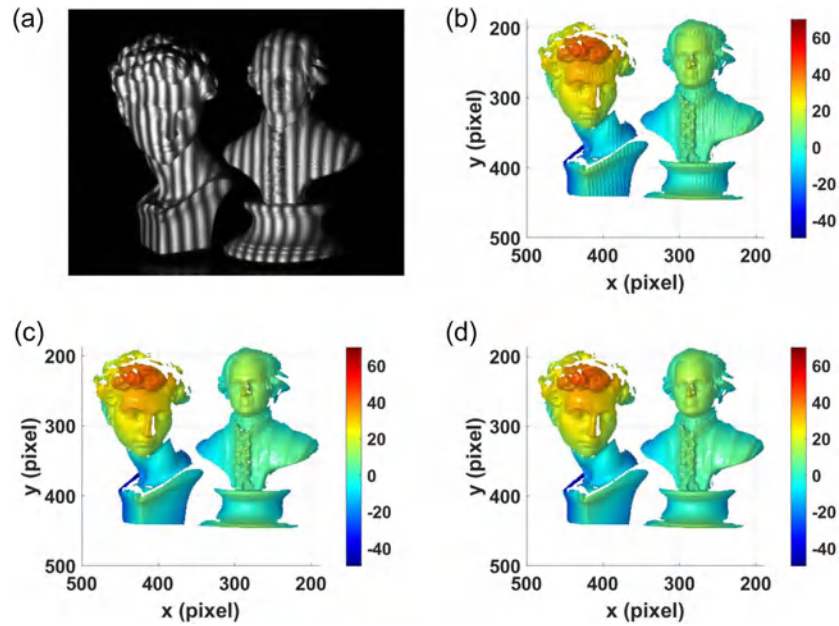


**Fig. 14.** 3D reconstructions of a ceramic plate with saturated fringe images. (a) The 3D result obtained by the traditional three-step PS algorithm (3PS). (b) The 3D result obtained by the proposed method (DL). (c) The 3D result obtained by the 12-step PS algorithm. (d) Comparison of the measurement errors of the three-step PS method and the proposed method.

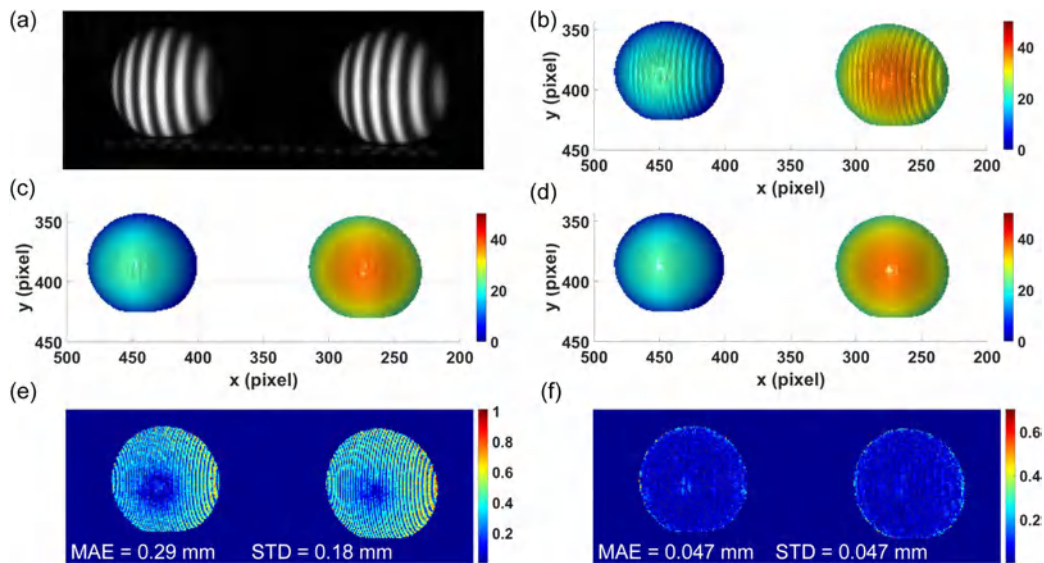
reduced to 0.052 mm and 0.039 mm, respectively, indicating the error reduction by 84.7% for MAE and by 79.5% for STD.

Next, we tested the performance of our method for a more complicated situation where the gamma distortion ( $\gamma = 2.2$ ) was coupled with the image saturation. Figure 15(a) shows one of the captured three-step PS patterns where the head of the left model was captured under effects of both the gamma distortion and the pixel saturation issues. As two non-sinusoidal factors work together, many wave artifacts can be seen in the 3D model reconstructed with the traditional three-step PS method [Fig. 15(b)]. In contrast, Figs. 15(c) and 15(d) display

the 3D results of our method and the 12-step PS method, respectively. We can see that the deep learning framework has successfully removed the influence of the gamma effect and the image saturation at the same time. In quantitative evaluation, Fig. 16 demonstrates the measurement results of a pair of ceramic spheres. Benefited from the deep learning, the coupling non-sinusoidal errors can be removed effectively. With the proposed strategy, the MAE and STD of the measured sphere can be decreased to 0.047 mm. Then, a ceramic plate was also measured. The results are shown in Figs. 17(a)–17(c). From the error distribution shown in Figs. 17(d), we can see



**Fig. 15.** 3D reconstructions under the coupling non-sinusoidal case where the gamma effect of 2.2 was coupled with the image saturation. (a) One of the captured three-step phase-shifting images. (b) The 3D result obtained by the traditional three-step PS algorithm. (c) The 3D result obtained by the proposed method. (d) The 3D result obtained by the 12-step PS algorithm.

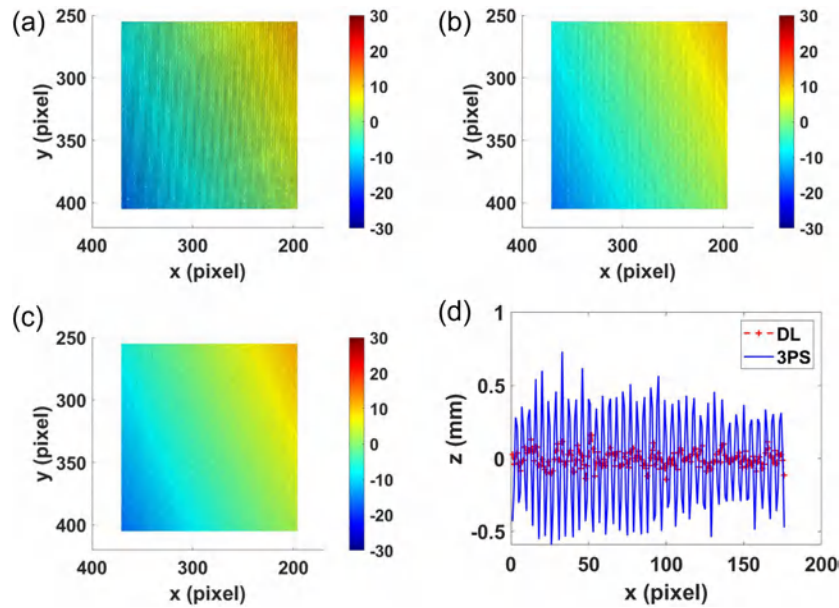


**Fig. 16.** 3D reconstructions of a pair of ceramic spheres in the coupling non-sinusoidal case where the gamma effect of 2.2 was coupled with the image saturation. (a) One of the captured three-step PS images. (b) The 3D result obtained by the traditional three-step PS algorithm. (c) The 3D result obtained by the proposed method. (d) The 3D result obtained by the 12-step PS algorithm. (e) The absolute error map of the three-step PS algorithm. (f) The absolute error map of the proposed method.

that the proposed approach can eliminate the periodic artifacts and recover the shape of the plate correctly. Numerically, the MAE and the STD of the three-step PS method are 0.28 mm and 0.16 mm, respectively. When the proposed method was applied, the MAE and the STD were reduced by 84% and 79% to 0.044 mm and 0.033 mm, respectively.

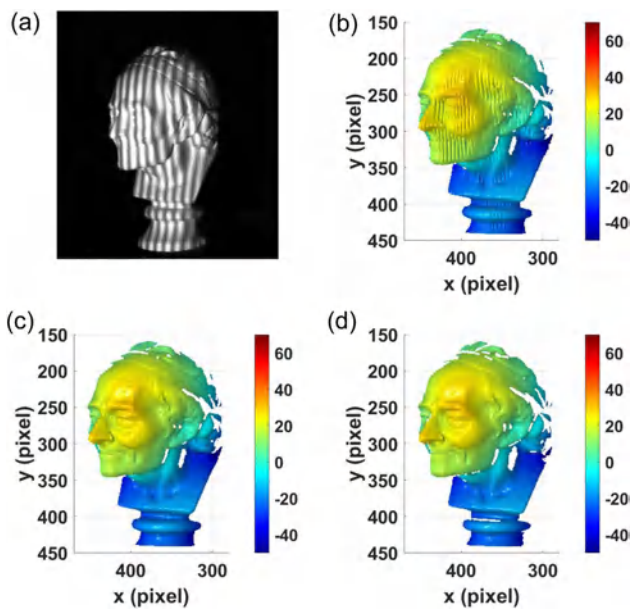
Last, we tested the second coupling non-sinusoidal case where the fringe images were captured under the slightly

defocusing projection and the image saturation. Figure 18(a) shows one of the captured three-step PS patterns in which the face was captured with defocused and saturated fringes. Figure 18(b) shows the 3D result obtained by the traditional three-step PS method; wrinkle errors due to both non-sinusoidal factors can be observed clearly. Figures 18(c) and 18(d) show the 3D reconstructions of the proposed deep neural network and the 12-step PS algorithm, respectively. As shown in



**Fig. 17.** 3D reconstructions of a ceramic plate in the coupling non-sinusoidal case where the gamma effect of 2.2 was coupled with the image saturation. (a) The 3D result obtained by the traditional three-step PS algorithm (3PS). (b) The 3D result obtained by the proposed method (DL). (c) The 3D result obtained by the 12-step PS algorithm. (d) Comparison of the measurement errors of the three-step PS method and the proposed method.

Fig. 18(c), these non-sinusoidal errors can be compensated with the proposed method effectively. A pair of ceramic spheres was then tested, and the results are shown in Fig. 19. We can see that the deep neural network is able to eliminate the ripple errors successfully and reduce the MAE and STD to



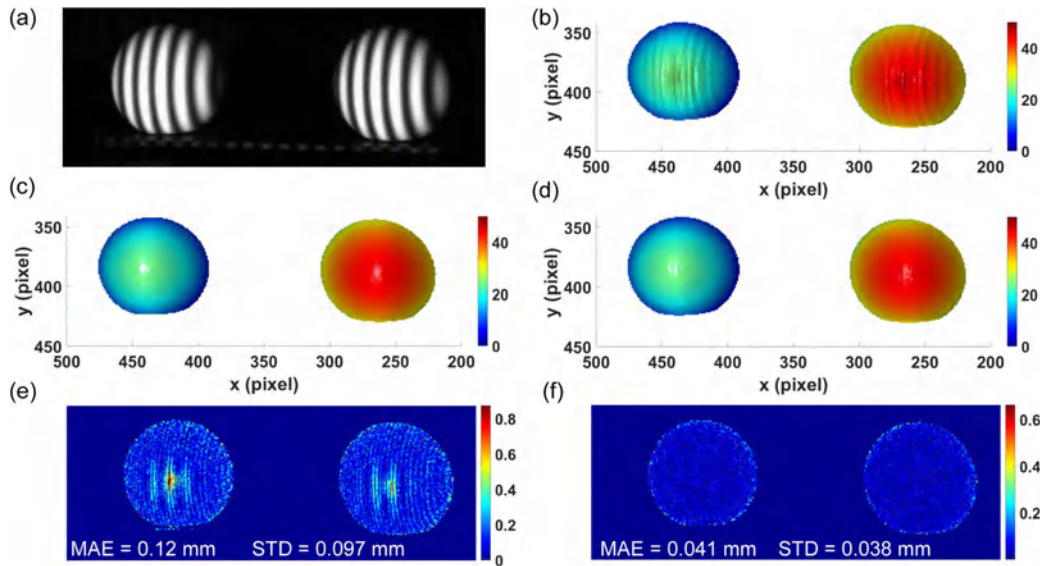
**Fig. 18.** 3D reconstructions under the coupling non-sinusoidal case where the images were projected through a slightly defocused projector and were captured with the pixel saturation. (a) One of the captured three-step phase-shifting images. (b) The 3D result obtained by the traditional three-step PS algorithm. (c) The 3D result obtained by the proposed method. (d) The 3D result obtained by the 12-step PS algorithm.

0.041 mm and 0.038 mm, respectively. Further, Figs. 20(a)–20(c) demonstrate 3D reconstructions of a ceramic plate by the traditional three-step PS method, our method, and the 12-step PS method, respectively. From the error distribution shown in Fig. 20(d), the MAE and the STD of the three-step PS method are 0.33 mm and 0.21 mm, respectively. When our method was applied, the MAE and the STD have been reduced to 0.047 mm and 0.039 mm, showing an accuracy improvement of more than 80%.

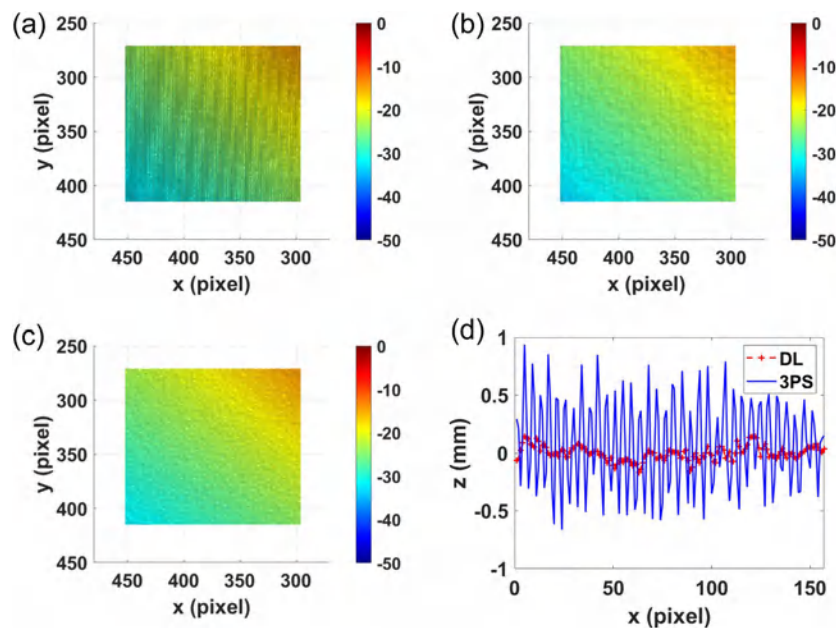
### 5. CONCLUSION

The fringe analysis is important to fringe projection profilometry, which has a high requirement on captured sinusoidal fringes. When the fringe is not a perfect sinusoid, the phase accuracy and the 3D reconstruction suffer. This paper focuses on several frequently encountered non-sinusoidal issues, including the gamma effect of digital projectors, residual high-order harmonics in binary defocusing projection, the image saturation, and more complex cases where the image saturation is coupled with the gamma effect and with the binary defocusing projection. Conventionally, these non-sinusoidal patterns can be represented by a generalized model and the corresponding phase errors can be relieved by increasing the number of phase shift in the PS algorithms. We proposed a unified deep learning technique that can analyze fringe images from all of the mentioned non-sinusoidal causes and their coupling scenarios. More importantly, to remove these phase errors without increasing the number of phase shift, we train a deep neural network that can mimic the phase correction of PS





**Fig. 19.** 3D reconstructions of a pair of ceramic spheres under the coupling non-sinusoidal case where the effect of the slightly defocusing projection was coupled with the image saturation. (a) One of the captured three-step PS images. (b) The 3D result obtained by the traditional three-step PS algorithm. (c) The 3D result obtained by the proposed method. (d) The 3D result obtained by the 12-step PS algorithm. (e) The absolute error map of the three-step PS algorithm. (f) The absolute error map of the proposed method.



**Fig. 20.** 3D reconstructions of a ceramic plate under the coupling non-sinusoidal case where the effect of the slightly defocusing projection was coupled with the image saturation. (a) The 3D result obtained by the traditional three-step PS algorithm (3PS). (b) The 3D result obtained by the proposed method (DL). (c) The 3D result obtained by the 12-step PS algorithm. (d) Comparison of the measurement errors of the three-step PS method and the proposed method.

algorithms with many steps (e.g., the 12-step PS method) by using PS fringe images captured with a few-step PS method (e.g., the three-step PS method). Experimental results have shown that compared with the traditional PS algorithm, the proposed method can effectively suppress the phase error

due to the gamma effect of projectors, insufficient defocusing of binary fringe projection, the image saturation, and two complex coupled non-sinusoidal cases without increasing the fringe images. We believe this method shows great potential for robust and accurate phase retrieval and 3D measurements.

**Funding.** National Natural Science Foundation of China (61722506, 62075096); Leading Technology of Jiangsu Basic Research Plan (BK20192003); Jiangsu Provincial “One Belt and One Road” Innovation Cooperation Project (BZ2020007); Final Assembly “13th Five-Year Plan” Advanced Research Project of China (30102070102); Equipment Advanced Research Fund of China (61404150202); Jiangsu Provincial Key Research and Development Program (BE2017162); Outstanding Youth Foundation of Jiangsu Province of China (BK20170034); National Defense Science and Technology Foundation of China (2019-JCJQ-JJ-381); “333 Engineering” Research Project of Jiangsu Province (BRA2016407); Fundamental Research Funds for the Central Universities (30920032101).

**Disclosures.** The authors declare no conflicts of interest.

## REFERENCES

1. K. Harding, “Industrial metrology: engineering precision,” *Nat. Photonics* **2**, 667–669 (2008).
2. J. M. Schmitt, “Optical coherence tomography (OCT): a review,” *IEEE J. Sel. Top. Quantum Electron.* **5**, 1205–1215 (1999).
3. J. Han, L. Shao, D. Xu, and J. Shotton, “Enhanced computer vision with Microsoft Kinect sensor: a review,” *IEEE Trans. Cybern.* **43**, 1290–1334 (2013).
4. P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, “A 3D face model for pose and illumination invariant face recognition,” in *6th IEEE International Conference on Advanced Video and Signal Based Surveillance* (IEEE, 2009), pp. 296–301.
5. M. M. P. A. Vermeulen, P. Rosielle, and P. Schellekens, “Design of a high-precision 3D-coordinate measuring machine,” *CIRP Ann.* **47**, 447–450 (1998).
6. F. Chen, G. M. Brown, and M. Song, “Overview of 3-D shape measurement using optical methods,” *Opt. Eng.* **39**, 10–22 (2000).
7. R. Leach, *Optical Measurement of Surface Topography* (Springer, 2011), Vol. **14**.
8. J. Geng, “Structured-light 3D surface imaging: a tutorial,” *Adv. Opt. Photonics* **3**, 128–160 (2011).
9. J. Salvi, J. Pagès, and J. Battle, “Pattern codification strategies in structured light systems,” *Pattern Recogn.* **37**, 827–849 (2004).
10. C. Zuo, S. Feng, L. Huang, T. Tao, W. Yin, and Q. Chen, “Phase shifting algorithms for fringe projection profilometry: a review,” *Opt. Lasers Eng.* **109**, 23–59 (2018).
11. L. Zhang, B. Curless, and S. Seitz, “Rapid shape acquisition using color structured light and multi-pass dynamic programming,” in *1st International Symposium on 3D Data Processing Visualization and Transmission* (IEEE Computer Society, 2002), pp. 24–36.
12. M. Schaffer, M. Grosse, B. Harendt, and R. Kowarschik, “High-speed three-dimensional shape measurements of objects with laser speckles and acousto-optical deflection,” *Opt. Lett.* **36**, 3097–3099 (2011).
13. S. Heist, P. Lutzke, I. Schmidt, P. Dietrich, P. Kühmstedt, A. Tünnermann, and G. Notni, “High-speed three-dimensional shape measurement using GOBO projection,” *Opt. Lasers Eng.* **87**, 90–96 (2016).
14. M. Takeda and K. Mutoh, “Fourier transform profilometry for the automatic measurement of 3-D object shapes,” *Appl. Opt.* **22**, 3977–3982 (1983).
15. X. Su and Q. Zhang, “Dynamic 3-D shape measurement method: a review,” *Opt. Lasers Eng.* **48**, 191–204 (2010).
16. Q. Kemao, “Two-dimensional windowed Fourier transform for fringe pattern analysis: principles, applications and implementations,” *Opt. Lasers Eng.* **45**, 304–317 (2007).
17. J. Zhong and J. Weng, “Spatial carrier-fringe pattern analysis by means of wavelet transform: wavelet transform profilometry,” *Appl. Opt.* **43**, 4993–4998 (2004).
18. L. Huang, Q. Kemao, B. Pan, and A. K. Asundi, “Comparison of Fourier transform, windowed Fourier transform, and wavelet transform methods for phase extraction from a single fringe pattern in fringe projection profilometry,” *Opt. Lasers Eng.* **48**, 141–148 (2010).
19. V. Srinivasan, H.-C. Liu, and M. Halioua, “Automated phase-measuring profilometry of 3-D diffuse objects,” *Appl. Opt.* **23**, 3105–3108 (1984).
20. P. S. Huang, Q. J. Hu, and F.-P. Chiang, “Double three-step phase-shifting algorithm,” *Appl. Opt.* **41**, 4503–4509 (2002).
21. P. Hariharan, B. Oreb, and T. Eiju, “Digital phase-shifting interferometry: a simple error-compensating phase calculation algorithm,” *Appl. Opt.* **26**, 2504–2506 (1987).
22. S. Zhang and S.-T. Yau, “High-speed three-dimensional shape measurement system using a modified two-plus-one phase-shifting algorithm,” *Opt. Eng.* **46**, 113603 (2007).
23. P. Jia, J. Kofman, and C. E. English, “Two-step triangular-pattern phase-shifting method for three-dimensional object-shape measurement,” *Opt. Eng.* **46**, 083201 (2007).
24. P. S. Huang, S. Zhang, and F.-P. Chiang, “Trapezoidal phase-shifting method for three-dimensional shape measurement,” *Opt. Eng.* **44**, 123601 (2005).
25. T. Anna, S. K. Dubey, C. Shakher, A. Roy, and D. S. Mehta, “Sinusoidal fringe projection system based on compact and non-mechanical scanning low-coherence michelson interferometer for three-dimensional shape measurement,” *Opt. Commun.* **282**, 1237–1242 (2009).
26. Y. Guan, Y. Yin, A. Li, X. Liu, and X. Peng, “Dynamic 3D imaging based on acousto-optic heterodyne fringe interferometry,” *Opt. Lett.* **39**, 3678–3681 (2014).
27. S. Yoneyama, Y. Morimoto, M. Fujigaki, and M. Yabe, “Phase-measuring profilometry of moving object without phase-shifting device,” *Opt. Lasers Eng.* **40**, 153–161 (2003).
28. C. Zuo, Q. Chen, G. Gu, S. Feng, and F. Feng, “High-speed three-dimensional profilometry for multiple objects with complex shapes,” *Opt. Express* **20**, 19493–19510 (2012).
29. S. Ma, C. Quan, R. Zhu, L. Chen, B. Li, and C. Tay, “A fast and accurate gamma correction based on Fourier spectrum analysis for digital fringe projection profilometry,” *Opt. Commun.* **285**, 533–538 (2012).
30. K. Liu, Y. Wang, D. L. Lau, Q. Hao, and L. G. Hassebrook, “Gamma model and its analysis for phase measuring profilometry,” *J. Opt. Soc. Am. A* **27**, 553–562 (2010).
31. S. Zhang and S.-T. Yau, “Generic nonsinusoidal phase error correction for three-dimensional shape measurement using a digital video projector,” *Appl. Opt.* **46**, 36–43 (2007).
32. Z. Li, Y. Shi, C. Wang, and Y. Wang, “Accurate calibration method for a structured light system,” *Opt. Eng.* **47**, 053604 (2008).
33. B. Pan, Q. Kemao, L. Huang, and A. Asundi, “Phase error analysis and compensation for nonsinusoidal waveforms in phase-shifting digital fringe projection profilometry,” *Opt. Lett.* **34**, 416–418 (2009).
34. H. Guo, H. He, and M. Chen, “Gamma correction for digital fringe projection profilometry,” *Appl. Opt.* **43**, 2906–2914 (2004).
35. T. Hoang, B. Pan, D. Nguyen, and Z. Wang, “Generic gamma correction for accuracy enhancement in fringe-projection profilometry,” *Opt. Lett.* **35**, 1992–1994 (2010).
36. C. Jiang, S. Xing, and H. Guo, “Fringe harmonics elimination in multi-frequency phase-shifting fringe projection profilometry,” *Opt. Express* **28**, 2838–2856 (2020).
37. B. Li, Y. Wang, J. Dai, W. Lohry, and S. Zhang, “Some recent advances on superfast 3D shape measurement with digital binary defocusing techniques,” *Opt. Lasers Eng.* **54**, 236–246 (2014).
38. H. Fujita, K. Yamatan, M. Yamamoto, Y. Otani, A. Suguro, S. Morokawa, and T. Yoshizawa, “Three-dimensional profilometry using liquid crystal grating,” *Proc. SPIE* **5058**, 51–60 (2003).
39. T. Yoshizawa and H. Fujita, “Liquid crystal grating for profilometry using structured light,” *Proc. SPIE* **6000**, 60000H (2005).
40. G. A. Ayubi, J. A. Ayubi, J. M. Di Martino, and J. A. Ferrari, “Pulse-width modulation in defocused three-dimensional fringe projection,” *Opt. Lett.* **35**, 3682–3684 (2010).

41. C. Zuo, Q. Chen, S. Feng, F. Feng, G. Gu, and X. Sui, "Optimized pulse width modulation pattern strategy for three-dimensional profilometry with projector defocusing," *Appl. Opt.* **51**, 4477–4490 (2012).
42. Y. Wang and S. Zhang, "Optimal pulse width modulation for sinusoidal fringe generation with projector defocusing," *Opt. Lett.* **35**, 4121–4123 (2010).
43. J. Sun, C. Zuo, S. Feng, S. Yu, Y. Zhang, and Q. Chen, "Improved intensity-optimized dithering technique for 3D shape measurement," *Opt. Lasers Eng.* **66**, 158–164 (2015).
44. W. Lohry and S. Zhang, "Genetic method to optimize binary dithering technique for high-quality fringe generation," *Opt. Lett.* **38**, 540–542 (2013).
45. S. Feng, L. Zhang, C. Zuo, T. Tao, Q. Chen, and G. Gu, "High dynamic range 3D measurements with fringe projection profilometry: a review," *Meas. Sci. Technol.* **29**, 122001 (2018).
46. S. Feng, Y. Zhang, Q. Chen, C. Zuo, R. Li, and G. Shen, "General solution for high dynamic range three-dimensional shape measurement using the fringe projection technique," *Opt. Lasers Eng.* **59**, 56–71 (2014).
47. S. Zhang and S.-T. Yau, "High dynamic range scanning technique," *Opt. Eng.* **48**, 033604 (2009).
48. Z. Song, H. Jiang, H. Lin, and S. Tang, "A high dynamic range structured light means for the 3D measurement of specular surface," *Opt. Lasers Eng.* **95**, 8–16 (2017).
49. S. Feng, Q. Chen, C. Zuo, and A. Asundi, "Fast three-dimensional measurements for dynamic scenes with shiny surfaces," *Opt. Commun.* **382**, 18–27 (2017).
50. C. Waddington and J. Kofman, "Saturation avoidance by adaptive fringe projection in phase-shifting 3D surface-shape measurement," in *International Symposium on Optomechatronic Technologies* (IEEE, 2010), pp. 1–4.
51. L. Zhang, Q. Chen, C. Zuo, and S. Feng, "High dynamic range 3D shape measurement based on the intensity response function of a camera," *Appl. Opt.* **57**, 1378–1386 (2018).
52. H. Lin, J. Gao, Q. Mei, Y. He, J. Liu, and X. Wang, "Adaptive digital fringe projection technique for high dynamic range three-dimensional shape measurement," *Opt. Express* **24**, 7703–7718 (2016).
53. Z. Cai, X. Liu, X. Peng, Y. Yin, A. Li, J. Wu, and B. Z. Gao, "Structured light field 3D imaging," *Opt. Express* **24**, 20324–20334 (2016).
54. V. Suresh, Y. Wang, and B. Li, "High-dynamic-range 3D shape measurement utilizing the transitioning state of digital micromirror device," *Opt. Lasers Eng.* **107**, 176–181 (2018).
55. H. Jiang, H. Zhao, and X. Li, "High dynamic range fringe acquisition: a novel 3-D scanning technique for high-reflective surfaces," *Opt. Lasers Eng.* **50**, 1484–1493 (2012).
56. L. Zhang, Q. Chen, C. Zuo, and S. Feng, "Real-time high dynamic range 3D measurement using fringe projection," *Opt. Express* **28**, 24363–24378 (2020).
57. L. Zhang, Q. Chen, C. Zuo, and S. Feng, "High-speed high dynamic range 3D shape measurement based on deep learning," *Opt. Lasers Eng.* **134**, 106245 (2020).
58. J. H. Bruning, D. R. Herriott, J. Gallagher, D. Rosenfeld, A. White, and D. Brangaccio, "Digital wavefront measuring interferometer for testing optical surfaces and lenses," *Appl. Opt.* **13**, 2693–2703 (1974).
59. Y. Yin, Z. Cai, H. Jiang, X. Meng, J. Xi, and X. Peng, "High dynamic range imaging for fringe projection profilometry with single-shot raw data of the color camera," *Opt. Lasers Eng.* **89**, 138–144 (2017).
60. M. Wang, G. Du, C. Zhou, C. Zhang, S. Si, H. Li, Z. Lei, and Y. Li, "Enhanced high dynamic range 3D shape measurement based on generalized phase-shifting algorithm," *Opt. Commun.* **385**, 43–53 (2017).
61. Y. Chen, Y. He, and E. Hu, "Phase deviation analysis and phase retrieval for partial intensity saturation in phase-shifting projected fringe profilometry," *Opt. Commun.* **281**, 3087–3090 (2008).
62. Z. Qi, Z. Wang, J. Huang, C. Xing, and J. Gao, "Error of image saturation in the structured-light method," *Appl. Opt.* **57**, A181–A188 (2018).
63. P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, and L. Bottou, "Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.* **11**, 3371–3408 (2010).
64. D. J. Im, S. Ahn, R. Memisevic, and Y. Bengio, "Denoising criterion for variational auto-encoding framework," in *AAAI Conference on Artificial Intelligence* (2017), pp. 2059–2062.
65. Y. Kiarashinejad, M. Zandehshahvar, S. Abdollahramezani, O. Hemmatyar, R. Pourabolghasem, and A. Adibi, "Knowledge discovery in nanophotonics using geometric deep learning," *Adv. Intell. Syst.* **2**, 1900132 (2020).
66. Y. Rivenson, Z. Göröcs, H. Günaydin, Y. Zhang, H. Wang, and A. Ozcan, "Deep learning microscopy," *Optica* **4**, 1437–1443 (2017).
67. Y. Rivenson, Y. Zhang, H. Günaydn, D. Teng, and A. Ozcan, "Phase recovery and holographic image reconstruction using deep learning in neural networks," *Light Sci. Appl.* **7**, 17141 (2018).
68. M. Lyu, W. Wang, H. Wang, H. Wang, G. Li, N. Chen, and G. Situ, "Deep-learning-based ghost imaging," *Sci. Rep.* **7**, 17865 (2017).
69. Y. Li, Y. Xue, and L. Tian, "Deep speckle correlation: a deep learning approach toward scalable imaging through scattering media," *Optica* **5**, 1181–1190 (2018).
70. C. S. Lee, A. J. Tyring, N. P. Deruyter, Y. Wu, A. Rokem, and A. Y. Lee, "Deep-learning based, automated segmentation of macular edema in optical coherence tomography," *Biomed. Opt. Express* **8**, 3440–3448 (2017).
71. S. Feng, Q. Chen, G. Gu, T. Tao, L. Zhang, Y. Hu, W. Yin, and C. Zuo, "Fringe pattern analysis using deep learning," *Adv. Photonics* **1**, 025001 (2019).
72. J. Qian, S. Feng, Y. Li, T. Tao, J. Han, Q. Chen, and C. Zuo, "Single-shot absolute 3D shape measurement with deep-learning-based color fringe projection profilometry," *Opt. Lett.* **45**, 1842–1845 (2020).
73. J. Shi, X. Zhu, H. Wang, L. Song, and Q. Guo, "Label enhanced and patch based deep learning for phase retrieval from single frame fringe pattern in fringe projection 3D measurement," *Opt. Express* **27**, 28929–28943 (2019).
74. T. Yang, Z. Zhang, H. Li, X. Li, and X. Zhou, "Single-shot phase extraction for fringe projection profilometry using deep convolutional generative adversarial network," *Meas. Sci. Technol.* **32**, 015007 (2020).
75. W. Yin, Q. Chen, S. Feng, T. Tao, L. Huang, M. Trusiak, A. Asundi, and C. Zuo, "Temporal phase unwrapping using deep learning," *Sci. Rep.* **9**, 20175 (2019).
76. J. Qian, S. Feng, T. Tao, Y. Hu, Y. Li, Q. Chen, and C. Zuo, "Deep-learning-enabled geometric constraints and phase unwrapping for single-shot absolute 3D shape measurement," *APL Photonics* **5**, 046105 (2020).
77. O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2015), pp. 234–241.
78. C. Zuo, T. Tao, S. Feng, L. Huang, A. Asundi, and Q. Chen, "Micro Fourier transform profilometry ( $\mu$ ftp): 3D shape measurement at 10,000 frames per second," *Opt. Lasers Eng.* **102**, 70–91 (2018).





# Fringe-pattern analysis with ensemble deep learning

Shijie Feng,<sup>a,b,c</sup> Yile Xiao,<sup>a,b,c</sup> Wei Yin,<sup>a,b,c</sup> Yan Hu,<sup>a,b,c</sup> Yixuan Li,<sup>a,b,c</sup> Chao Zuo<sup>Ⓢ,a,b,c,\*</sup> and Qian Chen<sup>a,b,\*</sup>

<sup>a</sup>Nanjing University of Science and Technology, Smart Computational Imaging Laboratory, Nanjing, China

<sup>b</sup>Nanjing University of Science and Technology, Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense, Nanjing, China

<sup>c</sup>Smart Computational Imaging Research Institute of Nanjing University of Science and Technology, Nanjing, China

**Abstract.** In recent years, there has been tremendous progress in the development of deep-learning-based approaches for optical metrology, which introduce various deep neural networks (DNNs) for many optical metrology tasks, such as fringe analysis, phase unwrapping, and digital image correlation. However, since different DNN models have their own strengths and limitations, it is difficult for a single DNN to make reliable predictions under all possible scenarios. In this work, we introduce ensemble learning into optical metrology, which combines the predictions of multiple DNNs to significantly enhance the accuracy and reduce the generalization error for the task of fringe-pattern analysis. First, several state-of-the-art base models of different architectures are selected. A  $K$ -fold average ensemble strategy is developed to train each base model multiple times with different data and calculate the mean prediction within each base model. Next, an adaptive ensemble strategy is presented to further combine the base models by building an extra DNN to fuse the features extracted from these mean predictions in an adaptive and fully automatic way. Experimental results demonstrate that ensemble learning could attain superior performance over state-of-the-art solutions, including both classic and conventional single-DNN-based methods. Our work suggests that by resorting to collective wisdom, ensemble learning offers a simple and effective solution for overcoming generalization challenges and boosts the performance of data-driven optical metrology methods.

Keywords: optical metrology; fringe-pattern analysis; deep learning; ensemble learning; three-dimensional measurement; phase retrieval.

Received Dec. 28, 2022; revised manuscript received Apr. 11, 2023; accepted for publication Apr. 20, 2023; published online May 17, 2023.

© The Authors. Published by SPIE and CLP under a Creative Commons Attribution 4.0 International License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI.

[DOI: [10.1117/1.APN.2.3.036010](https://doi.org/10.1117/1.APN.2.3.036010)]

## 1 Introduction

Optical metrology plays a significant role in many fields because of its merits of noninvasiveness, flexibility, and high accuracy. In optical metrology, fringe-pattern analysis is indispensable to many tasks, e.g., interferometry, fringe projection profilometry, and digital holography. According to the number of patterns used, fringe-pattern analysis can be generally classified into two categories: single-frame and multiframe methods. The Fourier-transform fringe-pattern analysis is a representative single-frame approach<sup>1</sup> that converts a fringe pattern into the frequency domain and extracts the phase information by filtering

the first order of the spectrum. This method is suitable for measuring dynamic scenes because it only needs a single fringe image. However, it tends to compromise on handling complex surfaces, owing to the spectrum aliasing issue. In contrast, the multiframe approaches, e.g., the  $N$ -step phase-shifting (PS) algorithm,<sup>2</sup> can achieve higher accuracy, since the phase demodulation can be carried out pixel by pixel along the temporal axis. Nevertheless, multiframe approaches usually suffer when facing fast-moving objects because of the need to capture multiple images. Hence, there is a contradiction between the efficiency and the accuracy of the fringe-pattern analysis.

Recently, many advances have emerged in the field of optical metrology that benefit from harnessing the power of deep learning.<sup>3,4</sup> Fringe-pattern analysis using deep learning has shown promising performance in measuring complex contours

\*Address all correspondence to Chao Zuo, [zuochoa@njjust.edu.cn](mailto:zuochoa@njjust.edu.cn); Qian Chen, [chenqian@njjust.edu.cn](mailto:chenqian@njjust.edu.cn)

using a single fringe image.<sup>5</sup> As a data-driven approach, it can exploit useful hidden clues that may be overlooked by traditional physical models, thus showing potential for resolving the contradiction between efficiency and accuracy in the phase demodulation. However, it is not trouble-free for this kind of approach. Usually, people adopt a single deep neural network (DNN) and depend on it completely to handle all possible measurements once it is trained. Actually, this is risky, as the DNN may only learn limited attributes of input data because of its fixed structure. Consequently, it tends to demonstrate high variance for unseen scenarios. Further, the DNN may converge to a local loss minimum during training, which further increases the risk of making unreliable predictions.

To handle these issues, ensemble deep learning has been developed,<sup>6,7</sup> which refers to a set of strategies where, rather than relying on a single model, several base models are combined to perform tasks. As different architectures can capture distinct information, better decisions can be made by combining different networks. Inspired by recent successful applications of ensemble deep learning, we demonstrate for the first time, to the best of our knowledge, that an ensemble of multiple deep-learning models can improve the accuracy and the stability of fringe-pattern analysis substantially. First, multiple state-of-the-art DNNs for fringe-pattern analysis are employed as base models. To train the base models, we propose a  $K$ -fold average ensemble method to divide training data into several groups so that each one can be trained multiple times by using different data. Then, the average of the predictions is calculated as the output of each base model. To further fuse the outputs of the base models, we develop an adaptive ensemble that trains an extra DNN to extract and combine useful features from these outputs adaptively and automatically during training. Experimental results show that the proposed approach can improve the phase accuracy and the generalization capability for unseen scenarios greatly compared with the traditional method using a single model.

## 2 Methods

In fringe-pattern analysis, a fringe image is often written as

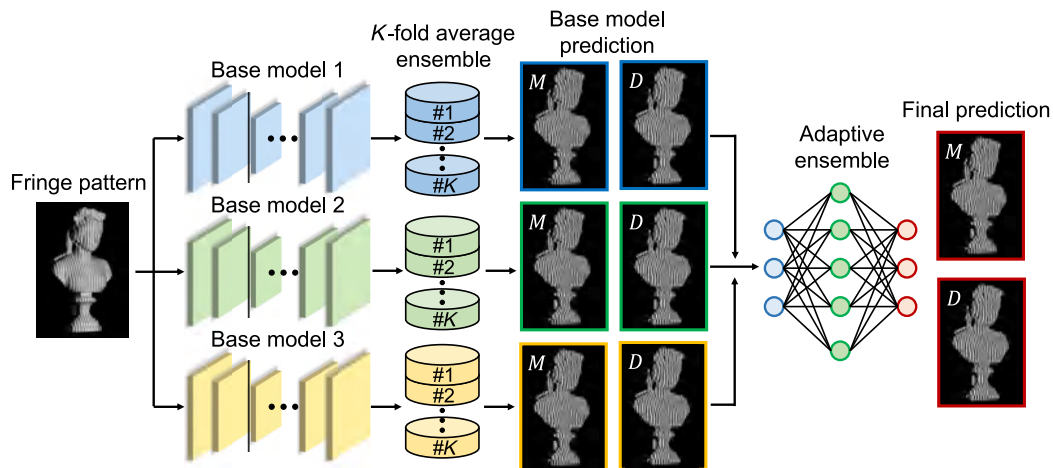
$$I(x, y) = A(x, y) + B(x, y) \cos \varphi(x, y), \quad (1)$$

where  $(x, y)$  is the pixel coordinate,  $A$  is the background signal,  $B$  is the amplitude, and  $\varphi$  is the phase to be measured. Conventionally, the phase is demodulated through an arctangent function,

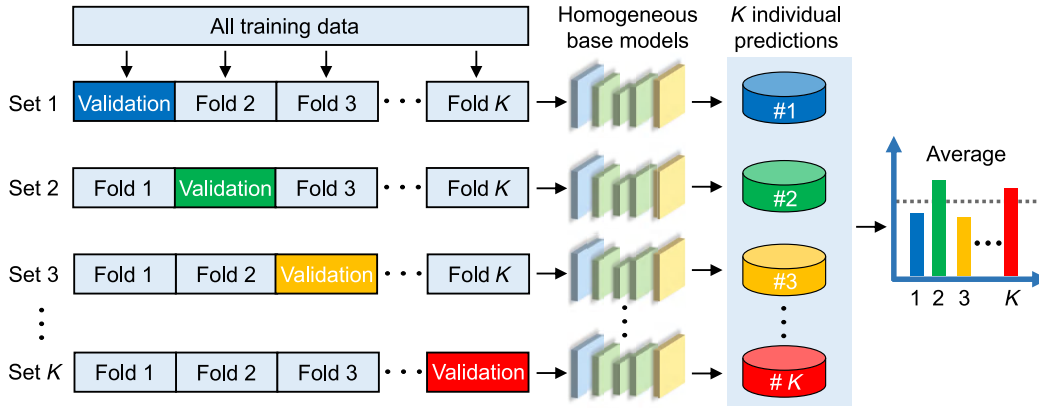
$$\varphi(x, y) = \arctan \frac{cB(x, y) \sin \varphi(x, y)}{cB(x, y) \cos \varphi(x, y)} = \arctan \frac{M(x, y)}{D(x, y)}, \quad (2)$$

where the numerator  $M$  represents the phase sine  $[\sin \varphi(x, y)]$  and the denominator  $D$  represents the phase cosine  $[\cos \varphi(x, y)]$ .  $c$  is a constant that is determined according to the phase demodulation approach. According to Eq. (2), a DNN can be constructed to learn to predict  $M$  and  $D$ . Then, the phase  $\varphi$  can be computed through the arctangent function.<sup>5</sup>

Instead of relying on a single model, we train several base models to analyze the same input fringe image and combine their outputs as the final prediction. Figure 1 demonstrates the diagram of the proposed framework. First, three state-of-the-art models for fringe-pattern analysis are selected as base models. The first two models are the U-Net<sup>8</sup> and the multipath DNN (MP DNN),<sup>5</sup> which are convolutional neural networks that are good at extracting local features. The third model is the Swin-Unet,<sup>9</sup> which is a vision transformer that shows the advantage of capturing global information. The structures of base models are detailed in the [Supplementary Material](#). As these models have different architectures, diverse attributes of the input data can be learned. To train the base models, we develop a  $K$ -fold average ensemble, whose schematic is shown in Fig. 2. The whole training data set is divided into  $K$  parts equally (i.e., from fold 1 to fold  $K$ ). Any  $K - 1$  parts of the data can



**Fig. 1** Diagram of the fringe-pattern analysis using ensemble deep learning. The input fringe image is processed by three base models. In each base model, a  $K$ -fold average ensemble is proposed to generate  $K$  sets of data to train  $K$  homogeneous models. Each homogeneous model outputs a pair of numerator  $M$  and denominator  $D$ . The mean is computed over  $K$  homogeneous models and is treated as the output of the base model. To further combine the predictions of the base models, an adaptive ensemble is developed that trains a DNN to fuse their predictions adaptively and gives the final prediction.



**Fig. 2** Diagram of the  $K$ -fold average ensemble approach. The whole data set is equally separated into  $K$  parts. We combine any  $K - 1$  parts of the data for training and leave the remaining part for validation. Then,  $K$  sets of data can be generated to train a base model, which yields  $K$  homogeneous models. Each one gives a prediction independently, and their average is calculated as the output of the  $K$ -fold average ensemble.

be merged and then used for training; the remaining one is used for validation. In this way, we can generate  $K$  sets of training data. As each of them is different, additional information can be provided. To train these base models, we use the following mean squared error loss function:

$$\text{Loss}(\theta^i) = \frac{1}{H \times W} \sum_{h=1}^H \sum_{w=1}^W (y_{h,w}^i - \hat{y}_{h,w}^i)^2, \quad (3)$$

where  $\theta^i$  represents the parameters of the  $i$ th base model; these are learned during the training process.  $H$  and  $W$  represent the height and width of the image in pixels, respectively. Omitting the pixel index,  $y^i$  is the output of the base model that consists of a pair of estimated numerator and denominator.  $\hat{y}^i$  is the ground-truth label that can be obtained by the PS algorithm. With the  $K$ -fold average ensemble,  $K$  homogeneous models can be trained for each base model. As each homogeneous model can give a prediction,  $K$  pairs of predictions can be obtained. In this work, the structures of these homogeneous models are the same. We use the He normal initialization to initialize the parameters of these networks.<sup>10</sup> As both the training data and the initial values of the parameters are different, the performance of each network will be different, which enhances the diversity in model prediction. To combine these predictions, their average is computed as

$$\bar{y}^i = \frac{1}{K} \sum_{k=1}^K y_k^i, \quad (4)$$

where  $y_k^i$  is the prediction of the  $k$ th homogeneous model regarding to the  $i$ th base model and  $\bar{y}^i$  is the output of the  $i$ th base model using the  $K$ -fold average ensemble.

To further combine the predictions of the base models, we develop an adaptive ensemble that adopts a MultiResUNet to fuse the features of different models adaptively.<sup>11</sup> The diagram of the adaptive ensemble is shown in Fig. 3. The feature extraction is enhanced by MultiRes blocks that use a series of  $3 \times 3$  convolutions, as shown in Fig. 3(b). This structure is equivalent to the  $5 \times 5$  and  $7 \times 7$  convolutions and has the advantage that it

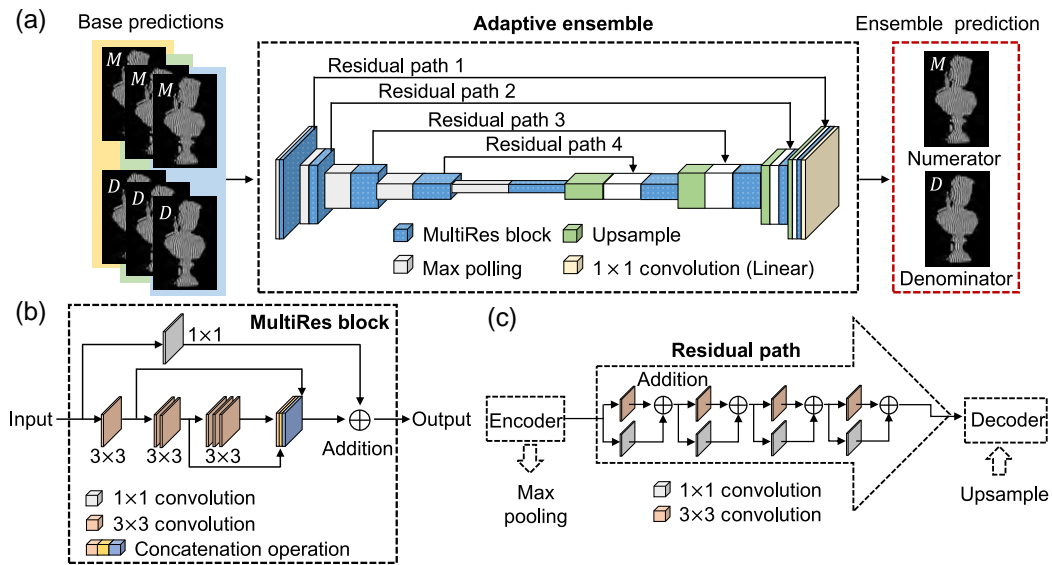
can not only learn features of various base predictions at different image scales but also saves memory and speeds up network training. In addition, instead of combining the features of encoders and decoders immediately, residual paths are constructed, where features of the encoder are processed by several convolutional layers, which can reduce the content gap between encoder and decoder features. To train the MultiResUNet, we also use the loss function shown in Eq. (3). During training, the MultiResUNet can learn proper weights for features extracted from each base prediction without manual intervention, thus making the fusion in an adaptive and automatic way.

### 3 Results

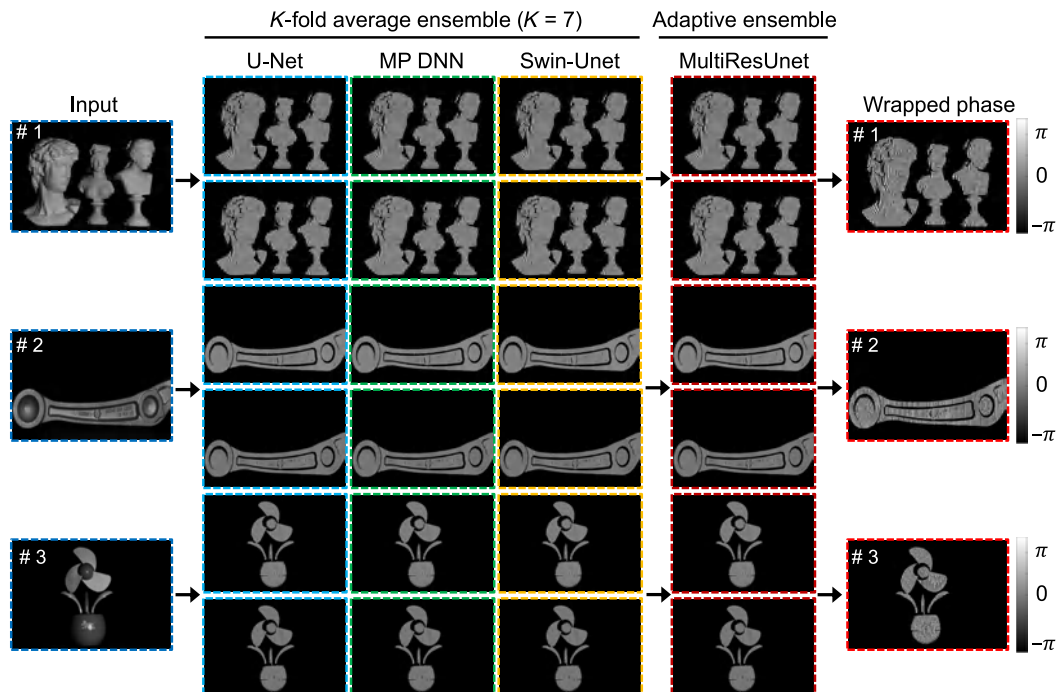
We validated the presented method under the scenario of fringe projection profilometry. The system consists of a camera (V611, Vision Research Phantom) and a projector (DLP 4100, Texas Instruments). The measured scene was illuminated by the projector with a sinusoidal fringe pattern, and the fringe image was captured by the camera from a different viewing point. To collect the training data, many fringe images of various objects were captured. To generate the ground-truth labels, the 12-step PS algorithm was applied. The captured fringe patterns are 8-bit gray-scale images. In the data preprocessing stage, the input fringe pattern was divided by 255 for normalization before being fed into the DNNs. Further details about the optical setup and the calculation of the ground-truth data are provided in the [Supplementary Material](#). For the adaptive ensemble, the training data were generated using the trained base models. All base models and the MultiResUNet were implemented by the Keras and computed on a graphic card (GTX Titan, NVIDIA).

To test the performance of our approach, we measured three different scenarios that were not seen by these networks during training. They are a set of statues, an industrial part made of aluminium alloy, and a desk fan made of plastic. The experimental results regarding each stage of our approach are shown in Fig. 4. Here, for better performance, a seven-fold average ensemble was used to train each base model. So, we divided the training data into seven parts and trained seven homogeneous models for each base model. Given an input fringe pattern, the homogeneous models gave predictions independently, and





**Fig. 3** Diagram of the proposed adaptive ensemble. (a) It trains a MultiResUNet to combine the predictions of base models. (b) Structure of the MultiRes block, where a series of  $3 \times 3$  convolutions is used to approximate the behaviors of  $5 \times 5$  convolution and  $7 \times 7$  convolution. (c) Structure of the residual path, where features of the encoder pass through a few convolutional layers before being fed into the decoder.



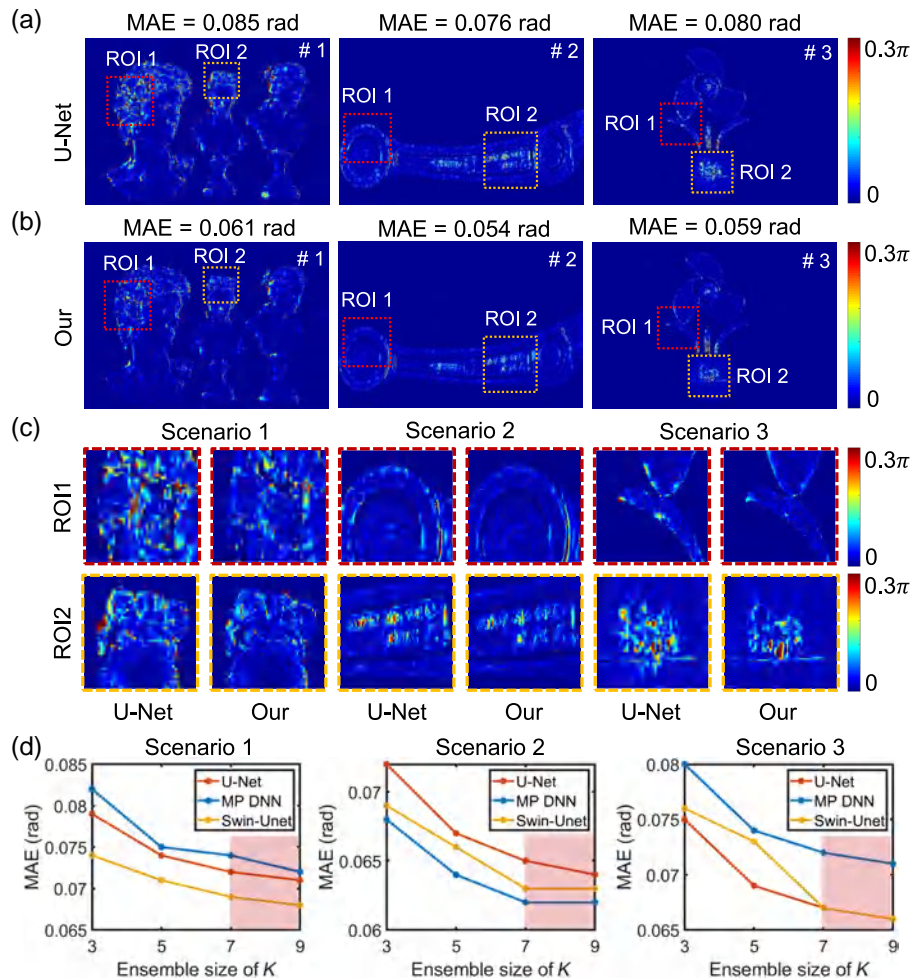
**Fig. 4** Experimental results of several unseen scenarios that include a set of statues, an industrial part, and a desk fan. The input is a fringe pattern. It is then fed into the U-Net, MP DNN, and Swin-U-Net, which are trained by the sevenfold average ensemble, respectively. By calculating the average, each base model outputs a pair of numerators and denominators. Then, the outputs of base models are processed by the adaptive ensemble, which combines the contribution of each base model and calculates the wrapped phase.

Eq. (4) was used to compute the average. As there were three base models, three pairs of numerators and denominators were obtained for each input image. These predictions were further combined by being fed into the adaptive ensemble that output the final prediction and calculated the wrapped phase.

For quantitative analysis, the ground-truth phase was obtained by the 12-step PS method. For comparison, the fringe image was also analyzed by a single U-Net; its absolute phase error is shown in Fig. 5(a). For the first scenario, we can see that the phase of smooth areas is retrieved accurately, while that of complex regions is measured with large errors. The mean absolute error (MAE) of the whole scene is 0.085 rad. The phase error of our approach is shown in Fig. 5(b). As can be seen, the phase error of the first scene has been reduced effectively. For detailed investigation, two regions of interest (ROIs), i.e., two complex regions around hairs, were selected. We can see that our method performs much better than the U-Net for handling the complex areas of depth variations and edges. Quantitatively, the MAE was greatly reduced to 0.061 rad when our method was used. For the second scenario, the MAE of the U-Net is 0.076 rad, and obvious errors can be observed around the edges and the small raised letters on the surface of the object,

as can be seen in Figs. 5(b) and 5(c). When our approach was applied, these phase errors were apparently reduced, and the MAE of the scene has been reduced to 0.054 rad. Last, for the third scenario, our method also outperformed the U-Net, as the MAE decreased significantly from 0.080 to 0.059 rad, demonstrating the accuracy improvement by 26%.

To further validate the proposed method, we investigated the effect of the ensemble size of the  $K$ -fold average ensemble. Different  $K$  were tested for these base models; the results are shown in Fig. 5(d). We find that a similar trend can be observed for these base models. The MAE decreases with the increase of  $K$ , and it tends to be stable when  $K$  is larger than seven. Therefore, the sevenfold average ensemble was used in our work. Moreover, we also compared the accuracy of each base model under the cases of the single model and the seven-fold average ensemble. Table 1 shows their MAEs for the tested scenarios. From the performance of a single DNN, we find different models demonstrate different performances. For example, the U-Net shows the smallest MAE for the third scenario, while the MAE for the second scenario is the largest among the three models. When the seven-fold average ensemble was utilized, the ensembles outperformed the single model as



**Fig. 5** Comparison of the proposed method with the U-Net. (a) and (b) The absolute phase error maps of the U-Net and our method, respectively. (c) Selected ROIs of the phase error for the two methods. (d) The performance of different  $K$ -fold average ensembles.

**Table 1** Quantitative validation of the proposed approach.

Method	MAE of #1 (rad)	MAE of #2 (rad)	MAE of #3 (rad)
U-Net (single)	0.085	0.076	0.080
MP DNN (single)	0.089	0.074	0.085
Swin-Unet (single)	0.081	0.075	0.081
U-Net (seven-fold)	0.072	0.065	0.067
MP DNN (seven-fold)	0.074	0.062	0.072
Swin-Unet (seven-fold)	0.069	0.063	0.067
Adaptive ensemble	0.061	0.054	0.059

the MAEs were reduced. After further combining the outputs of the base models by the adaptive ensemble, we obtained the smallest MAE of 0.061, 0.054, and 0.059 rad for these scenes, respectively. From this experiment, we can see that different DNNs have different advantages, and it is hard for a single DNN to demonstrate excellent performance for all scenarios. It is worth noting that the model accuracy and generalization capability can be improved significantly by the proposed approach, which combines the strengths of diverse models. More experimental results are provided in the [Supplementary Material](#).

## 4 Conclusions

In this work, we have proposed a novel fringe-pattern analysis method using ensemble deep learning, which can exploit the contributions of multiple state-of-the-art DNNs. The  $K$ -fold average ensemble approach is developed to manipulate the training data set into different groups. Each base model is trained several times with different groups of data. Within each base model, the output is computed by taking the average over the predictions of all homogeneous models. To further fuse the predictions of the base models, we have proposed an adaptive ensemble that can train a DNN to combine these predictions adaptively and automatically. Experimental results have shown that our work can leverage the strength of multiple base models to boost performance, which is superior to the method that only uses a single DNN. Furthermore, deep-learning techniques have been widely applied in various optical metrology applications, such as phase unwrapping, 3D reconstruction, and image denoising. However, a single model with a fixed architecture may only extract limited information from input data. We believe that the idea of utilizing the collective wisdom demonstrated here can also be extended to these applications because more DNNs of different structures can extract diverse information from input data, which is advantageous for making reliable predictions. We believe this work has great potential in inspiring powerful and practical optical metrology techniques in the future.

## Acknowledgments

This work was supported by the National Key R&D Program of China (Grant Nos. 2022YFB2804600 and 2022YFB2804605), the National Natural Science Foundation of China (Grant Nos. 62075096 and U21B2033), the Leading Technology of

Jiangsu Basic Research Plan (Grant No. BK20192003), the “333 Engineering” Research Project of Jiangsu Province (Grant No. BRA2016407), the Jiangsu Provincial “Belt and Road Initiative” Cooperation Project (Grant No. BZ2020007), the Fundamental Research Funds for the Central Universities (Grant No. 30921011208), and the National Major Scientific Instrument Development Project (Grant No. 62227818).

## References

1. M. Takeda and K. Mutoh, “Fourier transform profilometry for the automatic measurement of 3-D object shapes,” *Appl. Opt.* **22**(24), 3977–3982 (1983).
2. C. Zuo et al., “Phase shifting algorithms for fringe projection profilometry: a review,” *Opt. Lasers Eng.* **109**, 23–59 (2018).
3. G. Barbastathis, A. Ozcan, and G. Situ, “On the use of deep learning for computational imaging,” *Optica* **6**(8), 921–943 (2019).
4. C. Zuo et al., “Deep learning in optical metrology: a review,” *Light: Sci. Appl.* **11**(1), 39 (2022).
5. S. Feng et al., “Fringe pattern analysis using deep learning,” *Adv. Photonics* **1**(2), 025001 (2019).
6. M. A. Ganaie et al., “Ensemble deep learning: a review,” *CoRR*, <https://arxiv.org/abs/2104.02395> (2021).
7. M. S. S. Rahman et al., “Ensemble learning of diffractive optical networks,” *Light: Sci. Appl.* **10**(1), 14 (2021).
8. O. Ronneberger, P. Fischer, and T. Brox, “U-net: convolutional networks for biomedical image segmentation,” *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
9. H. Cao et al., “Swin-unet: Unet-like pure transformer for medical image segmentation,” <https://doi.org/10.48550/arXiv.2105.05537> (2021).
10. K. He et al., “Delving deep into rectifiers: surpassing human-level performance on imagenet classification,” in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 1026–1034 (2015).
11. N. Itezhaz and M. S. Rahman, “MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation,” *Neural Networks* **121**, 74–87 (2020).

**Shijie Feng** received his PhD in optical engineering at Nanjing University of Science and Technology. He is working as an associate professor at Nanjing University of Science and Technology. His research interests include phase measurement, high-speed 3D imaging, fringe projection, machine learning, and computer vision.

**Yile Xiao** is pursuing his MS degree at Nanjing University of Science and Technology. His research interests include phase measurement, high-speed 3D imaging, fringe projection, and deep learning.

**Wei Yin** received his PhD from Nanjing University of Science and Technology. His research interests include deep learning, high-speed 3D imaging, fringe projection, and computational imaging.

**Yan Hu** received his PhD from Nanjing University of Science and Technology. His research interests include high-speed microscopic imaging, 3D imaging, and system calibration.

**Yixuan Li** is a PhD student at Nanjing University of Science and Technology. Her research interests include phase measurement, high-speed 3D imaging, fringe projection, and deep learning.

**Chao Zuo** received his BE and PhD degrees from Nanjing University of Science and Technology (NJUST) in 2009 and 2014, respectively. He was working as a research assistant at the Centre for Optics and Lasers Engineering, Nanyang Technological University, from 2012 to 2013. Currently, he is working as a professor in the Department of Electronic and Optical Engineering and principal investigator of the Smart Computational Imaging Laboratory, NJUST. His research interests



include computational imaging and high-speed 3D sensing and has authored over 160 peer-reviewed journal publications. He has been selected for the Natural Science Foundation of China for Excellent Young Scholars and the Outstanding Youth Foundation of Jiangsu Province, China. He is the fellow of SPIE and Optica.

**Qian Chen** received his BS, MS, and PhD degrees from Nanjing University of Science and Technology. Currently, he is working as a

professor and vice-principal at Nanjing University of Science and Technology. He has been selected as Changjiang Scholar Distinguished Professor. With broad research interests in photoelectric imaging and information processing, he has authored more than 200 journal papers. His research team develops novel technologies and systems for non-interferometric quantitative phase imaging and high-speed 3D sensing and imaging with particular applications in national defense, industry, and bio-medicine.

OPEN

# Temporal phase unwrapping using deep learning

 Wei Yin<sup>1,2,3</sup>, Qian Chen<sup>1,2\*</sup>, Shijie Feng<sup>1,2,3</sup>, Tianyang Tao<sup>1,2,3</sup>, Lei Huang<sup>4</sup>, Maciej Trusiak<sup>5</sup>, Anand Asundi<sup>6</sup> & Chao Zuo<sup>1,2,3\*</sup>

The multi-frequency temporal phase unwrapping (MF-TPU) method, as a classical phase unwrapping algorithm for fringe projection techniques, has the ability to eliminate the phase ambiguities even while measuring spatially isolated scenes or the objects with discontinuous surfaces. For the simplest and most efficient case in MF-TPU, two groups of phase-shifting fringe patterns with different frequencies are used: the high-frequency one is applied for 3D reconstruction of the tested object and the unit-frequency one is used to assist phase unwrapping for the wrapped phase with high frequency. The final measurement precision or sensitivity is determined by the number of fringes used within the high-frequency pattern, under the precondition that its absolute phase can be successfully recovered without any fringe order errors. However, due to the non-negligible noises and other error sources in actual measurement, the frequency of the high-frequency fringes is generally restricted to about 16, resulting in limited measurement accuracy. On the other hand, using additional intermediate sets of fringe patterns can unwrap the phase with higher frequency, but at the expense of a prolonged pattern sequence. With recent developments and advancements of machine learning for computer vision and computational imaging, it can be demonstrated in this work that deep learning techniques can automatically realize TPU through supervised learning, as called deep learning-based temporal phase unwrapping (DL-TPU), which can substantially improve the unwrapping reliability compared with MF-TPU even under different types of error sources, e.g., intensity noise, low fringe modulation, projector nonlinearity, and motion artifacts. Furthermore, as far as we know, our method was demonstrated experimentally that the high-frequency phase with 64 periods can be directly and reliably unwrapped from one unit-frequency phase using DL-TPU. These results highlight that challenging issues in optical metrology can be potentially overcome through machine learning, opening new avenues to design powerful and extremely accurate high-speed 3D imaging systems ubiquitous in nowadays science, industry, and multimedia.

Many imaging systems, such as fringe projection profilometry (FPP)<sup>1–3</sup>, optical interferometry<sup>4,5</sup>, synthetic aperture radar (InSAR)<sup>6,7</sup>, X-ray crystallography<sup>8</sup>, and magnetic resonance imaging<sup>9</sup>, make use of the phase to produce the physiological and physical information of the measured objects. For instance, in FPP, the phase is proportional to the surface profile; in optical interferometry, the phase can be exploited to infer profile, fast displacement, and vibration of the object's surface. In these existing imaging methods and systems, it generally need to perform the arctangent function for phase retrieval thus resulting in the wrapped phase with  $2\pi$  phase jumps, so the operation of phase unwrapping is necessary to eliminate the phase ambiguities and convert the wrapped phases into the absolute ones<sup>10–15</sup>.

Numerous phase unwrapping algorithms have been proposed and can be divided into two categories with regard to the working domains: spatial phase unwrapping (SPU)<sup>10,11</sup> and temporal phase unwrapping (TPU)<sup>12</sup>. Under the assumption of spatial continuity, SPU calculates the relative fringe order of the center pixel on a single

<sup>1</sup>School of Electronic and Optical Engineering, Nanjing University of Science and Technology, No. 200 Xiaolingwei Street, Nanjing, Jiangsu Province, 210094, China. <sup>2</sup>Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing University of Science and Technology, Nanjing, Jiangsu Province, 210094, China. <sup>3</sup>Smart Computational Imaging (SCI) Laboratory, Nanjing University of Science and Technology, Nanjing, Jiangsu Province, 210094, China. <sup>4</sup>Brookhaven National Laboratory, NSLS II 50 Rutherford Drive, Upton, New York, 11973-5000, United States. <sup>5</sup>Institute of Micromechanics and Photonics, Warsaw University of Technology, 8 Sw. A. Boboli Street, Warsaw, 02-525, Poland. <sup>6</sup>Centre for Optical and Laser Engineering (COLE), School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore, 639798, Singapore. \*email: [chenqian@njust.edu.cn](mailto:chenqian@njust.edu.cn); [zuochao@njust.edu.cn](mailto:zuochao@njust.edu.cn)

wrapped phase map by analyzing the phase information of its neighboring pixels, thus it cannot successfully measure discontinuities and isolated objects. Conversely, TPU approaches can realize pixel-wise absolute phase unwrapping via the temporal analysis of more than one wrapped phase maps with different frequencies even under the conditions of truncated or spatially isolated areas. Currently, there are three representative approaches to TPU: multi-frequency (hierarchical) approach (MF-TPU), multi-wavelength (heterodyne) approach, and number-theoretical approach. We have analyzed and discussed the unwrapping success rate and anti-noise performance of these TPU algorithms in a comparative review, revealing that the MF-TPU approach provides the highest unwrapping reliability and best noise-robustness among others<sup>12</sup>.

The subsequent content of this paper will be focused on the MF-TPU approach, with an emphasis on the application of high-speed FPP<sup>16,17</sup>. In such a context, to improve the measurement efficiency, it is necessary to make MF-TPU as reliable as possible while using a minimum number of projection patterns<sup>18</sup>. For the simplest and most efficient case in MF-TPU, two groups of phase-shifting fringe patterns with different frequencies are used: the high-frequency one is applied for 3D reconstruction of the tested object and the unit-frequency one is used to assist phase unwrapping for the wrapped phase with high frequency. The final measurement precision or sensitivity is determined by the number of fringes used within the high-frequency pattern, under the pre-condition that its absolute phase can be successfully recovered without any fringe order errors. However, due to the non-negligible noises and other error sources in actual measurement, the frequency of the high-frequency fringes is generally restricted to about 16, resulting in limited measurement accuracy<sup>12</sup>. On the other hand, using an additional intermediate set of fringe patterns (totally 3 sets of phase-shifting patterns) can unwrap the phase with higher frequency or higher success rate<sup>18</sup>. As a result, the increased number of required patterns reduces the measurement efficiency of FPP, which is not suitable for measuring dynamic scenes.

In this work, we demonstrated that a trained deep neural network can greatly improve the ability of TPU compared with conventional MF-TPU. This learning-based framework uses only two (one unit-frequency, one high-frequency) wrapped phases calculated using 3-step phase-shifting fringe patterns as input, and directly outputs an unwrapped version of the same phase map with high reliability. Deep learning<sup>19</sup> is a method based on the representation of data in machine learning for data analysis and prediction and have been applied to various fields such as automatic drive, face recognition, and mechanical translation, where they have produced results that surpass the performance of traditional algorithms and are comparable or superior in some cases to human experts. Recently, machine learning-based methods have been further successfully applied to solving challenging problems in computational imaging<sup>20–24</sup> and the analysis of nanostructures devices<sup>25–27</sup>, such as phase retrieval<sup>20</sup>, lensless on-chip microscopy<sup>21</sup>, fringe pattern analysis<sup>22</sup>, computational ghost imaging<sup>23,24</sup>, and the assist design of electromagnetic nanostructures<sup>26</sup>.

Inspired by the great successes of deep learning techniques for these fields, here we adopt deep neural networks to beat the TPU problem, which can substantially improve the unwrapping reliability compared with MF-TPU even in the presence of different types of error sources. To validate the proposed approach, we recover the absolute phases of various tested objects by projecting fringe patterns with different frequencies, such as 1, 8, 16, 32, 48, and 64, all of which demonstrate the successful removal of phase unwrapping errors arising from the intensity noise, low fringe modulation, intensity nonlinearity, and motion artifacts. Furthermore, as far as we know, our method was demonstrated experimentally that the high-frequency phase with 64 periods can be directly and reliably unwrapped from one unit-frequency phase, facilitating high-accuracy high-speed 3D surface imaging with use of only 6 projected patterns without exploring any prior information and geometric constraint. These results highlight that machine learning is able to potentially overcome challenging issues in optical metrology, and provides new possibilities to design powerful high-speed FPP systems.

## Methods

**Phase-shifting profilometry (PSP).** In a typical FPP system, sinusoidal fringe-based FPP methods are more prevalent to a great variety of practical applications and can be generally divided into two main categories for phase extraction: Fourier transform profilometry (FTP)<sup>28</sup> and Phase-shifting profilometry (PSP)<sup>29</sup>. Numerous dynamic 3D measurement techniques have been developed based on FTP, which have the advantage to provide the phase map utilizing only a single high-frequency fringe pattern<sup>16,30</sup>. However, suffering from frequency band overlapping problem, these methods generally yield coarse wrapped phase with low quality which limits its measurement precision for dynamic 3D acquisition. In addition, not just limited to Fourier transform, the windowed Fourier transform (WFT) and the wavelet transform (WT) can also be applied for the phase retrieval and enhancing 3D measurement accuracy even in the case of complex surfaces and depth discontinuities<sup>31</sup>. Different from FTP, PSP can realize pixel-by-pixel phase measurements with higher accuracy unaffected by ambient light, but it needs to project at least three fringe patterns to obtain a phase map theoretically<sup>29</sup>. In this work, the standard 3-step phase-shifting fringe patterns with shift offset of  $2\pi/3$  are adopted and represented as

$$I_n^p(x^p, y^p) = 0.5 + 0.5 \cos(2\pi f x^p - 2\pi n/3), \quad (1)$$

where  $I_n^p(x^p, y^p)$  ( $n = 0, 1, 2$ ) represent fringe patterns to be projected,  $f$  is the frequency of fringe patterns. After projected onto the object surfaces, the deformed fringe patterns captured by the camera can be described as

$$I_n^c(x, y) = A(x, y) + B(x, y) \cos(\Phi(x, y) - 2\pi n/3), \quad (2)$$

where  $A(x, y)$ ,  $B(x, y)$ , and  $\Phi(x, y)$  are the average intensity, the intensity modulation, and the phase distribution of the measured object. According to the least-squares algorithm, the wrapped phase  $\phi(x, y)$  can be obtained as<sup>32–34</sup>:



$$\phi(x, y) = \tan^{-1} \frac{\sqrt{3}(I_1^c(x, y) - I_2^c(x, y))}{2I_0^c(x, y) - I_1^c(x, y) - I_2^c(x, y)}. \quad (3)$$

Due to the truncation effect of the arctangent function, the obtained phase  $\phi(x, y)$  is wrapped within the range of  $(-\pi, \pi]$ , and its relationship with  $\Phi(x, y)$  is:

$$\Phi(x, y) = \phi(x, y) + 2\pi k(x, y), \quad (4)$$

where  $k(x, y)$  represents the fringe order of  $\Phi(x, y)$ , and its value range is from 0 to  $N - 1$ .  $N$  is the period number of the fringe patterns (i.e.,  $N = f$ ). In FPP, the core challenge for the absolute phase recovery is to obtain  $k(x, y)$  for each pixel in the phase map quickly and accurately.

**Multi-frequency temporal phase unwrapping (MF-TPU).** In temporal phase unwrapping (TPU), the wrapped phase  $\phi(x, y)$  is unwrapped with the aid of one (or more) additional wrapped phase map with different frequency. For instance, two wrapped phases  $\phi_h(x, y)$  and  $\phi_l(x, y)$  are both retrieved from phase-shifting algorithms by using Eq. (3), ranging from  $-\pi$  to  $\pi$ . It is easy to find that the two absolute phases  $\Phi_h(x, y)$  and  $\Phi_l(x, y)$  corresponding to  $\phi_h(x, y)$  and  $\phi_l(x, y)$  have the following relationship:

$$\begin{cases} \Phi_h(x, y) = \phi_h(x, y) + 2\pi k_h(x, y), \\ \Phi_l(x, y) = \phi_l(x, y) + 2\pi k_l(x, y), \\ \Phi_h(x, y) = (f_h/f_l)\Phi_l(x, y), \end{cases} \quad (5)$$

where  $f_h$  and  $f_l$  are the frequency of high-frequency fringes and low-frequency fringes. Based on the principle of MF-TPU,  $k_h(x, y)$  can be calculated by the following formula:

$$k_h(x, y) = \frac{(f_h/f_l)\Phi_l(x, y) - \phi_h(x, y)}{2\pi}. \quad (6)$$

Since the fringe order  $k_h(x, y)$  is integer, ranging from 0 to  $f_h - 1$ , Eq. (6) can be adapted as

$$k_h(x, y) = \text{Round} \left[ \frac{(f_h/f_l)\Phi_l(x, y) - \phi_h(x, y)}{2\pi} \right], \quad (7)$$

where  $\text{Round}()$  is the rounding operation. When  $f_l$  is 1, there will be no phase ambiguity so that  $\Phi_l(x, y)$  is inherently an unwrapped phase. Theoretically, for MF-TPU, this single-period phase can be to directly assist phase unwrapping of  $\phi_h(x, y)$  with relatively higher frequency. However, the phase unwrapping capability of MF-TPU is greatly constrained due to the influence of noise in practice. Assuming phase errors in the wrapped phase maps  $\phi_h(x, y)$  and  $\Phi_l(x, y)$  are  $\Delta\phi_h(x, y)$  and  $\Delta\phi_l(x, y)$  respectively, from Eq. (6) we have:

$$\Delta k(x, y) = \frac{(f_h/f_l)\Delta\phi_l(x, y) - \Delta\phi_h(x, y)}{2\pi}, \quad (8)$$

Let  $\Delta\phi_{\max} = \max(|\Delta\phi_h(x, y)|, |\Delta\phi_l(x, y)|)$ , from Eq. (8) we can find the upper bound of  $\Delta k(x, y)$ :

$$\Delta k_{\max}(x, y) = \left| \frac{(f_h/f_l)\Delta\phi_l(x, y) - \Delta\phi_h(x, y)}{2\pi} \right| = \Delta\phi_{\max} \frac{f_h + f_l}{2\pi f_l}. \quad (9)$$

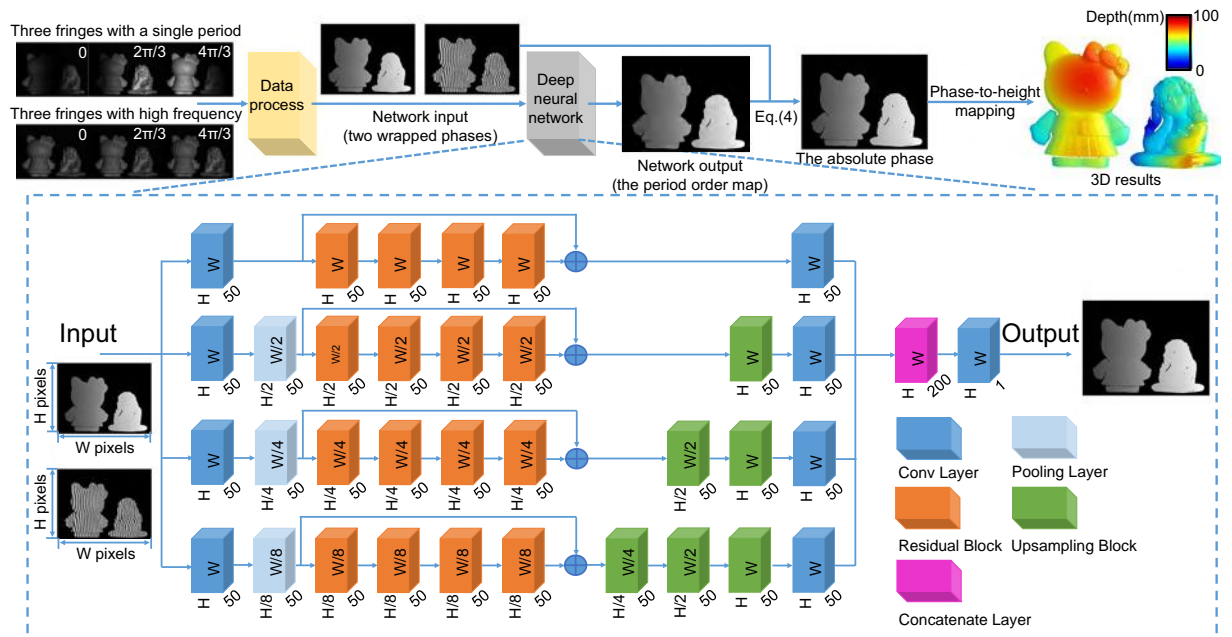
To avoid errors in determining the fringe orders, from Eqs. (7) and (9) we have:

$$\Delta k_{\max}(x, y) = \Delta\phi_{\max} \frac{f_h + f_l}{2\pi f_l} < 0.5. \quad (10)$$

Subsequently, we can confirm the boundary of  $\Delta\phi_{\max}(x, y)$ :

$$0 \leq \Delta\phi_{\max}(x, y) < \frac{\pi f_l}{f_h + f_l}. \quad (11)$$

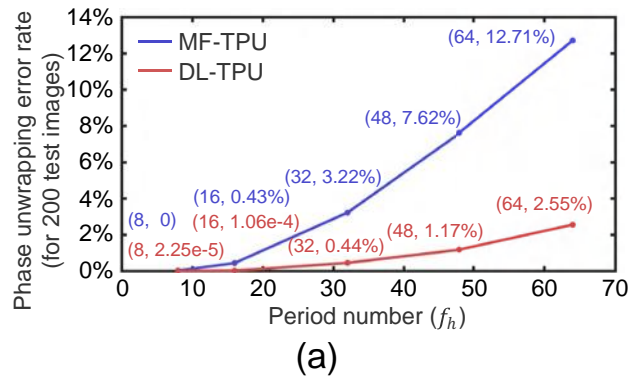
Notably, Eq. (11) defines the range of  $\Delta\phi_{\max}$  where the absolute phase can be correctly recovered. Otherwise, error will occur in determining the exact  $k_h(x, y)$ . In MF-TPU, since the frequency of the low-frequency fringes is fixed to 1, it can be found from Eq. (11) that the higher the frequency of the high-frequency fringes, the narrower the range of  $\Delta\phi_{\max}$ , and the worse the reliability of the phase unwrapping. Consequently, for a normal FPP system, MF-TPU can only reliably unwrap the phase with about 16 periods due to the non-negligible noises and other error sources in actual measurement. Thus, it generally exploits multiple ( $>2$ ) sets of phases with different frequencies to hierarchically unwrap the wrapped phase step by step, and finally arrives at the absolute phase with high frequency instead of only using the phase with a single period. Obviously, MF-TPU, which consumes additional time for projecting patterns with intermediate frequencies, is not a good choice to realize high-speed, high-precision 3D shape measurement based on FPP.



**Figure 1.** The diagram of the proposed method. The whole framework is composed of data process, deep neural network, and phase-to-height mapping. Data process is performed to extract phases and remove the background from fringe images according to Eq. (3) and Supplementary Eq. S1. Deep neural network, consisting of convolutional layers, pooling layers, residual blocks, upsampling blocks, and concatenate layer, is used to predict the period order map  $k_h(x, y)$  from the input data ( $\Phi_l(x, y)$  and  $\phi_h(x, y)$ ). Then, using Eq. (4),  $\Phi_h(x, y)$  is obtained and converted into 3D results after phase-to-height mapping.

**Deep-learning based temporal phase unwrapping (DL-TPU).** Aiming at this problem, we choose to use the deep neural networks (DNN) to overcome the limitations of MF-TPU, and the specific diagram of the proposed method is shown as in Fig. 1. The input data of the network are the two wrapped phases of the single period and high frequency, which is the same as the two-frequency TPU. To realize the highest unwrapping reliability, we adopt the residual network as the basic skeleton of our neural network<sup>35</sup>, which can speed up the convergence of deep networks and improve network performance by adding layers with considerable depth. Then, we introduce the multi-scale pooling layer to down-sampling the input tensors, which can compress and extract the main features of the tensors for reducing the computation complexity and preventing the over-fitting. Correspondingly, it is inconsistent for the tensors sizes in the different paths after the processing of the pooling layer. Therefore, upsampling blocks will be used to make the sizes of the tensors in the respective paths uniform (see Supplementary Section 1 for details)<sup>36</sup>. In summary, our network mainly consists of convolution layers, residual blocks, pooling layers, upsampling blocks, and concatenate layers. To maximize the efficiency of the model, after repeatedly adjusted the hyper-parameters of the network (number of layers and nodes), we found that in the whole network the number of residual blocks for each path should be set to 4, and the basic filter numbers of the convolution layers should be 50. The tensor data of each path in the network will be performed 1, 1/2, 1/4, and 1/8 downsampling operations by adopting pooling layers with different scales respectively, and then different numbers of upsampling blocks will be adopted to make the sizes of the tensors in the corresponding paths uniform. Besides, it has been found that implementing shortcuts between residual blocks contributes to making the convergence of the network more stable. Furthermore, to avoid over-fitting as the common problem of the deep neural network, L2 regularization is adopted in each convolution layer of residual blocks and upsampling blocks instead of all convolution layers of the proposed network, which can enhance the generalization ability of the network.

Although the purpose of building the network is to achieve phase unwrapping and obtain the absolute phase, there is no need to directly set the absolute phase as the network's label. Since  $\Phi_h(x, y)$  is simply the linear combination of  $k_h(x, y)$  and  $\phi_h(x, y)$  according to Eq. (4),  $\Phi_h(x, y)$  can be obtained immediately if  $k_h(x, y)$  is known. Once  $k_h(x, y)$  is set as the output data of the network, the purpose of our network is to implement semantic segmentation<sup>37</sup>, which is a pixel-wise classification. It is easy to understand that the complexity of the network will be greatly reduced so that the loss of the network will converge faster and more stable, and the prediction accuracy of the network is effectively improved. Different from the traditional SPU and TPU that the phase unwrapping is performed by utilizing the phase information solely in the spatial or temporal domain, it should be noted that our proposed method based on deep neural network is able to learn feature extraction and data screening, thus can exploit the phase information in the spatial and temporal domain simultaneously, providing more degrees of freedom and possibilities to achieve significantly better unwrapping performance (refer to Supplementary Section 3 for details).

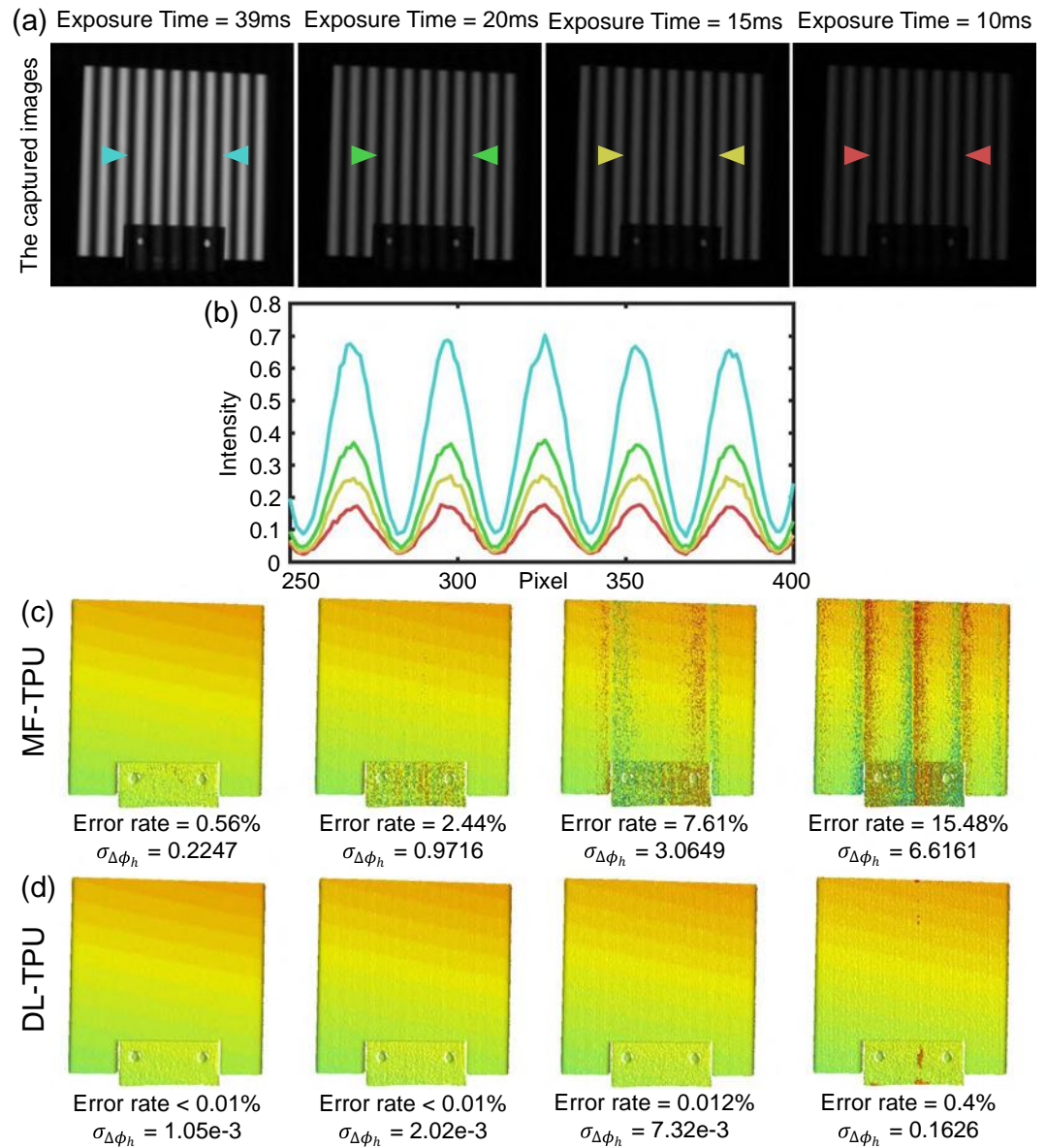


error rate	MF-TPU	DL-TPU
0% — 20%		
$f_h = 8$		
$f_h = 16$		
$f_h = 32$		
$f_h = 48$		
$f_h = 64$		

**Figure 2.** (a) Comparison of the average error rates of phase unwrapping with different high frequencies (such as 8, 16, 32, 48 and 64) on the testing dataset using MF-TPU and DL-TPU. (b) Comparison of the 3D reconstruction results after phase unwrapping with different high frequencies (such as 8, 16, 32, 48 and 64) for a representative sample on the testing dataset using MF-TPU and DL-TPU.

Then, using Eq. (4),  $\Phi_h(x, y)$  is obtained and converted into 3D results after phase-to-height mapping. In preparation for phase-to-height mapping, the projection matrices of the camera and projector need to be obtained through system calibration<sup>38,39</sup>. Besides, in order to speed up the reconstruction, we suggest phase-to-height mapping to be implemented with a graphics processing unit<sup>40</sup> or several look-up tables<sup>41</sup>, which can greatly save the time cost of the 3D reconstruction.

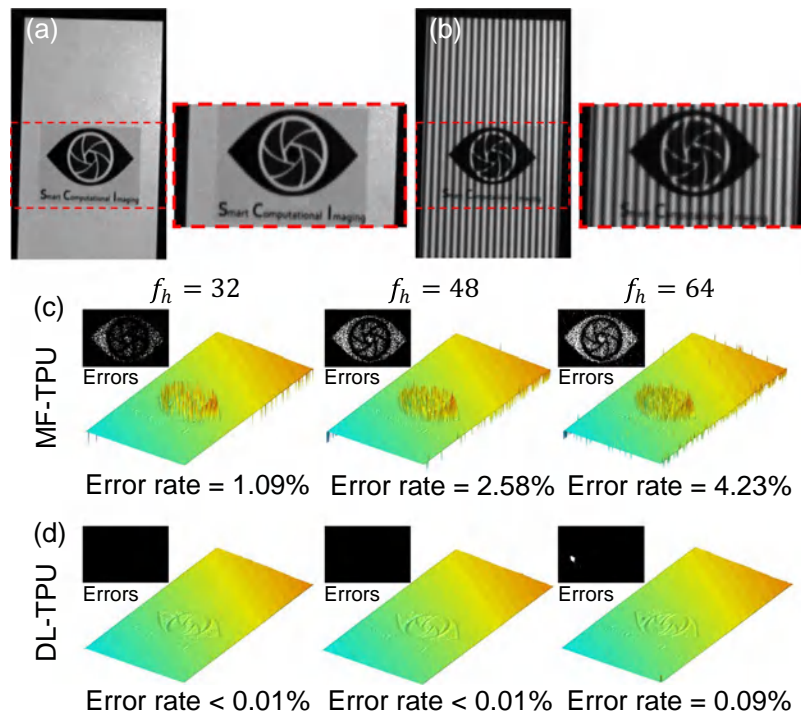




**Figure 3.** (a) The captured images ( $f_h = 32$ ) of a standard ceramic plate under different exposure times. (b) Comparison of intensity in line 230 of the captured images. (c,d) Comparison of the 3D reconstruction results after phase unwrapping under different exposure times using MF-TPU and DL-TPU.

## Results

**Quantitative comparison with MF-TPU.** In the first experiment, to verify the actual performance of the proposed DL-TPU, the trained DNN models for phase unwrapping with different high-frequency fringes are utilized to make predictions on the testing dataset (200 image pairs) (refer to Supplementary Section 2 for details), and MF-TPU is also implemented for comparison. In order to quantitatively analyze the accuracy of phase unwrapping for DL-TPU and MF-TPU, the phases with different high frequencies are independently unwrapped by the two algorithms, and the average error rates for phase unwrapping on the testing dataset are calculated and plotted against  $f_h$  in Fig. 2(a). It should be noted that these results are calculated only by comparing the differences between the obtained phases and the label's phases for each valid point from the testing dataset (refer to Supplementary Section 2 for identifying the valid points). The label's phases can be correctly acquired as the 'ground-truth' phase by exploiting multiple sets of phases with different frequencies to hierarchically unwrap the wrapped phase step by step. It can be seen from Fig. 2(a) that with the increase of  $f_h$  the reconstructed phases of MF-TPU are completely obviated, with a substantial increase of phase unwrapping error rate from 0 to 12.71%. The result shows again that MF-TPU cannot successfully unwrap a phase map when  $f_h \geq 16$  due to the non-negligible noises and other error sources in actual measurement. However, our approach always provides acceptable results, with more than 95% of all valid pixels being properly unwrapped. These experimental results confirm



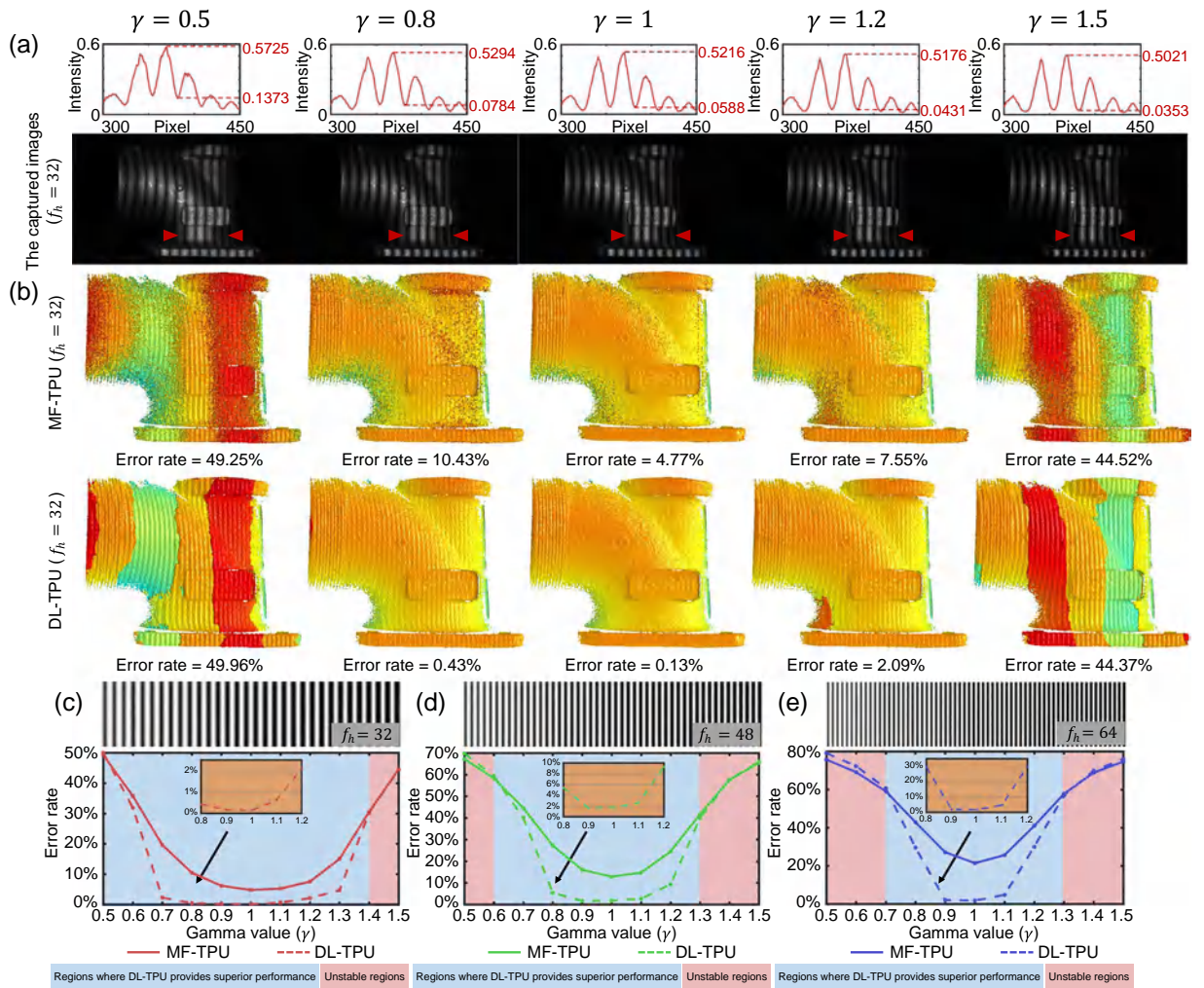
**Figure 4.** (a) The tested object with the low-modulation logo. (b) The captured fringe image ( $f_h = 64$ ). (c,d) Comparison of the 3D reconstruction results after phase unwrapping for the low-quality region using MF-TPU and DL-TPU.

that compared with MF-TPU our method can achieve much better unwrapping results and decrease the phase unwrapping errors by almost an order of magnitude.

In order to reflect the specific performance of DL-TPU and MF-TPU more intuitively, the 3D reconstruction results after phase unwrapping for a representative sample on the testing dataset are illustrated and compared in Fig. 2(b), and the phase unwrapping error rates can be obviously seen in the background. It can be found from Fig. 2(b) that our approach provides the smallest phase unwrapping errors and the significant improvement of phase measurement quality with the period number  $f_h$  as expected. It can be further observed that the fringe order errors are mostly concentrated on the dark regions and object edges where the fringe quality is low. Different from MF-TPU, phase unwrapping errors caused by the low signal-to-noise ratio (SNR) region of phases is significantly reduced by using DL-TPU. For these low SNR region, the remaining phase errors have the characteristics of accumulation and can be easily further corrected by some compensation algorithm for fringe order errors<sup>42–44</sup> (refer to Supplementary Section 4 for details of these compensation algorithms). Consequently, the trained models can substantially decrease error points to provide better phase unwrapping results (even  $f_h = 64$ ) and lower error rates, which demonstrates the capability and reliability of DL-TPU for phase unwrapping.

**Performance analysis under different types of phase errors.** *Intensity noise.* In the following series of experiments, we will further verify the superiority of DL-TPU in the presence of different types of phase errors. In high-speed 3D measurement, the quality of the fringe patterns is poorer than that of the static measurement because it is projected and captured with limited exposure time. To emulate the practical measurement conditions, we measure a standard ceramic plate using DL-TPU ( $f_h = 32$ ) but artificially adjust the camera's exposure time to 39 ms, 20 ms, 15 ms, and 10 ms. To better analyze and compare the reliability of the accuracy results for phase unwrapping, the absolute phase map obtained using the 12-step phase-shifting algorithm and combining with a highly redundant multi-frequency temporal phase unwrapping strategy (with different frequencies including 1, 8, 16, and 32) can serve as the reference phase. Next, the error rate of phase unwrapping and the variance of the phase error  $\sigma_{\Delta\phi_h}$  for different approaches are easily calculated by making a comparison between the unwrapped phase and the reference phase for each valid point.

Obviously, as the exposure time decreases, the quality of the phase measurement drops significantly presented in Fig. 3(a,b). Since the exposure time is a key factor affecting the speed and quality of phase measurement, the shorter the exposure time the algorithm can withstand, the faster the measurement can be achieved with six projection patterns in FFP. Therefore, a more robust phase unwrapping method is essential to eliminate the phase ambiguity introduced by reduced exposure times and make phase unwrapping correct. In Fig. 3(c,d), it can be found that DL-TPU can always provide higher success rate of phase unwrapping and lower phase error  $\sigma_{\Delta\phi_h}$  compared with MF-TPU, making it more appropriate for the high-speed 3D shape measurement applications.



**Figure 5.** (a) The captured fringe images ( $f_h = 32$ ) and the comparison of intensity in line 363 of the corresponding images under different degrees of intensity gamma distortion. (b) The 3D reconstruction results after phase unwrapping under different degrees of intensity gamma distortion using MF-TPU and DL-TPU when  $f_h$  is 32. (c–e) The error rates of phase unwrapping with different high frequencies (such as 32, 48 and 64) under different degrees of intensity gamma distortion using MF-TPU and DL-TPU.

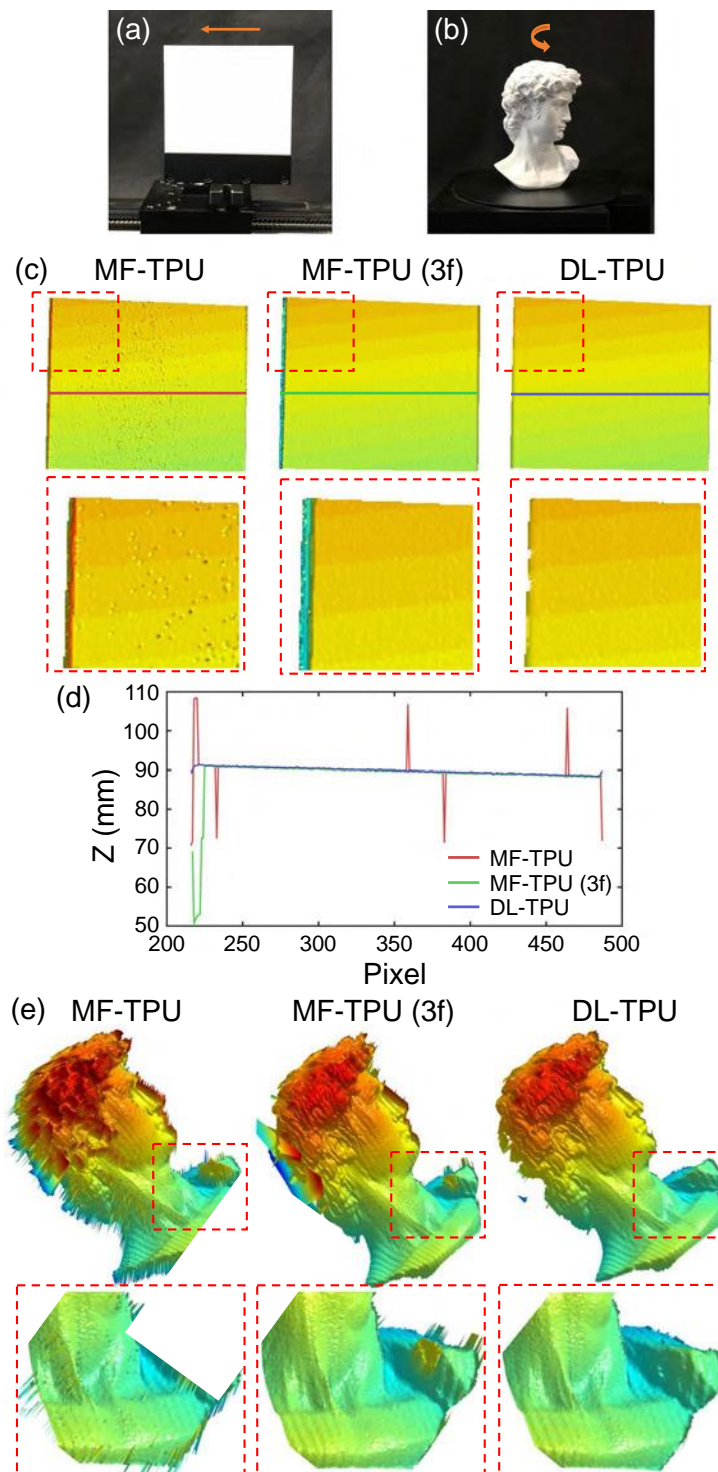
**Low fringe modulation.** Another attractive attribute of DL-TPU is its good tolerance to noise that can significantly suppress phase unwrapping errors in low-fringe-modulation areas, which frequently appear in practical measurement for the surfaces of complex objects, like the tested object shown in Fig. 4(a,b). For the low-modulation logo region, conventional MF-TPU results provide spinous results teemed with significant delta-spike artifacts, as shown in Fig. 4(c). In contrast, the DNN approach successfully overcomes the low-SNR problem and produces smooth measurement results with negligible errors, as shown in Fig. 4(d). This experimental result confirms once again that DL-TPU can provide superior capability and stability of phase unwrapping for suppressing unwrapping errors caused by low fringe modulation.

**Intensity nonlinearity.** In this section, we test the proposed DL-TPU under different degrees of intensity gamma distortion. The gamma distortion, or so called intensity nonlinearity, is a common error source in FPP due to the nonlinear response of the commercial projector, introducing high-order harmonics to the projected fringe patterns. The intensity of the fringes with the gamma distortion can be expressed as

$$I_n^{p,\gamma}(x^p, y^p) = \{0.5 + 0.5 \cos(2\pi f_x^p - 2\pi n/3)\}^\gamma, \tag{12}$$

where  $\gamma$  represents the nonlinearity parameter of projector that means the nonlinear response of the commercial projector. Then, we choose an industrial workpiece of metal as the measured object to validate the resistance of DL-TPU to the gamma distortion. A set of fringe patterns with different nonlinearity intensities, ranging from 0.5 to 1.5, are generated using Eq. (12) and projected onto the measured object in Fig. 5(a). It can be found from the 3D results shown in Fig. 5(b) that MF-TPU cannot provide acceptable phase unwrapping results even under low-level gamma distortions. On the contrary, DL-TPU is able to achieve a close to ideal phase unwrapping result even when  $\gamma$  is 0.8. It should be also noticed that, when  $\gamma$  is as low as 0.5 or as high as 1.5, both of the two





**Figure 6.** (a,b) The objects with fast translation movement and rapid rotary motion. (c) Comparison of the 3D results of phase unwrapping for the fast translation movement using MF-TPU, MF-TPU (3f), and DL-TPU. (d) The 3D result comparison in line 250 for the fast translation movement. (e) Comparison of the 3D results of phase unwrapping for the rapid rotary motion using MF-TPU, MF-TPU (3f), and DL-TPU.

approaches can produce meaningful results since the phase errors artificially introduced is much larger than the “safe line” without triggering phase unwrapping errors, so that the success/error rate of unwrapping is about fifty-fifty. In Fig. 5(c–e), for phase unwrapping with different high frequencies (such as 32, 48 and 64) under different degrees of intensity gamma distortion, the statistics curves of phase unwrapping for MF-TPU are shown as the solid lines, and the results are significantly improved by using DL-TPU as shown by dashed lines. These

results verify that our method can significantly reduce the fringe order errors of phase unwrapping and produce high-quality absolute phases even under a certain degree of gamma distortion in the FPP system.

**Application to high-speed 3D surface imaging.** Finally, our system, which can project and capture the fringe images at the speed of 25 Hz, is applied to imaging some classical dynamic scenes for fast 3D reconstruction: objects with fast translational movement and rapid rotatory motion. In Fig. 6(a), a standard ceramic plate, fixed on precise displacement platform, is performed to periodic translational movement at the speed of 1.25 cm/s. In traditional MF-TPU, it is more much difficult to recovery the high-frequency absolute phase using only one unit-frequency phase in Fig. 6(c) due to the unavoidable noises in actual measurement. Therefore, to guarantee a stable phase unwrapping success rate for the high-frequency phase, three sets of phase-shifting fringe patterns, so-called MF-TPU (3f) in which the frequency of the second set of fringe patterns is 8, are used to achieve high-accuracy but inefficient phase unwrapping. When measuring dynamic scenes, the relative motion between the object and the phase-shifting fringe patterns sequentially projected will cause motion artifacts and thus introduce additional phase errors into the initial phase map which is non-negligible and becomes more severe because of projecting more patterns as presented in Fig. 6(c). However, without the assistance of additional patterns, it illustrates the reliability and efficiency of DL-TPU from Fig. 6(c) that the trained models can still achieve better phase unwrapping results. We try to take one cross-section on the 3D results of the ceramic plate to compare DL-TPU with MF-TPU and MF-TPU (3f). From the comparison results shown in Fig. 6(d), it can be found that our approach provides the highest unwrapping reliability and best noise-robustness compared with other methods.

And then, for measuring the rapid rotatory motion, the statue of David rotates in a counter-clockwise direction at the rotation rate of 3 rpm as shown in Fig. 6(b). Undoubtedly, in Fig. 6(e), the experiment yielded a result similar to that of the fast translational motion. It can be found from these results that the 3D profile information with high quality of the ceramic plate and the David statue are accurately acquired during the entire movement of the tested objects, again demonstrating the unwrapping stability of the proposed method for implementing high-precision, fast absolute 3D shape measurement.

## Discussion

In this work, we have demonstrated that a trained deep neural network can greatly improve the ability of TPU with high-frequency fringes acquired by a common FPP system. This high-performance TPU (so-called DL-TPU) can be achieved based on a deep neural network after appropriate training. Compared with MF-TPU, DL-TPU can effectively recover the absolute phase from two wrapped phases with different frequencies by exploiting both spatial and temporal phase information in an integrated way. It can substantially improve the reliability of phase unwrapping even when high-frequency fringe patterns are used. We have further experimentally demonstrated for the first time, to our knowledge, that the high-frequency phase obtained from 64-period 3-step phase-shifting fringe patterns can be directly and reliably unwrapped from one unit-frequency phase, facilitating high-accuracy high-speed 3D surface imaging with use of only 6 projected patterns without exploring any prior information and geometric constraint. After that, various experiments have been designed to access the phase unwrapping capability of the proposed approach under the conditions of intensity noise, low fringe modulation, and intensity nonlinearity. Experimental results have verified that TPU using deep learning provides significantly improved unwrapping reliability to realize the absolute 3D measurement for objects with complex surfaces. Besides, for the applications to high-speed FPP, it has also been observed that the deep learning-based approach is much less affected by motion artifacts in dynamic measurement and can successfully reconstruct the surface profile of the moving and rotating objects at high speed. These results highlight that machine learning is able to potentially overcome challenging issues in optical metrology, and provides new possibilities and flexibilities to design more powerful high-speed FPP systems. Although the TPU and FPP have been the main focus of this research, we envisage that the similar deep learning framework might also be applicable to other 3D surface imaging modalities, including, e.g., stereo vision<sup>45</sup>, DIC<sup>46</sup>, spatial-temporal stereo<sup>47</sup>, spatial-temporal correlation<sup>48</sup>, among others.

Received: 19 May 2019; Accepted: 9 December 2019;

Published online: 27 December 2019

## References

- Gorthi, S. S. & Rastogi, P. Fringe projection techniques: whither we are? *Opt. Lasers Eng.* **48**, 133–140 (2010).
- Geng, J. Structured-light 3d surface imaging: a tutorial. *Adv. Opt. Photonics* **3**, 128–160 (2011).
- Feng, S. *et al.* High dynamic range 3d measurements with fringe projection profilometry: a review. *Meas. Sci. Technol.* **29**, 122001 (2018).
- Vest, C. M. Holographic interferometry. *New York, John Wiley Sons, Inc.* 476 (1979).
- Gahagan, K. *et al.* Measurement of shock wave rise times in metal thin films. *Phys. review letters* **85**, 3205 (2000).
- Bamler, R. & Hartl, P. Synthetic aperture radar interferometry. *Inverse problems* **14**, R1 (1998).
- Curlander, J. C. & McDonough, R. N. *Synthetic aperture radar*, vol. 396 (1991).
- Momose, A. Demonstration of phase-contrast x-ray computed tomography using an x-ray interferometer. *Nucl. Instruments Methods Phys. Res. Sect. A: Accel. Spectrometers, Detect. Assoc. Equip.* **352**, 622–628 (1995).
- Haacke, E. M. *et al.* *Magnetic resonance imaging: physical principles and sequence design*, vol. 82 (1999).
- Su, X. & Chen, W. Reliability-guided phase unwrapping algorithm: a review. *Opt. Lasers Eng.* **42**, 245–261 (2004).
- Flynn, T. J. Two-dimensional phase unwrapping with minimum weighted discontinuity. *JOSA A* **14**, 2692–2701 (1997).
- Zuo, C., Huang, L., Zhang, M., Chen, Q. & Asundi, A. Temporal phase unwrapping algorithms for fringe projection profilometry: A comparative review. *Opt. Lasers Eng.* **85**, 84–103 (2016).
- Schofield, M. A. & Zhu, Y. Fast phase unwrapping algorithm for interferometric applications. *Opt. Lett.* **28**, 1194–1196 (2003).
- Pritt, M. D. Phase unwrapping by means of multigrad techniques for interferometric sar. *IEEE Transactions on Geosci. Remote. Sens.* **34**, 728–738 (1996).

15. Chavez, S., Xiang, Q.-S. & An, L. Understanding phase maps in mri: a new outline phase unwrapping method. *IEEE transactions on medical imaging* **21**, 966–977 (2002).
16. Su, X. & Zhang, Q. Dynamic 3-d shape measurement method: a review. *Opt. Lasers Eng.* **48**, 191–204 (2010).
17. Zhang, S. High-speed 3d shape measurement with structured light methods: A review. *Opt. Lasers Eng.* **106**, 119–131 (2018).
18. Zhang, M. *et al.* Robust and efficient multi-frequency temporal phase unwrapping: optimal fringe frequency and pattern sequence selection. *Opt. Express* **25**, 20381–20400 (2017).
19. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nat.* **521**, 436 (2015).
20. Sinha, A., Lee, J., Li, S. & Barbastathis, G. Lensless computational imaging through deep learning. *Opt.* **4**, 1117–1125 (2017).
21. Rivenson, Y., Zhang, Y., Günaydin, H., Teng, D. & Ozcan, A. Phase recovery and holographic image reconstruction using deep learning in neural networks. *Light. Sci. & Appl.* **7**, 17141 (2018).
22. Feng, S. *et al.* Fringe pattern analysis using deep learning. *Adv. Photonics* **1**, 025001 (2019).
23. Shimobaba, T. *et al.* Computational ghost imaging using deep learning. *Opt. Commun.* **413**, 147–151 (2018).
24. Lyu, M. *et al.* Deep-learning-based ghost imaging. *Sci. reports* **7**, 17865 (2017).
25. Kiarashinejad, Y., Abdollahramezani, S., Zandehshahvar, M., Hemmatyar, O. & Adibi, A. Deep learning reveals underlying physics of light-matter interactions in nanophotonic devices. *arXiv preprint arXiv:1905.06889* (2019).
26. Kiarashinejad, Y., Abdollahramezani, S. & Adibi, A. Deep learning approach based on dimensionality reduction for designing electromagnetic nanostructures. *arXiv preprint arXiv:1902.03865* (2019).
27. Hemmatyar, O., Abdollahramezani, S., Kiarashinejad, Y., Zandehshahvar, M. & Adibi, A. Full color generation with fano-type resonant hfo<sub>2</sub> nanopillars designed by a deep-learning approach. *arXiv preprint arXiv:1907.01595* (2019).
28. Su, X. & Chen, W. Fourier transform profilometry: a review. *Opt. Lasers Eng.* **35**, 263–284 (2001).
29. Zuo, C. *et al.* Phase shifting algorithms for fringe projection profilometry: A review. *Opt. Lasers Eng.* **109**, 23–59 (2018).
30. Takeda, M. & Mutoh, K. Fourier transform profilometry for the automatic measurement of 3-d object shapes. *Appl. Opt.* **22**, 3977–3982 (1983).
31. Huang, L., Kemao, Q., Pan, B. & Asundi, A. K. Comparison of fourier transform, windowed fourier transform, and wavelet transform methods for phase extraction from a single fringe pattern in fringe projection profilometry. *Opt. Laser Eng.* **48**, 141–148 (2010).
32. Srinivasan, V., Liu, H.-C. & Halioua, M. Automated phase-measuring profilometry of 3-d diffuse objects. *Appl. Opt.* **23**, 3105–3108 (1984).
33. De Groot, P. Derivation of algorithms for phase-shifting interferometry using the concept of a data-sampling window. *Appl. Opt.* **34**, 4723–4730 (1995).
34. Surrel, Y. Design of algorithms for phase measurements by the use of phase stepping. *Appl. Opt.* **35**, 51–60 (1996).
35. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
36. Shi, W. *et al.* Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1874–1883 (2016).
37. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440 (2015).
38. Li, Z., Shi, Y., Wang, C. & Wang, Y. Accurate calibration method for a structured light system. *Opt. Eng.* **47**, 053604 (2008).
39. Zhang, Z. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis machine intelligence* **22** (2000).
40. Feng, S., Chen, Q. & Zuo, C. Graphics processing unit-assisted real-time three-dimensional measurement using speckleembedded fringe. *Appl. Opt.* **54**, 6865–6873 (2015).
41. Liu, K., Wang, Y., Lau, D. L., Hao, Q. & Hassebrook, L. G. Dual-frequency pattern scheme for high-speed 3-d shape measurement. *Opt. Express* **18**, 5229–5244 (2010).
42. Zheng, D., Da, F., Kemao, Q. & Seah, H. S. Phase-shifting profilometry combined with gray-code patterns projection: unwrapping error removal by an adaptive median filter. *Opt. Express* **25**, 4700–4713 (2017).
43. Zuo, C. *et al.* Micro fourier transform profilometry ( $\mu$ ftp): 3d shape measurement at 10,000 frames per second. *Opt. Lasers Eng.* **102**, 70–91 (2018).
44. Yin, W. *et al.* High-speed 3d shape measurement using the optimized composite fringe patterns and stereo-assisted structured light system. *Opt. Express* **27**, 2411–2431 (2019).
45. Lazaros, N., Sirakoulis, G. C. & Gasteratos, A. Review of stereo vision algorithms: from software to hardware. *Int. J. Optomechanics* **2**, 435–462 (2008).
46. Pan, B. Digital image correlation for surface deformation measurement: historical developments, recent advances and future goals. *Meas. Sci. Technol.* **29**, 082001 (2018).
47. Zhang, L., Curless, B. & Seitz, S. M. Spacetime stereo: Shape recovery for dynamic scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, II–367 (2003).
48. Harendt, B., Große, M., Schaffer, M. & Kowarschik, R. 3d shape measurement of static and moving objects with adaptive spatiotemporal correlation. *Appl. Opt.* **53**, 7507–7515 (2014).

## Acknowledgements

This work was supported by National Natural Science Foundation of China (61722506, 61705105, 11574152), National Key R&D Program of China (2017YFF0106403), Final Assembly “13th Five-Year Plan” Advanced Research Project of China (30102070102), Equipment Advanced Research Fund of China (61404150202), The Key Research and Development Program of Jiangsu Province (BE2017162), Outstanding Youth Foundation of Jiangsu Province (BK20170034), National Defense Science and Technology Foundation of China (0106173), “333 Engineering” Research Project of Jiangsu Province (BRA2016407), Fundamental Research Funds for the Central Universities (30917011204), China Postdoctoral Science Foundation (2017M621747), Jiangsu Planned Projects for Postdoctoral Research Funds (1701038A), National Science Center Poland (NCN) (2017/25/B/ST7/02049), Polish National Agency for Academic Exchange (PPN/BEK/2018/1/00511), Faculty of Mechatronics Warsaw University of Technology statutory funds.

## Author contributions

C.Z. proposed the idea. W.Y. and S.F. developed the theoretical description of the method and built the architecture of deep learning to perform phase unwrapping. C.Z., W.Y. and S.F. performed experiments. C.Z. and W.Y. analyzed the data. C.Z. and Q.C. supervised the research. All authors contributed to writing the manuscript.



### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41598-019-56222-3>.

**Correspondence** and requests for materials should be addressed to Q.C. or C.Z.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

## 深度学习技术在条纹投影三维成像中的应用

冯世杰, 左超, 尹维, 陈钱\*

(南京理工大学电子工程与光电技术学院, 江苏南京 210094)

**摘要:** 条纹投影(结构光)三维成像是一种广泛使用的三维成像手段。近年来,集成式的三维传感器发展迅速,特别是基于结构光原理的三维传感器器件已逐渐成为高端智能手机必不可少的一个重要传感单元。然而随着应用需求的不断增多,人们对条纹投影三维成像这项技术的效率、精度、稳定性等方面的要求也越来越高。同时近年来,深度学习技术的飞速发展已经为光学成像技术的发展开启了一扇新的大门,并且从这扇大门中人们注意到伴随着人工智能概念的引入,条纹投影技术的发展也正在经历着新的突破。首先简要介绍了条纹投影三维成像的基本理论。随后举例分析通过运用深度学习技术,起初基于物理模型的条纹投影技术也可成为一种在“数据”驱动下实现的技术,而且在这种情况下,它展现出了超越传统算法的潜力。最后从神经网络模型、训练数据、训练方法等方面,讨论该领域面临的挑战与未来的研究方向。

**关键词:** 条纹投影; 三维成像; 深度学习; 相位恢复

**中图分类号:** O439      **文献标志码:** A      **DOI:** 10.3788/IRLA202049.0303018

## Application of deep learning technology to fringe projection 3D imaging

Feng Shijie, Zuo Chao, Yin Wei, Chen Qian\*

(School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

**Abstract:** Fringe projection(structured light) 3D imaging is a widely used 3D imaging method. In recent years, the integrated three-dimensional sensor has developed rapidly, especially the three-dimensional sensor based on the principle of structured light has gradually become an essential sensor unit for high-end smart phones. However, with the increasing requirements from applications, people have higher and higher requirements on the efficiency, accuracy, stability and other aspects for the fringe projection technique. At the same time, the rapid development of deep learning technology has opened a new door for the development of optical imaging technology, and from this door we notice that with the introduction of the concept of artificial intelligence, the development of fringe projection technology is also experiencing a new breakthrough. In this paper, the basic theory of fringe projection 3D imaging was introduced. Then, by using the deep learning technology, the fringe projection technology based on the physical model can become a technology driven by "data", and in this case, it showed the potential to surpass the traditional algorithm. Finally, the challenges and future research directions in this field from the aspects of neural network model, training data, training methods and so on were discussed.

收稿日期:2019-12-03; 修订日期:2020-01-06

基金项目:国家自然科学基金(61722506, 61705105, 11574152); 总装“十三五”装备预研项目(30102070102); 总装“十三五”装备预研共用技术和领域基金(61404150202); 国防科技项目基金(0106173); 江苏省杰出青年基金(BK20170034); 江苏省重点研发计划项目(BE2017162); 江苏省“333工程”科研项目(BRA2016407); 江苏省光谱成像与智能感知重点实验室开放基金(3091801410411)

作者简介:冯世杰(1989-),男,副教授,博士,主要从事三维成像与计算光学成像方面的研究。Email: shijiefeng@njust.edu.cn

通讯作者:陈钱(1964-),男,教授,博士生导师,博士,主要从事三维成像、光电成像等方面的研究。Email: chenqian@njust.edu.cn

**Key words:** fringe projection; 3D imaging; deep learning; phase retrieval

## 0 引言

人类所处的物理世界空间是三维的,对三维信息的获取和处理技术体现了人类对客观世界的把握能力,因而从某种程度上来说它是体现人类智慧的一个重要标志。传统光探测器仅对被测场景的二维强度敏感而无法感知其三维形貌与深度信息。人类虽可通过自己的双眼来感知三维的世界,但无法对客观事物的三维形貌进行准确量化的描述。三维成像与传感技术作为感知真实三维世界的重要信息获取手段,为重构物体真实几何形貌及后续的三维建模、检测、识别等方面提供数据基础。近年来,随着计算机技术、光学和光电技术的发展,以光信号为载体的光学三维传感技术,融合光电子学、图像处理、计算机视觉与现代信号处理等多学科为一体,已发展成为光学计量和信息光学的最重要的研究领域和研究方向之一。

三维信息获取与处理技术以各种不同的风貌与特色渗透到我们身边的众多领域之中<sup>[1-4]</sup>。在工业设计中,基于三维数字化模型的逆向设计方法可快速获得现有成熟产品的准确和完整的计算机模型,大大缩短产品或模具的研发周期。在虚拟现实领域,大量景物的三维彩色模型化数据已被用于国防、模拟训练、科学试验、3D动画的建构。在医学整形领域,三维数字化技术已广泛用于面部软组织形态修复、外科检测、假牙假肢的量身定做。文物保护领域中,三维彩色数字化技术能以不损伤物体的手段,获得文物的三维信息和表面色彩、纹理,便于长期保存与再现。但在某些领域,如三维测量加工、机器人导航、快速逆向成型、自动化生产线控制、产品质量监控等,仅仅捕获待测物体的三维信息是不够的,三维数据获取的速度与效率直接关系到制造系统的响应能力、产品研发生产能力、以及产品质量保证能力。此外诸如在压模件尺寸监测、冲压板几何形状和形变检测、机车冲撞试验、压力波传播、不连续边界的应力集中、汽车制导中障碍检测、流体力学、流程可视化、运动力学、高速旋转等,这些高速瞬态过程的三维数据快速记录与准确定量再现将有助于描绘和分析动态过程

中物体表面三维形态的变化,并为进一步提取与被测物体相关的结构、形变、应力等物理参量提供数据基础。

条纹投影三维成像因其非接触、高精度、全场测量、点云重建效率高等优点,已成为目前三维传感技术中的主流光学方法<sup>[5-7]</sup>。然而现有研究工作大多集中在静态物体或缓变场景的形貌测量上,通过投影多组光栅条纹并结合格雷码/时间相位展开方法以获得绝对相位信息。这不可避免地延长了数据获取的时间,使其难以对动态物体或者变化场景达到快速响应。如何快速、准确、无歧义地获取目标,特别是运动目标的三维形貌信息是当前条纹投影轮廓术领域的一个亟待解决的问题。该问题直接制约着数字光栅投影技术的适用对象与应用范围,也逐渐成为该领域的研究热点之一<sup>[8-11]</sup>。

2016年,以围棋界AlphaGo击败李世石开始<sup>[12]</sup>,以深度学习为代表的人工智能(AI)技术全面进入了大众的视野,对于它的讨论变得更为火热起来;整个业界普遍认为,它很可能带来下一次科技革命,并且在未来可预见的10多年里,将深刻地改变人们的生活。正如当时的预测,目前人工智能已经在计算机视觉、图像语音处理等多个领域的技术上取得了全面的突破<sup>[13-19]</sup>。与此同时,深度学习技术也在光学成像、计算成像、全息显微等领域逐步渗透<sup>[20-27]</sup>,且展现出巨大的潜力。对基于条纹投影的三维成像而言,深度学习技术已成功应用于条纹图像的包裹相位求解、空域/时域包裹相位展开、条纹去噪、超快三维测量等方面。这些应用向大众展现了在人工智能的辅助下,条纹投影技术在效率、精度等方面取得的新突破。

文中首先将回顾条纹投影三维成像的基本原理。随后将列举深度学习技术在条纹投影三维成像中的典型应用。最后,从神经网络的可解释性、神经网络结构设计、神经网络训练数据获取等五个方面,分析与总结利用深度学习技术实现条纹投影成像面临的挑战和未来的走向。

## 1 基本原理

条纹投影三维成像技术通过将立体视觉中一个



摄像机替换成光源发生器(如投影仪)而实现,原理如图 1 所示。光源向被测物体投影按照一定规则和模式编码的图像,形成主动式三维形态测量。编码图案受到物体表面形状的调制而产生形变,而带有形变的结构光被另外位置的相机拍摄到,通过相机投影光源之间的位置关系和结构光形变的程度可以确定出物体的三维形貌。条纹投影技术本质上区别于干涉测量技术,但它采用的条纹形式和干涉测量中两束相干光干涉产生的原理相类似。相比于立体视觉法,其最大优点在于求解物体初相位时是点对点的运算,即在原理上中心点的相位值不受相邻点光强值的影响,从而避免了物面反光率不均匀或观察视角的偏差引起的误差,测量精度可以达到几十或几个微米。

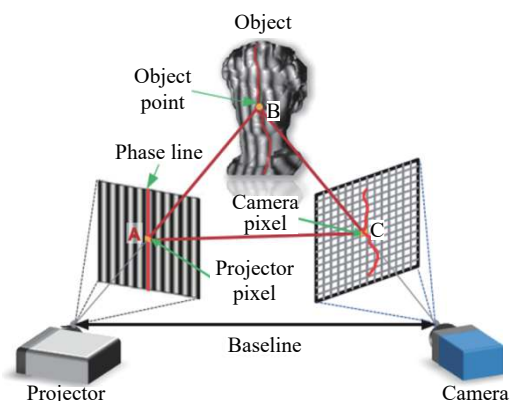


图 1 条纹投影三维成像原理图

Fig.1 Diagram of fringe projection 3D imaging

条纹投影技术大体上包含系统标定与三维成像两个方面。系统标定的目的在于获取相机与投影仪的内外参数,为相位与三维坐标转换提供参考系<sup>[28-29]</sup>。而另一部分三维成像的目的在于通过分析采集的光栅图像,求解相位信息,结合系统标定部分获得的参数进行相位深度之间的转换,完成三维模型重建。文中将简要回顾三维成像部分的基本原理。该部分可细分为三个步骤:条纹分析、相位展开、相位与三维坐标转换。

### 1.1 条纹分析

条纹投影技术通常采用正弦条纹图像作为照明图案对被测表面进行编码,采集的条纹图案一般可表示为:

$$I(x,y) = A(x,y) + B(x,y)\cos\phi(x,y) \quad (1)$$

式中:  $(x,y)$  为像素坐标;  $A$  为背景光强;  $B$  为调制度;

$\phi$  为相位。傅立叶轮廓术<sup>[30,8]</sup>是一种常用的条纹分析方法,通过利用带通滤波器提取光栅频谱的正负一级谱,可获得:

$$I'(x,y) = \frac{1}{2}B(x,y)e^{i\phi(x,y)} \quad (2)$$

随后,利用反正切函数计算相位 $\phi$

$$\phi(x,y) = \arctan \frac{\text{Re}[I'(x,y)]}{\text{Im}[I'(x,y)]} \quad (3)$$

需要注意的是傅立叶轮廓术是一种基于空间滤波的相位计算方法。尽管效率高,但通常假设被测表面为平滑表面,并且需要投影光栅的空间频率足够高<sup>[30]</sup>。

相移轮廓术<sup>[31]</sup>是另一种经常使用的光栅条纹分析法,以使用最为广泛的  $N$  步相移法为例,相机拍摄一系列具有  $2\pi/N$  相对相移的光栅图像

$$I_n(x,y) = A(x,y) + B(x,y)\cos[\phi(x,y) + 2\pi n/N] \quad (4)$$

式中:  $n$  为相移指数 ( $n = 1, 2, \dots, N$ )。当拍摄的图像大于三幅时 (即  $N \geq 3$ ), 利用最小二乘法<sup>[32]</sup>, 可计算物体相位 $\phi$ :

$$\phi(x,y) = \arctan \frac{\sum_{n=1}^N I_n \sin\left(\frac{2\pi n}{N}\right)}{\sum_{n=1}^N I_n \cos\left(\frac{2\pi n}{N}\right)} \quad (5)$$

与傅立叶轮廓术相比,相移法的优势在于相位解算精度高。更进一步,随着相移步数的增加,光栅图像的噪声<sup>[31]</sup>、系统的非线性(如投影仪的  $\Gamma$ )<sup>[33]</sup>以及光栅图像的饱和问题<sup>[34]</sup>对相位计算造成的影响都将减小。

### 1.2 相位展开

无论是傅立叶轮廓术(公式(3)),还是相移轮廓术(公式(5)),解调得到的相位均是包裹相位。它的空间分布是截断的,存在  $2\pi$  相位跳变。为了获得连续的真实空间相位分布,需要对其进行相位展开

$$\Phi(x,y) = \phi(x,y) + 2\pi k(x,y) \quad (6)$$

式中:  $\Phi(x,y)$  为去包裹相位或展开相位;  $k(x,y)$  为光栅条纹的级次。

相位展开算法目的在于确定光栅条纹的级次  $k(x,y)$ 。根据求取条纹级次的原理不同,常见的相位展开方法可以被分为空域展开法<sup>[35]</sup>与时域展开法两类<sup>[36]</sup>。空域相位展开是指利用相邻像素的相位值所提供的约束来计算绝对相位值,但该方法依赖于被测

物体表面连续的假设。如果被测场景中包含多个孤立物体,或者被测物存在处于不连续表面边界的相邻像素的绝对相位值相差超过 $2\pi$ ,则存在条纹级次歧义,从而无法正确展开。与空间相位展开相比,时间相位展开中每个像素的条纹级次都在时间轴上独立计算,无需参考邻近像素,因此可以展开任意复杂形状表面的包裹相位值。但就相位展开的效率而言,时间相位展开通常还需要至少一幅额外的参考相位图。

### 1.3 相位与三维坐标转换

若将投影仪看做“反相机”来处理,根据双目视觉原理<sup>[37]</sup>,对于相机存在如下投影关系:

$$\alpha^c(x, y, 1)^T = K^c [R^c, t^c](X, Y, Z, 1)^T \quad (7)$$

对于投影仪存在如下投影关系

$$\alpha^p(x^p, y^p, 1)^T = K^p [R^p, t^p](X, Y, Z, 1)^T \quad (8)$$

将展开后的相位 $\Phi(x, y)$ 作为线索,可构建相机坐标与投影仪坐标之间的关系:

$$x^p = \frac{\Phi(x, y)}{2\pi f_0} w^p \quad (9)$$

式中: $\alpha$ 为缩放因子; $K$ 为内参; $R$ 为旋转矩阵; $t$ 为平移向量; $f_0$ 为光栅频率; $w^p$ 为投影仪分辨率。在预先矫正系统的畸变后,通过联立公式(7)~(9),可获得的相机像素 $(x, y)$ 对应的三维坐标 $(X, Y, Z)^T$ 。

至此,笔者简要回顾了条纹投影的基本原理。这些基本原理构成了条纹投影技术的物理模型。传统的条纹投影技术是在“物理(模型)”驱动下的技术。下面将介绍通过运用深度学习技术,条纹投影技术也可成为一种在“数据”驱动下的技术,并且在这种情况下,它展现出了超越传统算法的能力。

## 2 基于深度学习的条纹分析

### 2.1 基于深度学习的单幅光栅条纹分析

光栅条纹分析的目的在于解调光栅中蕴含的与深度信息相关的相位信息。单幅光栅条纹分析,也就是空域相位解调法,具有天然的高时域分辨率优势。传统的单帧条纹分析法包括:傅立叶分析法<sup>[30]</sup>、加窗傅立叶法<sup>[38]</sup>、小波分析法<sup>[39]</sup>等。由于所有能运用的解调手段只能局限于一张信息量有限的光栅图像之中,传统的单帧条纹分析方法一般只适合处理表面高度变化平缓的物体,对轮廓陡变、不连续以及物体细

节丰富的区域较为敏感。针对轮廓复杂的物体,难以实现高精度、高分辨率的相位测量。此外这类算法在实施过程中通常需要操作者手动地设定与调节算法参数,因此它们难以完全实现全自动化操作。由于相位解调的结果与算法参数设定有较大关系,对于初学者而言,他们往往难以迅速获得理想的相位解调结果。

为了克服这些问题,Feng 等人<sup>[40]</sup>提出了一个基于深度学习框架的单帧光栅条纹分析法。该方法的思想在于仅采用一张条纹图像作为输入,利用深度神经网络来模拟时域相移法的相位解调过程。如图 2 所示,结合光栅图像的公式(1),通过构建两个卷积神经网络(CNN1 和 CNN2),CNN1 负责从输入条纹图像(I)中提取背景信息(A)。随后 CNN2 利用提取的背景图像(A)和原始输入图像(I)生成所需相位的正弦部分(M)与余弦部分(D)。最后,将该输出的正弦弦结果带入反正切函数计算得到最终的相位分布。为了给深度学习树立一个“优秀”的学习对象,该文作者以标准 12 步相移算法作为学习目标,通过对各类不同的大量样本进行训练,两个卷积神经网络学习各类型条纹图像中相位相关特征的提取。经过适当的训练之后,它们就可以被用于对单幅条纹图像进行全自动分析并且输出对应的高精度相位分布。

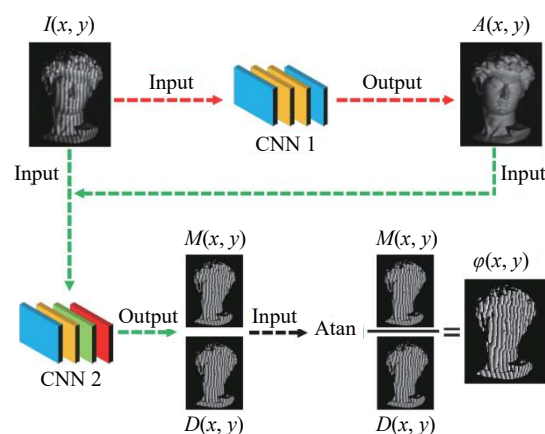


图 2 利用深度学习解调单幅条纹图像中的相位信息流程图<sup>[40]</sup>

Fig.2 Flowchart of phase calculation from a single fringe image using deep neural network<sup>[40]</sup>

实验结果表明,对于复杂表面,基于深度学习的条纹分析技术能够达到传统傅立叶变换法和加窗傅

立叶变换法难以实现的相位解调精度(相位误差降低 50% 以上),且能够有效保持物体边界与轮廓的细节,总体测量效果接近于 12 步相移法(如图 3 所示)。由此可见,该方法为实现“高精度、高效率、全自动”

的条纹投影或相位恢复提供了一条切实可行的方案。仅采用单一条纹图像作为输入,深度神经网络即可快速生成对应的高精度相位分布。整个过程全自动、无需人工干预。

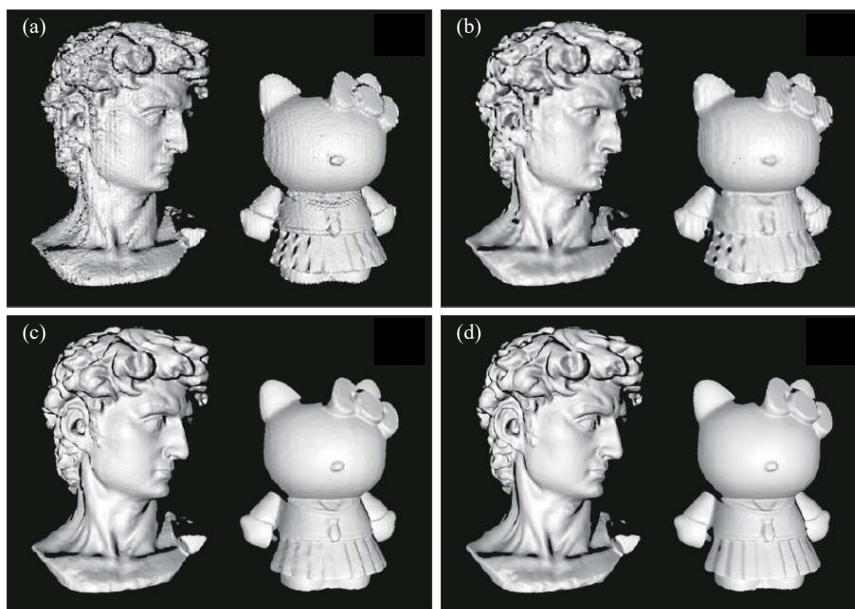


图 3 三维重建结果对比<sup>[40]</sup>。(a) 傅立叶变换法<sup>[30]</sup>, (b) 加窗傅立叶变换法<sup>[38]</sup>, (c) 基于深度学习的条纹分析法, (d) 12 步相移法

Fig.3 Comparison of 3D reconstruction results<sup>[40]</sup>. (a) Fourier transform profilometry, (b) windowed Fourier transform profilometry, (c) fringe analysis based on deep learning, and (d) 12-step phase-shifting profilometry

## 2.2 基于标签增强与区域分块的深度学习条纹分析

同样为了实现针对单幅光栅图像的包裹相位恢复, Shi 等人<sup>[41]</sup>提出了一种基于标签增强与区域分块的深度学习条纹分析技术。Shi 等人建议将原始大小(如 512×512 分辨率)的图片,划分成更小且具有邻域交叠(如 40×40 分辨率)的小图片作为神经网络的输入数据进行相位恢复的训练。由于图片更小,神经网络的训练对于设备的硬件要求可有所降低。相位恢复方面,该方法同样利用深度学习模型进行相位计算的中间变量(光栅条纹的余弦信息)提取。随后,该方法对得到的中间变量进行 Hilbert 变换与反正切函数计算,获取最终的包裹相位信息。作为一种监督式的神经网络学习,为了使神经网络能更好地学习与模仿正确的包裹相位解调,研究人员需要尽可能地制作高精度的标签数据。为达到这一目的, Shi 等人提出首先通过四步相移法得到所需的标签数据,然后采用 Shearlet 变换法对得到标签数据进行滤波,实现光栅

中噪声信号的抑制。图 4 展示了该方法的流程图。

为了证明该方法的有效性, Shi 等人对运动的手掌进行了三维测量。他们选取了运动过程中的六个不同时刻,然后利用神经网络重建相位信息。作为对照,还采用了传统的傅立叶轮廓术(FT)进行相位提取。相位重建结果如图 5 所示。实验表明相比于传统的傅立叶变换轮廓术,该方法(DNN)可更为准确的提取运动手掌的相位信息。

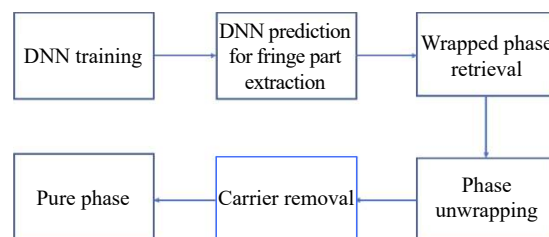


图 4 基于标签增强与区域分块的深度学习条纹分析的相位反演流程图<sup>[41]</sup>  
Fig.4 Flowchart of label enhanced and patch based deep learning fringe analysis for phase retrieval<sup>[41]</sup>



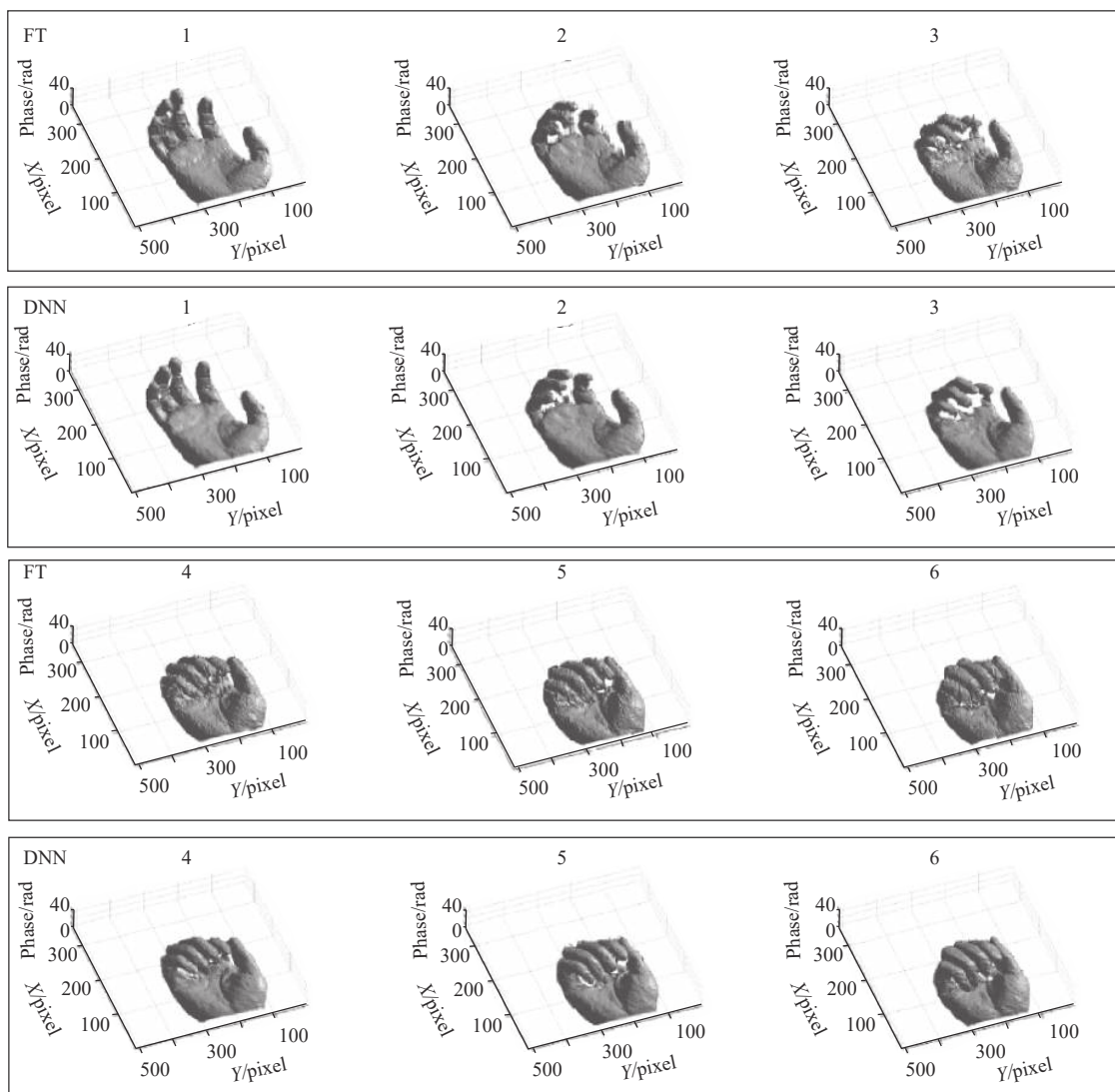


图 5 用 FT 法和 DNN 法对六个不同时刻的运动手掌进行了相位测量<sup>[41]</sup>

Fig.5 Phase measurement of hand movement at six different moments by FT and DNN methods<sup>[41]</sup>

### 2.3 基于深度学习的条纹图像去噪

对于基于条纹图像分析的相位恢复方法,如条纹投影、干涉测量、全息术等,噪声的存在会降低图像的条纹信号信噪比,进而影响相位恢复的准确性。Yan 等人<sup>[42]</sup>提出了利用深度学习算法来降低条纹图的噪声。图 6 显示了该方法的流程图。该方法的核心在于构建一个层数为 20 的深度卷积神经网络。图 6(a)为具有噪声的光栅图,它是整个神经网络的输入数据。该输入随后经过一系列的串联卷积神经网络,最后输出噪声得到抑制的光栅图(图 6(b)所示)。该网络的训练方式为监督式训练,使用的训练标签为不含任何噪声的仿真光栅图(图 6(c)所示)。

为验证该方法的有效性,Yan 等人利用训练好的神经网络预测了六组不同的含噪声条纹图。结果如图 7 所示。图 7(a)显示了含有噪声的原始条纹图,图 7(b)显示了不含噪声的标准条纹图,图 7(c)为利用深度学习法计算得到的去噪后条纹图。与标准结果相比,不难发现深度学习算法成功地学习了如何去去除噪声。此外,基于深度学习向前计算的优势,整个去噪算法的执行速度相比传统方法更快。作为训练数据的生成,该方法采用了计算机仿真的方式进行生成。尽管效率高,但是可节省操作人员的大量时间。但是实际的光栅图与仿真的光栅图存在差异,此差异将对算法的性能提出更高的要求。

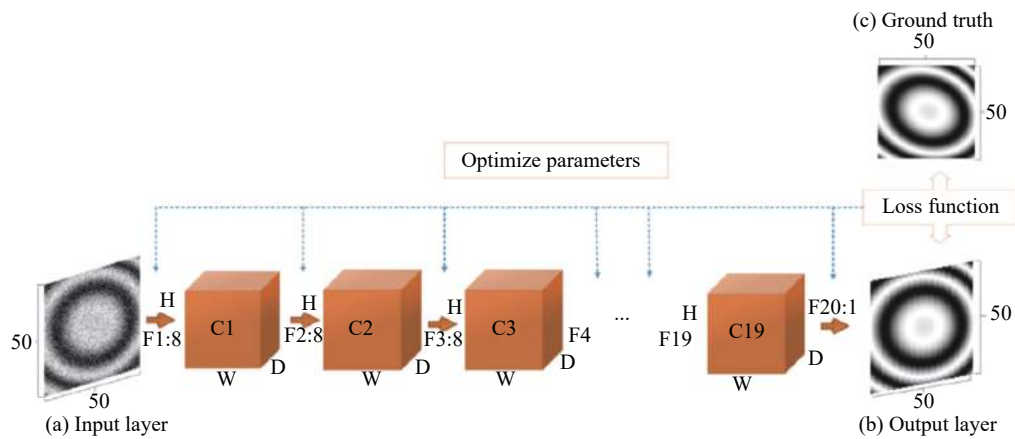


图 6 基于深度学习的条纹图像去噪方法原理图<sup>[42]</sup>

Fig.6 Diagram of fringe image denoising using deep learning<sup>[42]</sup>

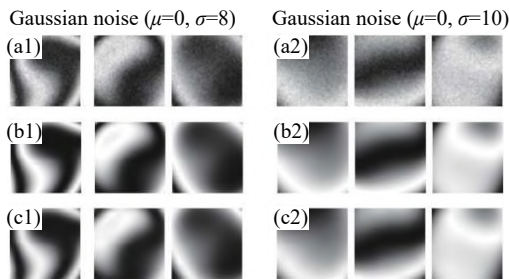


图 7 神经网络的测试结果<sup>[42]</sup>。(a1)、(a2) 带有噪声的仿真条纹图；(b1)、(b2) 不含噪声的条纹图；(c1)、(c2) 用深度学习去噪后的结果

Fig.7 Test results<sup>[42]</sup>。(a1)、(a2) Simulation fringe pattern with noise；(b1)、(b2) fringe pattern without noise；(c1)、(c2) denoised results with deep learning

### 3 基于深度学习的相位展开

如 1.2 节所述的基本原理, 相位展开法大体上分为空域相位展开和时域相位展开两类。按照这一方法分类, 基于深度学习的相位展开方法也可同样划分为空域法和时域法。

#### 3.1 基于深度学习的空域相位展开

##### 3.1.1 相位神经网络 PhaseNet

Spoorthi G.E.等人<sup>[43]</sup>提出了一个基于相位神经网络 (PhaseNet) 用于实现二维的空域相位展开。该方

法的原理如图 8 所示。该网络的输入为包裹相位, 通过构建的神经网络 DCNN, 使其输出条纹级次 (即包裹计数)。该网络由一个编码器、一个对应的解码器和一个像素级分类层组成。研究人员发现由于深度神经网络预测的条纹级次在包裹相位跳变周围区域和存在相位陡变的区域容易发生错误, 他们继续提出了一个基于聚类的条纹级次后处理方法。该方法通过合并互补的方式来增强相位空间分布的平滑度。最后, 原始的包裹相位结合优化后的条纹级次, 可计算最终的展开相位。

为了验证该方法, 研究人员首先利用展开后相位与条纹级次之间的关系, 仿真生成了大量的训练数据。然后利用这些数据训练构建神经网络。神经网络训练结束后, 研究人员又采用一组额外的仿真数据来测试该网络的表现。图 9(a) 显示了神经网络输入的包裹相位, 图 9(b) 和图 9(c) 分别为利用神经网络计算和条纹级次优化后得到展开相位和条纹级次。此外, Spoorthi G.E.等人还发现该方法对于包裹相位中的噪声具有很好地抑制作用。相比于 MATLAB 自带的相位展开函数以及基于质量导向的相位展开法, 该方法的展开相位误差更小。最后, 得益于深度学习

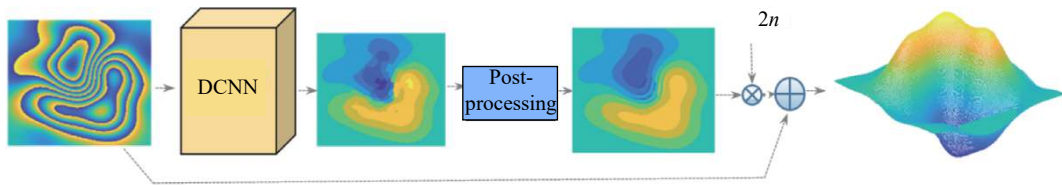


图 8 基于 PhaseNet 的相位展开原理图<sup>[43]</sup>

Fig.8 Schematic of phase unwrapping using PhaseNet<sup>[43]</sup>

方法的一个先天优势,即训练结束后,神经网络的执行是无迭代、无搜索的向前传播计算,该方法的计算速度也比传统的基于质量导向方法更快。但是值得注意的是,在训练和测试过程中,该方法使用的数据同样来自仿真。由于实际的包裹相位情况通常比仿真的相位更加复杂,采用该方法在处理实际的或者更为复杂的包裹相位时还需要更为深入地调试与优化。

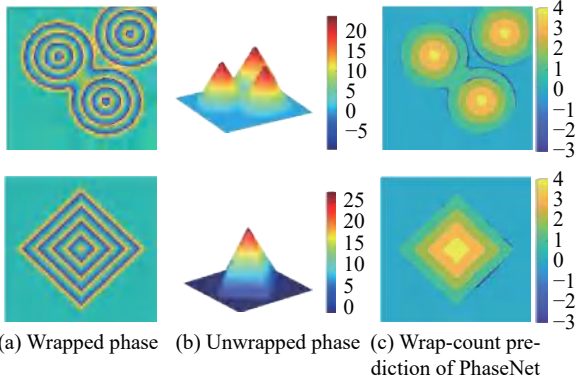


图 9 利用 PhaseNet 展开不同形状包裹相位得到的结果<sup>[43]</sup>。(a) 包裹相位; (b) 展开相位; (c) PhaseNet 输出的条纹级次

Fig.9 Results of different wrapped shapes using PhaseNet<sup>[43]</sup>. (a) Wrapped phase; (b) unwrapped phase; (c) fringe order with PhaseNet

### 3.1.2 一步相位去包裹法

为了解决相位展开过程中的噪声问题与采样不

足引起的混叠问题, Wang 等人<sup>[44]</sup>也利用深度学习技术构建了相位展开神经网络。与 PhaseNet 不同,该方法采用的是具有 U-Net 结构的神经网络,该结构适用于训练数据样本较小的神经网络。图 10 展示了该方法的训练和测试过程。与 PhaseNet 相比, PhaseNet 是利用神经网络预测条纹级次,再结合包裹相位计算展开后的相位。而该方法省去了计算条纹级次的这个中间步骤,直接预测包裹相位对应的去包裹相位。

实验中研究人员对动态的蜡烛火焰进行了相位恢复,结果如图 11 所示。在实验过程中,火焰受到风

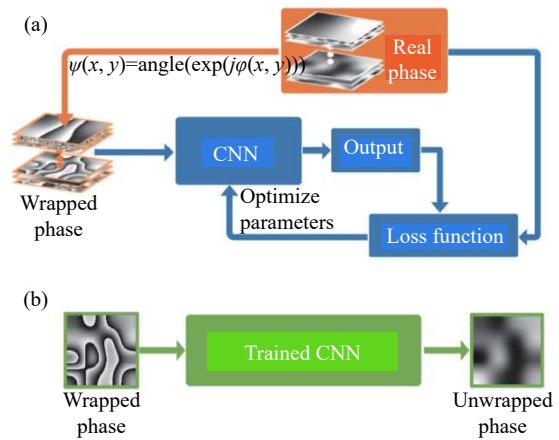


图 10 神经网络的训练与测试<sup>[44]</sup>。(a) 训练; (b) 测试

Fig.10 Schematics of the training and testing of the neural network<sup>[44]</sup>. (a) training; (b) testing

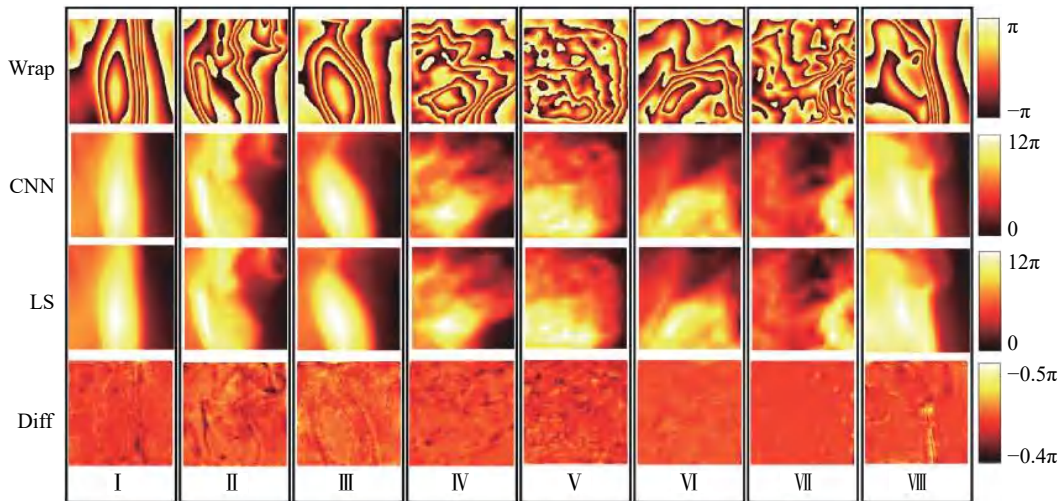


图 11 动态蜡烛火焰的包裹相位展开结果对比<sup>[44]</sup>。Wrap 表示包裹相位; CNN 表示该方法获得的展开相位; LS 表示最小二乘法获得的展开相位; Diff 为 CNN 法与 LS 法计算结果之间的差异

Fig.11 Comparison of results of phase unwrapping of dynamic candle flame<sup>[44]</sup>. Wrap represents the wrapped phase; CNN represents the phase unwrapped by this method; LS represents the phase unwrapped by the least square method; Diff represents the difference between the results of CNN and LS methods



扇的干扰,产生不同的相位分布。该图显示了动态蜡烛火焰的包裹相位、该方法 (CNN) 和 LS 方法在不同帧中重建的展开相位以及它们在不同帧中的相位展开差异。实验表明该方法可成功地重建神经网络在训练阶段中未见过对象的包裹相位。

### 3.2 基于深度学习的时域相位展开

时域相位展开较空域相位展开相比具有恢复不连续或孤立物体表面包裹相位的优势。为了实现这一优势,通常需要采集不同频率光栅对应的多幅包裹相位。时域相位展开有三种代表性的方法<sup>[36]</sup>:多频相位展开方法、多波长(外差)相位展开方法和数论相位展开方法。研究人员发现多频相位展开方法具有最高的展开可靠性和最佳的鲁棒性<sup>[36]</sup>。

通常为了提高测量的效率,笔者需要使用尽量少的的光栅图案。所以一种常见的做法是获取具有两个不同频率光栅的包裹相位。将它们简单地称为低频

相位和高频相位。对于多频相位展开方法,通常低频光栅的频率为 1,即投影光栅只包含一个正弦分布。由于三维重建模型最终来自于高频光栅,为了获得高精度的三维数据,需要尽可能地提升高频光栅的空间频率。但由于噪声等因素的影响,低频光栅相位的展开(辅助)能力有限,它难以正确展开频率大幅提升后的高频光栅包裹相位。

在不改变低频光栅的前提下,为了尽可能地提高可展开的高频光栅空间频率,Yin 等人<sup>[45]</sup>提出了基于深度学习的时域相位展开方法。如图 12 所示,首先利用三步相移法得到两个不同频率光栅对应的包裹相位。然后,将它们作为输入,送入构建的一个四路径的卷积神经网络。该网络经过训练后,可输出高频光栅包裹相位对应的条纹级次。最后结合高频包裹相位,进行相位展开,进而获得被测物体的三维数据。

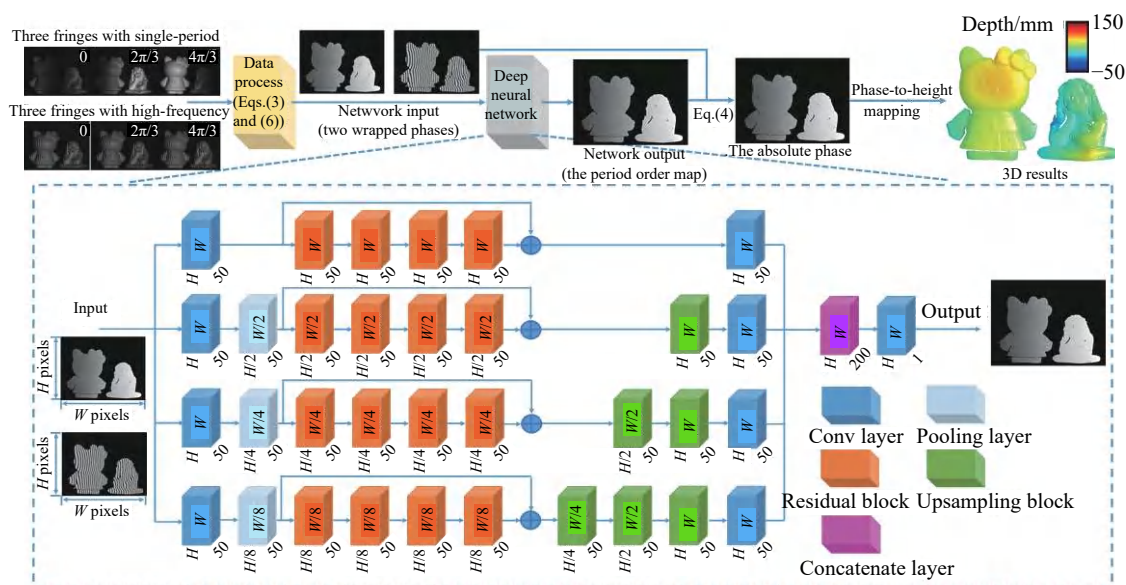


图 12 基于深度学习的时域相位展开方法的示意图<sup>[45]</sup>

Fig.12 Schematic of temporal phase unwrapping using deep learning<sup>[45]</sup>

Yin 等人比较了传统多频相位展开法 (MF-TPU) 与基于深度学习的时域相位展开法。图 13 显示了一组被测物体的相位展开后的 3D 重建结果,背景颜色的深浅显示了相位展开误差率的大小。当投影的高频光栅频率为 16 时,传统多频相位展开法开始表现出较为明显的去包裹错误。当频率持续增加时,相位展开错误率也随之明显增加。但对于基于深度学习

的时域相位展开法,相位展开的错误率并未随光栅频率增加而显著增加。根据该实验结果,即使当高频光栅频率达到 64,该方法的相位展开的正确率仍高于传统方法在频率为 16 时的正确率。由此可见,采用深度学习技术辅助后,时域多频相位展开的正确率可得到大幅提升。

Error rate	MF-TPU		Our method	
	0		20 %	
$f_h=8$				
$f_h=16$				
$f_h=32$				
$f_h=48$				
$f_h=64$				

图 13 针对不同的高频光栅包裹相位 (例如频率分别为 8、16、32、48 和 64), 比较多频相位展开方法 (图中 MF-TPU) 和基于深度学习的时域相位展开方法 (图中 Our method) 的相位展开结果<sup>[45]</sup>

Fig.13 Comparison between traditional MF-TPU and the deep learning based method for high-frequency phase unwrapping (for example, the frequencies are 8, 16, 32, 48 and 64 respectively)<sup>[45]</sup>

#### 4 基于深度学习的深度计算与系统误差标定

##### 4.1 基于深度学习的深度计算

在光学三维成像中, 基于单幅图像的测量方法在测量速度和对运动伪影的鲁棒性等方面均优于基于

多幅图像的结构光测量技术。Sam Vam Der Jeught 等人<sup>[46]</sup>提出了一种完全基于深度学习的单帧光栅解调方法, 该方法可直接从一幅变形的光栅中解调出被测表面的高度 (或深度) 信息。该方法首先通过计算机仿真的方式, 随机生成对应不同高度分布的扭曲光栅条纹。然后将这些的仿真数据输入构建的卷积神经网络, 其结构如图 14 所示。输入的光栅图顺次经过 10 个卷积神经网络, 最后输出对应的高度分布图。

为了训练该神经网络, Sam Vam Der Jeught 等人随机生成了 12 500 幅高度分布图和与它们对应的扭曲光栅图。其中的 10 000 组数据用来训练网络, 剩余的 2 500 组数据用来验证。在 Titan X 的 GPU 平台上, 整个训练耗时接近 12 h。图 15 给出了一组实验结果。该实验一共测试了三个对象: 球面、三角斜面和人脸头像。从第四列的误差分析来看, 对于球面和三角斜面这类变化较为简单的对象, 均方根误差 (RMSE) 误差较小, 而对于轮廓较为复杂的人脸模型, RMSE 误差较大且超过了 1%。

对于基于条纹投影的三维成像而言, 该方法提出的是一个端对端的深度学习训练模型。对于“端对端”的训练策略, 其优势在于将中间结果的计算过程 (如包裹相位计算和相位展开) 与最后的深度计算融合在了一起, 使得轮廓计算一步到位。尽管高效, 但由于部分中间结果, 比如包裹相位, 其存在固有的空间不连续性, 使得神经网络往往难以直接对齐进行准确的拟合。尽管这一过程隐藏在了这个“端对端”的大框架下, 但是从最后恢复的高度图来看, 该方法的测量精度仍有很大的提升空间。此外, 由于该方法同样是基于仿真数据进行的神经网络训练和验证, 当其处理真实拍摄的光栅图像时, 处理过程也许需要更为深入的优化。

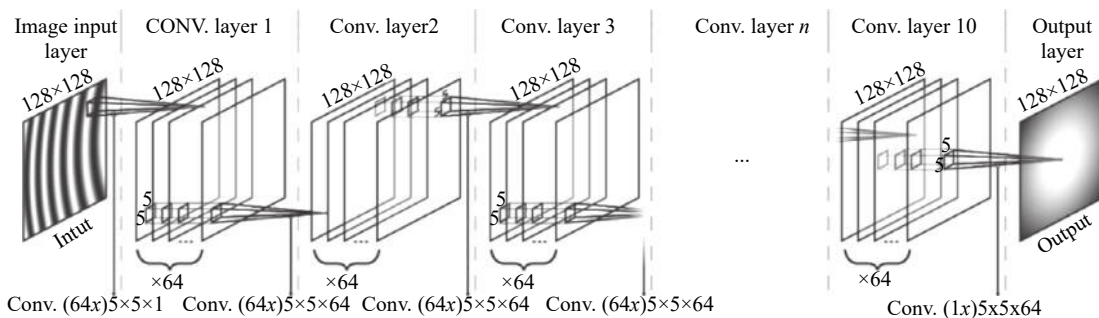


图 14 从单幅条纹图解调高度信息的神经网络结构图<sup>[46]</sup>

Fig.14 Neural network structure diagram of height estimation from a single fringe image<sup>[46]</sup>

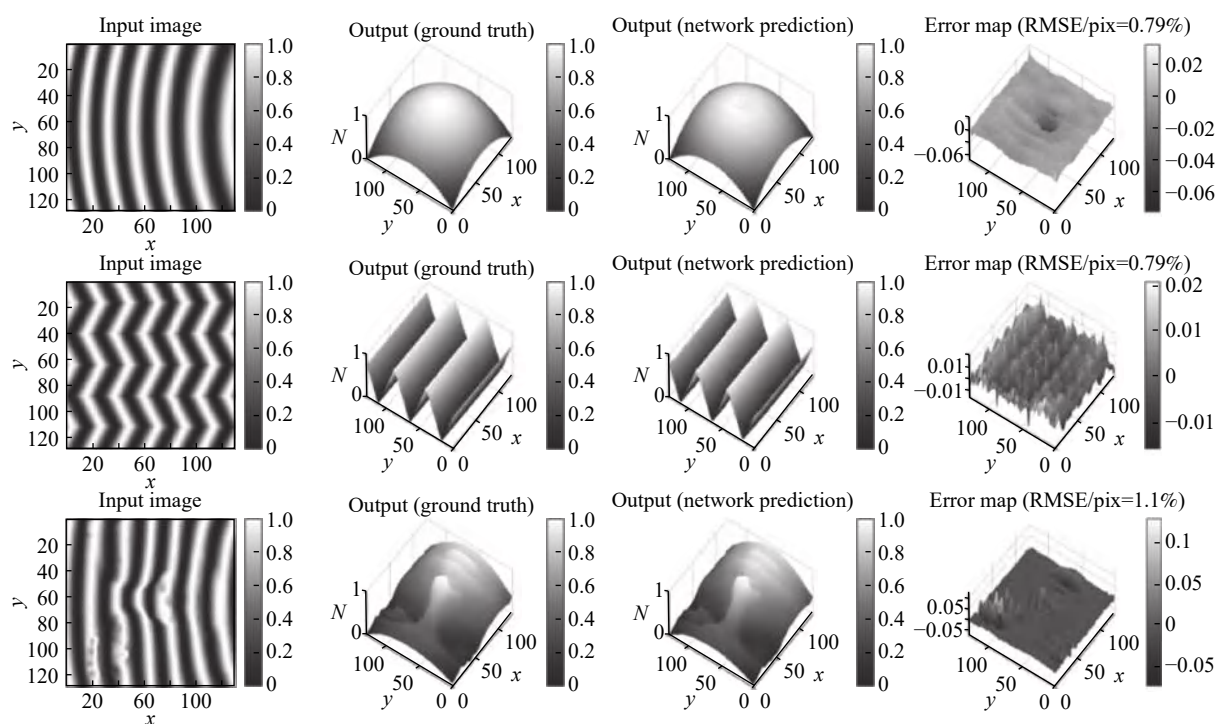


图 15 针对球面、三角斜面和人脸头像光栅图的实验结果图<sup>[46]</sup>。第一列为输入神经网络的条纹图；第二列为真实的高度分布；第三列为神经网络输出的高度分布；最后一列为根据第二列与第三列得出的误差分布图

Fig.15 Experimental results of spherical, triangular bevel and face image grating<sup>[46]</sup>. The first column is the fringe image of the input neural network; the second column is the true simulated height distribution; the third column is the height distribution of the output of the neural network; the last column is the error distribution map based on the second column and the third column

#### 4.2 基于深度学习的系统误差标定

系统标定作为条纹投影中重要的一环一直都是本领域的研究重点。条纹投影系统将双目视觉系统的一个相机替换成投影仪，构建了一种主动式的“双目”视觉三维成像系统。为了重构三维坐标，将投影仪当做“反相机”来处理，然后运用现有双目成像的原理。然而，投影仪的镜头与相机的镜头在设计与功能上存在一定差异。因此严格地说起来，有时投影仪的标定并非能够简单地套用相机标定的模型。这种套用带来的其中一个问题是投影仪镜头畸变矫正问题，即相机的畸变模型难以准确标定投影仪的畸变，致使重建的三维轮廓出现失真。

为了解决这一问题，LV 等人<sup>[47]</sup>提出了一种基于深度学习的投影仪镜头畸变影响矫正方法。如图 16 所示，该方法首先采用传统标定方法，对投影仪和相机的畸变进行矫正，随后利用深度学习矫正剩余的投影仪畸变对三维轮廓造成的影响。

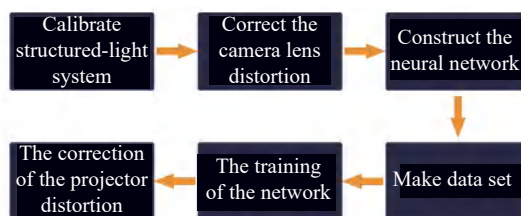


图 16 基于深度学习的投影仪畸变矫正流程图<sup>[47]</sup>

Fig.16 Flowchart for projector distortion correction with deep learning<sup>[47]</sup>

LV 等人提出了一个全连接神经网络，其输入为存在畸变残差的三维空间坐标 $[x, y, z]^T$ ，输出为该空间位置处的深度方向误差 $\Delta z$ 。通过该方法来补偿剩余畸变对三维重造成造成的影响。研究人员利用训练好的模型对平板测量进行的验证，结果如图 17 所示。可以看出，经过矫正后，峰谷误差 (PV) 得到了较大幅度的下降。但是需要指出的是，该方法获取训练的标签数据依赖于对存在残余误差的平面三维数据进行平面拟合，来获得理想的平面三维数据。如果需要更加准确地确定畸变造成的深度误差，也许需要一种精



度更高的方式来确定不同姿态下平板表面的真实三维数据。

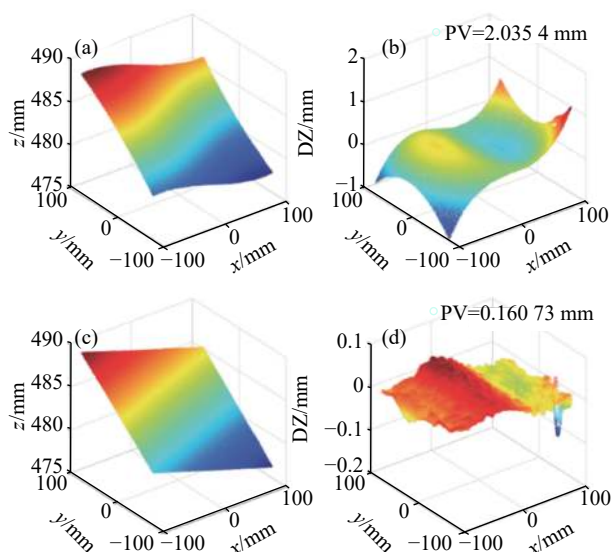


图 17 测试数据结果<sup>[47]</sup>。(a) 原始数据的三维形状; (b) 原始数据的误差分布; (c) 校正后的数据三维形状; (d) 校正后数据的误差分布  
Fig.17 Test results<sup>[47]</sup>. (a) 3D shape of the original data; (b) error distribution of the original data; (c) 3D shape of the corrected data; (d) error distribution of the corrected data

## 5 基于深度学习的超快三维成像

高速摄影技术作为图像获取技术的一个重要分支, 能够对各种瞬态过程进行记录, 广泛应用于军工、航空航天等领域<sup>[48]</sup>。尽管高速 CMOS 器件目前已能实现每秒万帧, 甚至百万帧的拍摄, 但仅能够获取二维平面图像数据。针对瞬态场景, 如何从二维平面图像中获取三维深度图像, 依旧是一个极具挑战性的世界性难题。为此 Feng 等人<sup>[49]</sup> 提出了微频移深度学习轮廓术, 研制了基于数字光栅投影的瞬态三维轮廓测量系统, 测量速度可达每秒 20,000 帧三维数据。

为了满足超快测量中相位信息的高效提取, 高速三维成像中使用数量更少的光栅条纹对运动物体进行编码, 可以减小物体运动对三维重建造成的干扰。同时为了确保三维重建的精度, 该方法最终使用了三种不同的高频率光栅。该方法的原理如图 18 所示, 首先利用深度学习算法计算这三幅光栅图中的相位信息, 其中一幅用于重构三维轮廓, 另外两幅用来辅助相位的绝对展开。最后根据标定的系统参数, 重构光栅图像中蕴含的三维轮廓数据。

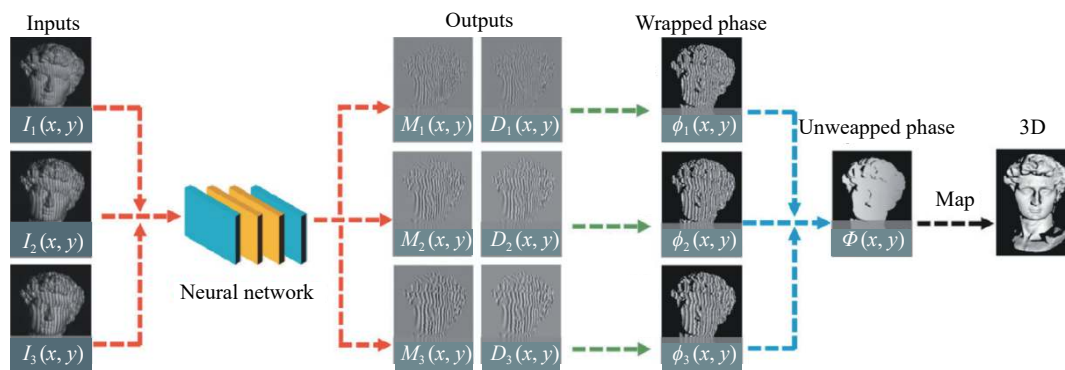


图 18 微频移深度学习轮廓术原理图<sup>[49]</sup>

Fig.18 Diagram of micro deep learning profilometry<sup>[49]</sup>

为了测试该方法的有效性, Feng 等人设置了一个瞬态场景。该场景由一个静态的石膏像和一个下落的乒乓球组成。相机拍摄速度为 20 000 帧/s, 记录了乒乓球落地与反弹的全过程。这两个物体均未在神经网络的训练过程中出现。图 19 的第一行显示了在不同时刻下拍摄的光栅图像。图 19 的第二行显示了该时刻对应的重构三维模型。可以看出整个乒乓球的下落过程不到 0.1 s, 根据三维重建的结果可知, 针

对不同的时刻, 该方法成功地恢复了具有不同运动状态的物体轮廓。相比于传统超快三维成像方法, 该方法表明得益于深度学习算法的强大运算能力, 可在光栅图像数量减少的前提下, 依然精确恢复物体的轮廓信息。因此基于人工智能的辅助, 基于条纹投影的超快三维成像可朝着更高的时间与空间分辨率方向发展。

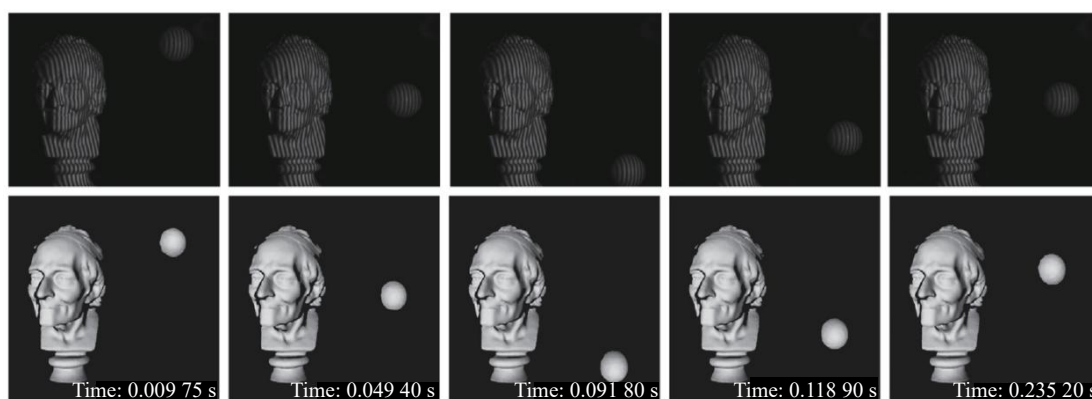


图 19 针对下落的乒乓球和静态石膏像进行的高速三维成像,速度为 20 000 帧/s<sup>[49]</sup>

Fig.19 High speed 3D imaging of a falling table tennis and static plaster at speed of 20 000 frame/s<sup>[49]</sup>

## 6 挑战与未来的方向

### 6.1 深度学习到底学到了什么?

如第一章基本原理所介绍,条纹投影技术的三维成像部分主要包括条纹分析、相位展开、相位深度映射这几个方面。通过第二章至第五章的介绍,笔者发现当前研究人员正尝试着用深度学习技术替代传统方法以实现上述几个方面中的某一项内容,或者全部内容(端对端的策略)。然而对于大多数研究人员而言,深度学习方法预测最终结果的过程仍是一个“黑箱子”——只能通过最终的测试结果来判断神经网络的优劣。由于难以把握神经网络的推演机理,使得优化和提升神经网络性能的目标沦为了大量的试错。尤其是对于大规模的神经网络,巨量的参数使得完成一次训练通常需要数个昼夜甚至更久。多次且无明确方向的试错易造成时间的大量浪费。

近年来越来越多的研究人员意识这个问题的重要性,为了解神经网络的学习过程,Zeiler等<sup>[50]</sup>提出了一种针对卷积神经网络的可视化方法。该方法通过对神经网络学习的特征进行可视化,为优化网络结构、提升预测的准确性提供了思路。

### 6.2 深度神经网络的架构设计与优化

针对具体的条纹投影应用(如计算包裹相位、相位展开、高动态范围成像等),到底什么样的神经网络合适?尽管从前人相似工作中能找到网络结构设计的灵感,但是在神经网络后期的调试与优化过程中,如何调整超参数(如神经网络的类型,卷积神经网络中滤波器的尺寸,抽取特征的数量等)使得能够在自己的应用上表现出色仍是一个难以回答的问题。通

过试错法进行超参数的调整尽管有一定效果,但时间成本过高。此外,当神经网络的规模足够大时,想要快速地输出结果对计算平台的硬件也是一种考验。对于固定的服务器而言,这种影响相对较小。但是对于移动终端或者穿戴设备,如手机、平板、智能手表等,通常难以将规模过大的神经网络部署到这些设备上,而这时需要考虑对网络结构进行压缩。

令人欣喜的是,近年来自动化机器学习(AutoML)成为深度学习技术领域的一个研究热点。自动机器学习的目标就是使计算机自动地做出上述的决策。自动机器学习采用:超参数优化<sup>[51]</sup>(Hyper-parameter Optimization)、元学习<sup>[52]</sup>(Meta Learning)、神经网络架构搜索<sup>[53]</sup>(Neural Architecture Search)等方式自动搜索理想网络结构与超参数。使用者只需提供训练数据,自动机器学习系统就能自动地决定最佳的训练方案。让不同领域的研究人员不必苦恼于学习各种机器学习的算法。

### 6.3 训练数据的获取与标注成本高

神经网络并非一个新概念,它实际上已具有几十年的历史。但是由于它是一种数据驱动的计算方法,几十年前的数据规模并未像今天一样地井喷式增长。因此当前迅速发展的互联网时代积累下的数据与算力释放了深度学习神经网络的潜力。

但就当前而言,对于条纹投影技术领域,训练数据的大规模获取与正确标注仍需要耗费大量的人力和物力成本。加之公开的数据集稀少,这都增加了深度学习技术的实施难度。尽管采用仿真的方式获取数据集可在一定程度上降低训练数据采集过程中的

成本。但是仿真数据受制于有限的预设参数,它并不能完全等于真实数据。而深度学习的强大能力就在于学习与发掘输入数据与输出数据之间的潜在联系。因此,如何快速获得准确可靠的训练数据是提高深度学习技术在条纹投影技术领域应用效率的一个重要问题。值得注意的是迁移学习将是解决这一问题的一个潜在方案。迁移学习<sup>[54]</sup>的初衷是节省人工标注样本的时间,让模型可以通过已有的标记数据向未标记数据迁移,从而训练出适用于未标记数据的运算模型。

#### 6.4 深度神经网络泛化能力的思考

泛化能力评价的是一个神经网络在完成训练后,在处理“从未遇见过”的输入数据时的表现。对于传统的条纹投影方法而已,得益于构建的数学模型普适通用,对于满足朗伯体假设条件的所有测量对象,均可获得较为理想的三维成像。但是如前所述,深度学习技术是以数据为导向的算法,它依赖于大量的训练数据为其良好的表现提供基础。因此当训练数据的类型较少时,深度神经网络往往难以抽取以及学习有效的图像特征映射。为了提升神经网络在处理全新场景的能力,大规模的训练数据通常是必不可少的。

但是,笔者认为关于神经网络的泛化能力应该能够一分为二的看待。这就像是“通才”与“专才”。“通才”掌握知识全面,但深度有所不足,且往往需要大量的时间累积以获得丰富的知识储备。而“专才”尽管只专注于部分领域,但能够做到精益求精。其实“通才”与“专才”都是社会发展或不可或缺的。

因此,对于条纹投影的应用而言,如果拟研制系统设计的潜在对象类型本身就较为单一,通过单方面地增加相同类型的训练数据就应该能对其性能提高发挥积极的效用。同时还能节省设备的开发周期,有利于专用系统的快速研发。笔者认为一切从实际出发,具体问题具体分析,才能最大限度地发挥深度学习技术的特长。

#### 6.5 “数据驱动”与“物理驱动”双引擎并存

深度学习的强大能力源于大量的训练数据支撑与驱动。因此本质上来说,这样的人工智能只能机械式的学习而缺乏推理能力。图灵奖得主、贝叶斯网络之父 Judea Pearl 曾指出当前的深度学习“不过只是曲线拟合”。以条纹投影中的条纹分析为例,根据第二

节中所述方法,目前基本的策略是两步走:先利用深度学习技术学习求解某项中间变量(比如条纹的实部信息与虚部信息),然后再将中间变量代入反正切函数计算最终的包裹相位。由于缺乏推理能力,神经网络不知道包裹相位具有不连续空间跳变性质的先验知识,难以训练神经网络直接计算包裹相位。

基于物理模型的算法仍是当今世界科技的核心。尽管在许多任务中,数据驱动模型算法表现已优于物理驱动模型算法,但“数据驱动”的可解释性仍是个挑战。对于条纹投影的应用,我们认为需要向当前以“数据驱动”的神经网络引入“物理模型”。只有把数据和物理结合起来,综合运用数据与物理两个世界的优势,才能更全面地揭示出问题的本质。

## 7 结 论

文中回顾并讨论了近年来基于深度学习的条纹投影三维成像技术的研究现状。尽管这一研究方向才刚起步,但对于已经经历了几十年发展历程的条纹投影技术而言,这无疑是一个具有强大潜力的新增长点。总的来说,在深度学习技术的辅助下,将条纹投影技术放在以“数据驱动”的神经网络模型中重新考虑后,笔者发现的优势包括:

(1) 相位测量效率的提升 当前面向运动物体的快速三维成像是条纹投影技术的一个热点研究方向。尽管通过补偿的方式可有效去除由物体运动引起的运动误差,但当物体运动过快时,这类补偿算法仍难以发挥期待的效果。而深度学习技术仅采用单幅光栅图像即可准确恢复物体的相位信息,从而减少了三维图像重建所需的条纹图像数量,提高了成像的效率。结合多视角几何理论,该方法有望成为快速三维成像的一种理想手段。

(2) 相位测量精度的提高 作为条纹投影技术而言,三维成像质量的优劣直接取决于相位质量的好坏。对于用于求解相位信息的神经网络,当其经过适当的训练,其计算得到的包裹相位比传统的单幅条纹分析方法获得的相位信息更加准确,有效降低相位误差,相位解调精度已接近相移法。

(3) 成像稳定性的提升 将深度学习应用于相位展开,无论是空域展开还是时域展开,经过深度神经网络的处理,原始包裹相位中的噪声均得到了较好的



抑制。这使得即使在信噪比不理想的情况下,依然能获得准确的去包裹相位。此外,将深度学习技术直接应用于条纹图像的去噪,也能较好地去除图像中的噪声。

尽管在深度学习的辅助下,条纹投影三维成像取得新的研究进展。但是人们依然需要意识到,深度学习技术目前还无法做到真正的人工智能,这其中还有很长的路要走。为了能够更好地将深度学习技术应用于条纹投影三维成像技术的研究之中,首先需要明白“深度学习到底学到了什么?”。由于难以把握神经网络的推演机理,为了提升神经网络的性能,大部分人能做的只有试错。因此急需理解神经网络到底是如何思考我们为其布置的任务,进而找到优化神经网络的有效线索,避免无明确方向的试错造成的时间浪费。

在不久的将来,借助于自动机器学习,人们完全可以期待深度神经网络根据自己部署的需求,通过自身的迭代优化,自动地给出最佳的网络架构设计与优化。自动的机器学习将进一步降低深度技术应用的门槛,为条纹投影技术研究与应用深度定制提供高效可靠的方案。

对于基于深度学习的条纹投影技术研究而言,目前的训练数据基本都需要实地采集与标注,这需要耗费大量的人力和物力成本。在仿真数据尚不能完全代替实拍数据的前提下,基于少量样本的迁移学习将是提高研究效率的一个有效手段。同时,为了保障训练的神经网络能够处理各种不同类型的物体,需要在训练过程中尽可能多的让神经网络接触不同的物体,以提升其泛化能力。但是对于某些专用设备的研制,我们也许能够反向运用这种泛化能力,利用少量的同类样本训练研究针对特定样本的专用算法。最后,为进一步提升神经网络的性能,可在神经网络模型的构建或者迭代过程中加入“物理驱动”的引擎,这样有利于神经网络更为全面地认识问题的本质。

综上所述,条纹投影三维成像技术是一个极具发展前景的三维图像获取技术。在人工智能的辅助下,基于深度学习的条纹投影三维成像在相位测量效率、相位测量精度与三维成像稳定性等方面得到显著提升。这将推动条纹投影技术的进一步快速发展,以及带动该技术在更多领域的深入应用。

## 参考文献:

- [1] Harding K. Industrial metrology: engineering precision [J]. *Nature Photonics*, 2008, 2(11): 667.
- [2] Luhmann T. Close range photogrammetry for industrial applications [J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2010, 65(6): 558–569.
- [3] Ma Y, Soatto S, Koseck J, et al. An Invitation to 3-D Vision: from Images to Geometric Models[M]. New York: Springer Science & Business Media, 2012, 26.
- [4] Jiang H, Zhao H, Li X. High dynamic range fringe acquisition: A novel 3-D scanning technique for high-reflective surfaces [J]. *Optics and Lasers in Engineering*, 2012, 50(10): 1484–1493.
- [5] Salvi J, Fernandez S, Pribanic T, et al. A state of the art in structured light patterns for surface profilometry [J]. *Pattern Recognition*, 2010, 43(8): 2666–2680.
- [6] Feng S, Zuo C, Tao T, et al. Robust dynamic 3-D measurements with motion-compensated phase-shifting profilometry [J]. *Optics and Lasers in Engineering*, 2018, 103: 127–138.
- [7] Zhang Z H. Review of single-shot 3D shape measurement by phase calculation-based fringe projection techniques [J]. *Optics and Lasers in Engineering*, 2012, 50(8): 1097–1106.
- [8] Su X, Zhang Q. Dynamic 3-D shape measurement method: a review [J]. *Optics and Lasers in Engineering*, 2010, 48(2): 191–204.
- [9] Wang Y, Liu Z, Jiang C, et al. Motion induced phase error reduction using a Hilbert transform [J]. *Optics Express*, 2018, 26(26): 34224.
- [10] Feng S, Chen Q, Zuo C, et al. Fast three-dimensional measurements for dynamic scenes with shiny surfaces [J]. *Optics Communications*, 2017, 382: 18–27.
- [11] Heist S, Lutzke P, Schmidt I, et al. High-speed three-dimensional shape measurement using GOBO projection [J]. *Optics and Lasers in Engineering*, 2016, 87: 90–96.
- [12] Borowiec S. AlphaGo seals 4-1 victory over Go grandmaster Lee Sedol [J]. *The Guardian*, 2016: 15.
- [13] Z'BONTAR J, Lecun Y. Stereo matching by training a convolutional neural network to compare image patches[J]. *The Journal of Machine Learning Research*, 2016, 17(1): 2287–2318.
- [14] Luo W, Schwing A G, Urtasun R. Efficient deep learning for stereo matching[C]//*IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016: 5695–5703.
- [15] Li S, Deng M, Lee J, et al. Imaging through glass diffusers using densely connected convolutional networks[J]. arXiv: 1711.06810[physics], 2017.

- [16] Moriya T, Roth H R, Nakamura S, et al. Unsupervised segmentation of 3D medical images based on clustering and deep representation learning[J]. arXiv: 1804.03830[cs], 2018: 71.
- [17] Li H, Wei T, Ren A, et al. Deep reinforcement learning: framework, applications, and embedded implementations[J]. arXiv: 1710.03792[cs], 2017.
- [18] Kuznetsov Y, Stuckler J, Leibe B. Semi-supervised deep learning for monocular depth map prediction[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, 2017: 2215-2223.
- [19] Kendall A, Grimes M, Cipolla R. PoseNet: A convolutional network for real-time 6-DOF camera relocalization[C]//2015 IEEE International Conference on Computer Vision (ICCV), 2015: 2938-2946.
- [20] Wang H, Rivenson Y, Jin Y, et al. Deep learning enables cross-modality super-resolution in fluorescence microscopy [J]. *Nature Methods*, 2019, 16(1): 103-110.
- [21] Rivenson Y, Zhang Y, GÜNAYDIN H, et al. Phase recovery and holographic image reconstruction using deep learning in neural networks [J]. *Light: Science & Applications*, 2018, 7(2): 17141.
- [22] Nguyen T, Xue Y, Li Y, et al. Deep learning approach for Fourier Ptychography microscopy [J]. *Optics Express*, 2018, 26(20): 26470.
- [23] Horisaki R, Takagi R, Tanida J. Learning-based imaging through scattering media [J]. *Optics Express*, 2016, 24(13): 13738.
- [24] Lyu M, Wang W, Wang H, et al. Deep-learning-based ghost imaging [J]. *Scientific Reports*, 2017, 7(1): 17865.
- [25] Nehme E, Weiss L E, Michaeli T, et al. Deep-STORM: super-resolution single-molecule microscopy by deep learning [J]. *Optica*, 2018, 5(4): 458-464.
- [26] Fang L, Cunefar D, Wang C, et al. Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search [J]. *Biomedical Optics Express*, 2017, 8(5): 2732-2744.
- [27] Li Y, Xue Y, Tian L. Deep speckle correlation: a deep learning approach toward scalable imaging through scattering media [J]. *Optica*, 2018, 5(10): 1181-1190.
- [28] Zhang S, Huang P S. Novel method for structured light system calibration [J]. *Optical Engineering*, 2006, 45(8): 083601.
- [29] Yin Y, Peng X, Li A, et al. Calibration of fringe projection profilometry with bundle adjustment strategy [J]. *Optics Letters*, 2012, 37(4): 542-544.
- [30] Takeda M, Mutoh K. Fourier transform profilometry for the automatic measurement of 3-D object shapes [J]. *Applied Optics*, 1983, 22(24): 3977-3982.
- [31] Zuo C, Feng S, Huang L, et al. Phase shifting algorithms for fringe projection profilometry: A review [J]. *Optics and Lasers in Engineering*, 2018, 109: 23-59.
- [32] Malacara D. *Optical Shop Testing*[M]. Hoboken, New Jersey: John Wiley & Sons, 2007, 59.
- [33] Hoang T, Pan B, Nguyen D, et al. Generic gamma correction for accuracy enhancement in fringe-projection profilometry [J]. *Optics Letters*, 2010, 35(12): 1992-1994.
- [34] Feng S, Zhang L, Zuo C, et al. High dynamic range 3D measurements with fringe projection profilometry: a review [J]. *Measurement Science and Technology*, 2018, 29(12): 122001.
- [35] Su X, Chen W. Reliability-guided phase unwrapping algorithm: a review [J]. *Optics and Lasers in Engineering*, 2004, 42(3): 245-261.
- [36] Zuo C, Huang L, Zhang M, et al. Temporal phase unwrapping algorithms for fringe projection profilometry: A comparative review [J]. *Optics and Lasers in Engineering*, 2016, 85: 84-103.
- [37] Hartley R, Zisserman A. *Multiple View Geometry in Computer Vision*[M]. Cambridge: Cambridge University Press, 2004: 673.
- [38] Kemao Q. Two-dimensional windowed Fourier transform for fringe pattern analysis: principles, applications and implementations [J]. *Optics and Lasers in Engineering*, 2007, 45(2): 304-317.
- [39] Zhong J, Weng J. Spatial carrier-fringe pattern analysis by means of wavelet transform: wavelet transform profilometry [J]. *Applied Optics*, 2004, 43(26): 4993-4998.
- [40] Feng S, Chen Q, Gu G, et al. Fringe pattern analysis using deep learning [J]. *Advanced Photonics*, 2019, 1(2): 1.
- [41] Shi J, Zhu X, Wang H, et al. Label enhanced and patch based deep learning for phase retrieval from single frame fringe pattern in fringe projection 3D measurement [J]. *Optics Express*, 2019, 27(20): 28929.
- [42] Yan K, Yu Y, Hu C, et al. Fringe pattern denoising based on deep learning [J]. *Optics Communications*, 2019, 437: 148-152.
- [43] Spoorthi G E, Gorthi S, Gorthi R K S S. PhaseNet: A deep convolutional neural network for two-dimensional phase unwrapping [J]. *IEEE Signal Processing Letters*, 2019, 26(1): 54-58.
- [44] Wang K, Li Y, Kemao Q, et al. One-step robust deep learning phase unwrapping [J]. *Optics Express*, 2019, 27(10): 15100.
- [45] Yin W, Chen Q, Feng S, et al. Temporal phase unwrapping using deep learning [J]. *Scientific Reports*, 2019, 9(1): 20175.
- [46] Van Der Jeught S, Dirckx J J J. Deep neural networks for single

- shot structured light profilometry [J]. *Optics Express*, 2019, 27(12): 17091.
- [47] Lv S, Sun Q, Zhang Y, et al. Projector distortion correction in 3D shape measurement using a structured-light system by deep neural networks [J]. *Optics Letters*, 2020, 45(1): 204–207.
- [48] Zuo C, Tao T, Feng S, et al. Micro Fourier Transform Profilometry ( $\mu$  FTP): 3D shape measurement at 10,000 frames per second [J]. *Optics and Lasers in Engineering*, 2018, 102: 70–91.
- [49] Feng S, Zuo C, Yin W, et al. Micro deep learning profilometry for high-speed 3D surface imaging [J]. *Optics and Lasers in Engineering*, 2019, 121: 416–427.
- [50] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[J]. arXiv: 1311.2901[cs], 2013.
- [51] Bergstra J, Bengio Y. Random search for hyper-parameter optimization [J]. *Journal of Machine Learning Research*, 2012, 13(2): 281–305.
- [52] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 1126–1135.
- [53] Zoph B, Le Q V. Neural architecture search with reinforcement learning[J]. arXiv preprint arXiv: 1611.01578, 2016.
- [54] Tan C, Sun F, Kong T, et al. A survey on deep transfer learning[C]//International Conference on Artificial Neural Networks, 2018: 270–279.



DOI: [10.29026/oea.2022.210021](https://doi.org/10.29026/oea.2022.210021)

# Deep-learning-enabled dual-frequency composite fringe projection profilometry for single-shot absolute 3D shape measurement

Yixuan Li<sup>1,2†</sup>, Jiaming Qian<sup>1,2†</sup>, Shijie Feng<sup>1,2</sup>, Qian Chen<sup>1,2\*</sup> and Chao Zuo<sup>1,2\*</sup>

Single-shot high-speed 3D imaging is important for reconstructions of dynamic objects. For fringe projection profilometry (FPP), however, it is still challenging to recover accurate 3D shapes of isolated objects by a single fringe image. In this paper, we demonstrate that the deep neural networks can be trained to directly recover the absolute phase from a unique fringe image that involves spatially multiplexed fringe patterns of different frequencies. The extracted phase is free from spectrum-aliasing problem which is hard to avoid for traditional spatial-multiplexing methods. Experiments on both static and dynamic scenes show that the proposed approach is robust to object motion and can obtain high-quality 3D reconstructions of isolated objects within a single fringe image.

**Keywords:** fringe projection profilometry (FPP); phase unwrapping; deep learning

Li YX, Qian JM, Feng SJ, Chen Q, Zuo C. Deep-learning-enabled dual-frequency composite fringe projection profilometry for single-shot absolute 3D shape measurement. *Opto-Electron Adv* 5, 210021 (2022).

## Introduction

The development of information technology has accelerated human life into the digital three-dimensional (3D) world. Among many 3D optical measurement technologies, fringe projection profilometry (FPP) stands out as one of the most promising 3D imaging methods due to its non-contact, high spatial resolution, high measurement accuracy, and good system flexibility<sup>1–5</sup>. Nowadays, FPP has been widely applied in intelligent manufacturing, cultural relic scanning, human-computer interaction and some other fields<sup>6–9</sup>. In some important applications, such as rapid reverse engineering and online quality control<sup>10,11</sup>, it is essential to obtain high-quality 3D in-

formation in continuously changing dynamic scenes<sup>12–14</sup>. For FPP, the projector projects a series of fringe patterns onto the target object, and then the camera captures these images modulated and deformed by the object. With the captured fringe patterns, the phase information of the measured object can be extracted through the fringe analysis algorithms. The most popular fringe analysis approaches are the Fourier transform (FT) methods<sup>15–19</sup> and the phase-shifting (PS) methods<sup>20,21</sup>. The FT approaches can utilize only a single high-frequency fringe pattern, where the phase information is recovered by applying a properly designed band-pass filter, such as the Hanning window, to extract phase-related spectrum

<sup>1</sup>Smart Computational Imaging (SCI) Laboratory, Nanjing University of Science and Technology, Nanjing 210094, China; <sup>2</sup>Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing University of Science and Technology, Nanjing 210094, China.

<sup>†</sup>These authors contributed equally to this work.

\*Correspondence: Q Chen, E-mail: [chenqian@njust.edu.cn](mailto:chenqian@njust.edu.cn); C Zuo, E-mail: [zuochao@njust.edu.cn](mailto:zuochao@njust.edu.cn)

Received: 13 February 2021; Accepted: 15 July 2021; Published online: 10 March 2022



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License.

To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022. Published by Institute of Optics and Electronics, Chinese Academy of Sciences.

information in the frequency domain. However, spectrum aliasing may cause low phase quality around discontinuities and isolated regions of the phase map. Unlike the FT methods, the PS technologies usually require three or more PS fringe patterns in the time domain to retrieve the phase map. Such methods are quite robust to ambient illumination and varying surface reflectivity, and can achieve pixel-wise phase measurement with high resolution and accuracy. For both FT and PS algorithms, the retrieved phase distribution is mathematically wrapped to principle values of arctangent function ranging between  $-\pi$  and  $\pi$ . Consequently, the phase value is wrapped whenever there is a  $2\pi$  jump. To solve the phase ambiguity problem and establish a unique pixel correspondence between the camera and the projector to ensure correct 3D reconstruction, phase unwrapping must be carried out. One of the most commonly used phase unwrapping methods are the temporal phase unwrapping (TPU) algorithms<sup>22</sup>, which can obtain the absolute phase with the assistance of multi-frequency fringe images. However, such sacrifice of time resolution using a large number of images seriously decreases the 3D measurement efficiency of FPP. Therefore, in order to measure dynamic scenes, researchers usually reduce the fringe patterns required for phase unwrapping, thus to improve the efficiency of per 3D reconstruction<sup>23,24</sup>. Ideally, the absolute depth is expected to be obtained by a single-shot fringe pattern.

The strategy of spatial frequency multiplexing is an effective single-frame 3D measurement technology<sup>25–32</sup>. The earliest idea of spatial frequency multiplexing was proposed by Takeda et al.<sup>25</sup>. By combining multiple sinusoids with different two-component spatial carrier frequencies into a fringe pattern, they developed single-shot spatial-frequency multiplexing for the FT technique<sup>15</sup> with the Chinese remainder theorem phase unwrapping technique (referred to as the G-S algorithm) to measure 3D objects with discontinuous and isolated surfaces. Another similar approach combined traditional multi-frame structure light pattern into a single composite pattern and can recover the depth data of moving or non-rigid object in real-time<sup>27</sup>. Although special fringe composite design is realized to separate the spectrum in the above work<sup>25,27</sup>, it is still unable to avoid spectrum aliasing entirely. Therefore, the resulting low phase quality around discontinuities and isolated regions of the phase map makes these methods unable to be applied in high-accuracy 3D measurement field. Liu et al.<sup>23</sup> proposed a

dual-frequency composite PS scheme, where high-frequency wrapped phase with high-quality obtained by the PS method is unwrapped according to a low-frequency phase through three look-up tables (LUTs) algorithm. Although higher-quality 3D measurement is allowed, the 5 fringe patterns required by the PS method increase the sensitivity to dynamic scenes.

In recent years, many studies have used deep learning as a tool to solve or improve the measurement efficiency issues in traditional FPP<sup>33–38</sup>. Feng et al.<sup>33,37</sup> proposed a fringe analysis approach using deep learning. By combining the physical model of the traditional PS method, high-quality phase information can be extracted from a single-frame fringe image. Shi et al.<sup>39</sup> proposed a deep learning-based fringe enhancement method to improve the phase imaging quality of the FT method. However, the above two methods can only achieve high-quality single-shot wrapped phase acquisition. In order to improve the efficiency of phase unwrapping, Yin et al.<sup>34</sup> applied deep learning to perform TPU. Although a large number of projected images required by traditional TPU are reduced, at least two phase maps with one frequency and high frequency are needed. Qian et al.<sup>36</sup> proposed a deep-learning-enabled geometric constraints and phase unwrapping method for single-shot absolute 3D shape measurement. Although robust phase unwrapping can be achieved on a single-frame projection, it is at the expense of increased hardware costs, where they used two cameras. Besides, they also combined deep learning and the color-coded technology to develop a single-shot absolute 3D shape measurement with color fringe projection profilometry<sup>38</sup>. However, this method will fail when measuring colored objects. In addition, there are also some end-to-end methods for linking fringe images and absolute depth information<sup>35,40,41</sup>. However, these methods may be difficult to obtain high-precision measurement results in practical applications, or may not guarantee stable fringe ambiguity removal.

Considering the traditional multi-frequency composite methods cannot guarantee single-frame high-accuracy 3D imaging, and inspired by the successful applications of deep learning in FPP, we propose a single-shot deep learning-based dual-frequency composite fringe projection profilometry, which can achieve spectrum-aliasing-free high-quality phase information retrieval, robust phase ambiguity removal and high-accuracy dynamic 3D shape measurement under the premise of only a single fringe projection image.

Different from the traditional end-to-end deep learning network that directly links the fringe image to absolute phase/depth<sup>35,40,41</sup>, we incorporate the concept of spatial frequency multiplexing in deep learning and design an unambiguous composite fringe image input to ensure that the networks have robust phase unwrapping performance. Besides, in order to provide the deep neural networks with the capability to overcome the serious spectrum aliasing problem that traditional spectrum separation technology cannot deal with, the fringe projection images without this problem are used to generate the aliasing-free labels. After proper training, the neural networks can directly recover robust absolute phase information through a composite fringe input image. Compared with traditional spatial frequency-multiplexing FT methods and deep learning techniques, our method can achieve higher quality phase information extraction as well as more robust phase unwrapping for objects with complex surface.

The remainder of this paper is organized as follows. In Section *Principle*, the basic principle of dual-frequency composite fringe projection profilometry, the acquisition of deep learning training data, the proposed deep learning-based composite fringe projection profilometry (DCFPP) method and the network architectures are introduced respectively. In Section *Experiments and results*, experimental verifications and comparison results are presented in detail. In the final Section *Conclusions*, we draw conclusions.

## Principle

### Single-shot dual-frequency composite fringe projection profilometry

In FPP, to achieve 3D measurement for high-speed dynamic scenes, it is necessary to minimize the number of projection frame per 3D reconstruction<sup>31</sup>. In this work, we aim at challenging the physical limit of the number of fringe patterns required for 3D imaging, and retrieval 3D data from a single frame.

Generally, phase unwrapping is a crucial step in FPP, which establishes the unique correspondence between different views, thereby allowing absolute 3D reconstruction. Meanwhile, it is also the operation that most affects 3D measurement efficiency<sup>42</sup>. Therefore, the key to achieve single-shot 3D shape measurement is to remove phase ambiguity through single-frame fringe image. One of the conventional single-shot phase unwrapping meth-

ods is the spatial phase unwrapping algorithm<sup>43</sup>, which can directly recover the absolute phase from only single wrapped phase map through the phase values of spatially adjacent pixels. However, this method cannot uniquely determine the period numbers for the cases of large discontinuities or spatially isolated surfaces. Inspired by the recent successful applications of deep learning techniques on FPP, we consider applying deep neural networks to perform single-shot absolute phase acquisition. Since the reliability of deep learning largely depends on the raw input information, if the input itself is ambiguous, the network is by no means always reliable<sup>44</sup>. Thus, in order to robustly eliminate the phase singularity, we must design an unambiguous input pattern. To this end, refer to the traditional temporal phase unwrapping (TPU) algorithms<sup>22</sup>, which project a series of fringe patterns with different frequencies and determine the pixel-wise fringe orders through the unique wrapped phase distribution in the time domain, we superimpose the time domain information of different frequencies into the spatial domain to generate a composite fringe pattern. As the phase unwrapping network input, the composite pattern should have sufficient capability to resist phase ambiguity, in other words, multi-frequency information separated from the composite fringe pattern should achieve the unambiguous phase unwrapping. In this work, we design a dual-frequency composite fringe coding strategy, where two vertical sinusoidal fringe patterns with different frequencies are added. The composed fringe pattern  $I_{cp}^p(x, y)$  (Fig. 1(a)) can be expressed as Eq. (1):

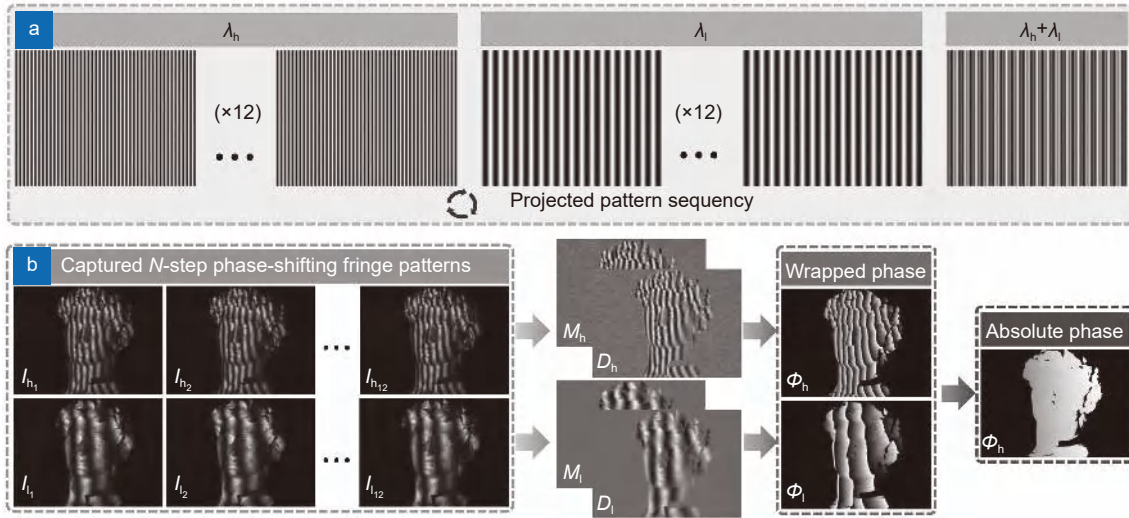
$$I_{cp}^p(x, y) = a^p(x, y) + b^p(x, y)[\cos(2\pi x/\lambda_h) + \cos(2\pi x/\lambda_l)], \quad (1)$$

where  $(x, y)$  is the image pixel coordinate,  $a^p$  denotes the mean intensity,  $b^p$  represents the amplitude,  $\lambda_h$  and  $\lambda_l$  are the wavelengths of the two hybrid sinusoidal fringe patterns with high and low frequencies, respectively. After illuminating the object with the composite fringe pattern  $I_{cp}^p$  through a digital projector, the intensities of the captured image can be expressed as:

$$I_{cp}(x, y) = A(x, y) + B(x, y)[\cos\Phi_h(x, y) + \cos\Phi_l(x, y)], \quad (2)$$

where  $A(x, y)$  is the average intensity relating to the pattern brightness and background illumination,  $B(x, y)$  is the intensity modulation relating to the pattern contrast and surface reflectivity. Besides, the captured composite fringe image contains two phase information of high and





**Fig. 1 | The process of generating training data.** (a) The projection mode includes dual-frequency 12-step phase-shifting fringe projection patterns and dual-frequency composite fringe pattern. (b) A set of captured images and the corresponding labels contain numerator term, denominator term, and absolute phase map.

low frequency, which are  $\Phi_h$  and  $\Phi_l$  respectively. In conventional fringe analysis methods, the extracted initial phase is the wrapped phase  $\varphi_h$  and  $\varphi_l$  with  $2\pi$  phase discontinuities due to the arctangent function<sup>45</sup>. Thus, phase unwrapping is required to remove the fringe ambiguities and correctly extract the absolute depth of the object<sup>42</sup>. The absolute phase maps corresponding to the wrapped phases of the hybrid sinusoidal fringe image can be represented as:

$$\begin{cases} \Phi_h(x, y) = \varphi_h(x, y) + 2\pi k_h(x, y) \\ \Phi_l(x, y) = \varphi_l(x, y) + 2\pi k_l(x, y) \end{cases}, \quad (3)$$

where  $k_h$  and  $k_l$  are the integer fringe order corresponding to the wrapped phases  $\varphi_h$  and  $\varphi_l$ ,  $k_h, k_l \in [0, K - 1]$ ,  $K$  denotes the number of the used fringes.

For two wrapped phase maps with different wavelengths, theoretically, we can use the traditional number-theoretical approach<sup>22</sup>, which is one of the TPU algorithms, to perform absolute phase unwrapping. The basic idea of this method relies on the fact that for suitable chosen fringe wavelengths  $\lambda_h$  and  $\lambda_l$ , their least common multiple  $LCM(\lambda_h, \lambda_l)$  determines the maximum range on the absolute phase axis within which the combination of wrapped phase values  $(\varphi_h, \varphi_l)$  is unique<sup>46,47</sup>. For a projection pattern with  $W \times H$  resolution, the selected two different wavelengths  $\lambda_h$  and  $\lambda_l$  should satisfy the following inequality to exclude phase ambiguity:

$$LCM(\lambda_h, \lambda_l) \geq W, \quad (4)$$

where  $LCM()$  represents the least common multiple function. That is to say, if  $LCM(\lambda_h, \lambda_l)$  (called the unam-

biguous range) can exceed the lateral resolution of the projected pattern, the phase ambiguity of the whole field can be eliminated. Specially, when the selected wavelengths are relatively prime, the  $LCM()$  function can be simplified to the multiplication of two wavelengths and the range of unambiguous phase becomes  $\lambda_h \lambda_l$ . After examining that the pairs of wrapped phase values are unique, the fringe orders  $k_h$  and  $k_l$  of the two phase maps can be determined.

Since the two sets of fringe patterns have different wavelengths ( $\lambda_h$  and  $\lambda_l$ ), their absolute phase map should have the following relationship:

$$\frac{\Phi_h(x, y)}{\Phi_l(x, y)} = \frac{\lambda_l}{\lambda_h}. \quad (5)$$

Combining Eqs. (3) and (5), we can get the following relation:

$$\frac{\lambda_l \varphi_h(x, y) - \lambda_h \varphi_l(x, y)}{2\pi} = k_l(x, y) \lambda_h - k_h(x, y) \lambda_l. \quad (6)$$

According to the number theory method, the fringe order pairs  $(k_h, k_l)$  can be determined by the pre-computed lookup table (LUT) which establishes the unique correspondence between the left side (called *Stair* function) and  $(k_h, k_l)$ :

$$(k_h, k_l) = LUT[Stair(x, y)]. \quad (7)$$

And the *Stair* function can be expressed as:

$$Stair(x, y) = \text{round}\left(\frac{\lambda_l \varphi_h(x, y) - \lambda_h \varphi_l(x, y)}{2\pi}\right), \quad (8)$$

where the  $\text{round}(\cdot)$  represents a rounding function.

Refer to the optimal dual-frequency selection approaches<sup>48–50</sup>, the high-frequency is as high as possible to allow high-accuracy measurement, while the low-frequency cannot be too low to ensure the stability of the phase unwrapping, and the relative minimum gap of the combined frequencies should be as large as possible to improve the fault tolerance rate of phase unwrapping, we finally select the frequency combination of a high-frequency fringe with wavelength of 19 pixels and a low-frequency fringe with wavelength of 51 pixels to synthesize a single-frame composite fringe pattern. It can perform unambiguous phase unwrapping of points within the range of 969 pixels, which means that the whole field of the projected pattern can carry out absolute phase unwrapping.

### Generate training data

The purpose of the data-driven-based deep learning network is to apply a large number of training data including the input values (the samples) and the ground-truth values (the targets/labels) to train a model, the output values predicted by which can be infinitely close to the ground-truth value. In this work, we aim at utilizing deep learning to predict high-quality the absolute phase map through a single fringe image.

In order to make the trained deep neural network overcome the problem of spectrum aliasing, we use dual-frequency 12-step phase-shifting fringe patterns (Fig. 1(a)) to generate high-quality, high-precision, and spectrum-aliasing-free network labels. In particular, the selected two frequencies/wavelength  $\lambda_h$  and  $\lambda_l$  are the same as the composite dual-frequency/dual-wavelength of the composite fringe pattern. The captured high-frequency and low-frequency sinusoidal fringe images can be expressed as:

$$\begin{aligned} I_{hn}^c(x, y) &= A_h(x, y) + B_h(x, y) \cos \left[ \varphi_h(x, y) + \frac{2\pi(n-1)}{12} \right] \\ I_{ln}^c(x, y) &= A_l(x, y) + B_l(x, y) \cos \left[ \varphi_l(x, y) + \frac{2\pi(n-1)}{12} \right], \end{aligned} \quad (9)$$

where  $I_{hn}^c$  and  $I_{ln}^c$  represent the intensity of the  $n$ th captured image with high and low frequencies respectively,  $n=1, 2, \dots, 12$ ,  $A_h$  and  $A_l$  are the average intensity of  $I_{hn}^c$  and  $I_{ln}^c$ ,  $B_h$  and  $B_l$  are the corresponding amplitude intensity maps. Then, the wrapped phase  $\varphi_h$  and  $\varphi_l$  can be obtained through the least-squares algorithm:

$$\begin{aligned} \varphi_h(x, y) &= \arctan \frac{\sum_{n=1}^{12} I_{hn}^c(x, y) \sin(2\pi(n-1)/12)}{\sum_{n=1}^{12} I_{hn}^c(x, y) \cos(2\pi(n-1)/12)} \\ &= \arctan \frac{M_h(x, y)}{D_h(x, y)} \\ \varphi_l(x, y) &= \arctan \frac{\sum_{n=1}^{12} I_{ln}^c(x, y) \sin(2\pi(n-1)/12)}{\sum_{n=1}^{12} I_{ln}^c(x, y) \cos(2\pi(n-1)/12)} \\ &= \arctan \frac{M_l(x, y)}{D_l(x, y)}, \end{aligned} \quad (10)$$

where set  $M_h$  and  $D_h$  as the numerator term and the denominator term of the arctangent function of wrapped phase  $\varphi_h$ , and set  $M_l$  and  $D_l$  as the numerator term and the denominator term of the arctangent function of wrapped phase  $\varphi_l$ . In order to eliminate the ambiguity of the high-frequency wrapped phase, we use the number-theoretical method (Eqs. (3), (7) and (8)) to unwrap the high-frequency wrapped phase into an absolute phase  $\Phi_h$ .

It should be emphasized that for the following three reasons, we do not adopt an end-to-end network structure that directly link the input fringe images to the output absolute phase/depth, but choose a network structure that predicting the numerator and denominator map of the wrapped phase arctangent function and a low-accuracy absolute phase map. 1) Since a single-frequency fringe image is insufficient to eliminate the phase/depth ambiguity in FPP while the multi-frequency fringe images can effectively remove this ambiguity through the TPU algorithm<sup>22</sup>, we use a single dual-frequency composite fringe image. As the network input, this composite image can not only retain the characteristics of a single frame projection, but also can be decomposed into two fringe images with different wavelengths/frequencies, which effectively removes the ambiguity of phase retrieval in essence and ensures that the absolute 3D shape measurement is not affected by any assumptions and prior knowledge, such as continuous surface, limited measurement range, geometric constraints. 2) Since the difficulty of establishing an accurate correspondence between the fringe intensity information and the high-accuracy absolute phase value, especially when the surface of the measured object contains sharp edges, discontinuities or large reflectivity variations, a simple input-output network structure only can usually obtain compromised imaging accuracy<sup>36</sup>. Based on this consideration, we use deep learning to predict a

rough absolute phase containing the correct fringe order information from the designed composite fringe image. 3) Our deep neural network is trained to predict the numerator and denominator of the arctangent function, to bypass the difficulties associated with reproducing abrupt  $2\pi$  phase wraps, and thus, obtain a high-quality phase information<sup>33</sup>.

Therefore, in this work, the output of the network we constructed includes the numerator and denominator used to calculate high-quality phase information, as well as the rough absolute phase that provides the fringe order information. The labels of the training data corresponding to these outputs are the numerator  $M_h$ , the denominator  $D_h$ , and the high-frequency absolute phase  $\Phi_h$ . Figure 1(a) is our projection mode, and Fig. 1(b) shows set of fringe images and the labels generated from these images.

In addition, in order to enhance the network learning ability, we set an appropriate modulation threshold to mask the invalid points of the training data maps by using the modulation function  $B(x, y)$  (Eq. (11)) and the Mask function (Eq. (12)):

$$B(x, y) = \frac{2}{N} \sqrt{M_h(x, y)^2 + D_h(x, y)^2}, \quad (11)$$

$$Mask(x, y) = \begin{cases} B(x, y), & B(x, y) \geq Thr \\ 0, & B(x, y) < Thr \end{cases}. \quad (12)$$

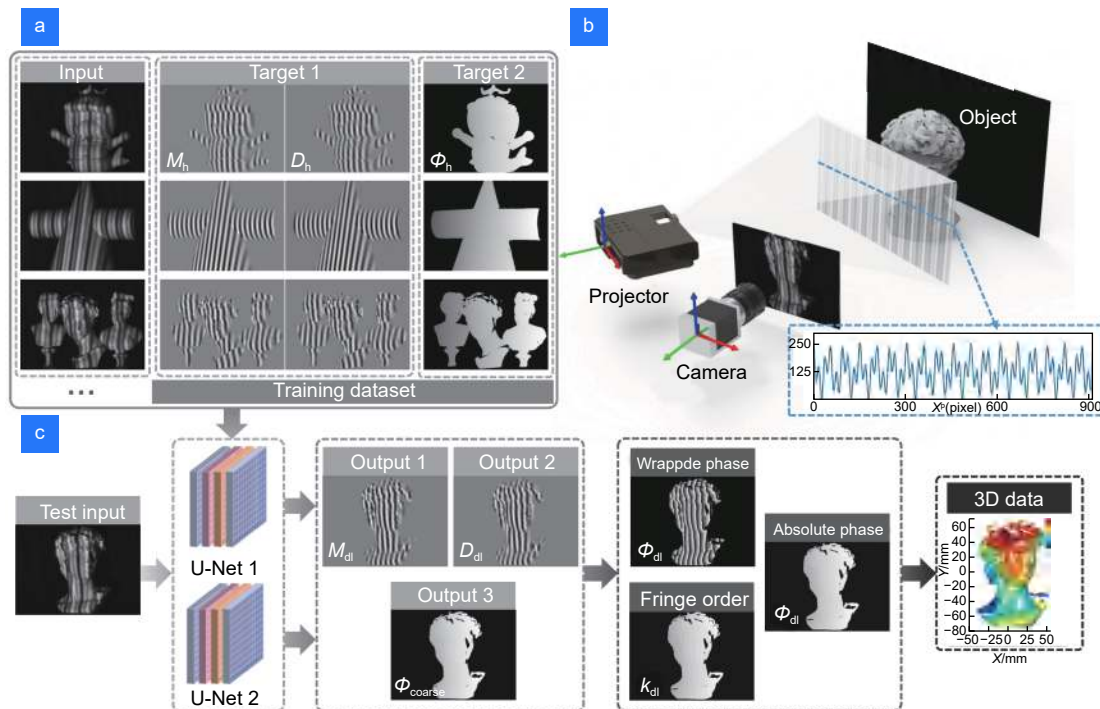
The value of threshold  $Thr$  is set to 8, which is suitable for most of our measurement scenarios in this work.

### Deep-learning-based single-shot composite fringe projection profilometry (DCFPP)

Our purpose is to propose a single-shot fringe projection profilometry using deep learning, which can robustly recover high-quality absolute phase information from a composite fringe image, thus to perform high-accuracy 3D reconstruction. The flowchart of our proposed approach (DCFPP) is shown in Fig. 2.

**Step 1:** Selection of wavelength combination strategy to generate a composite fringe pattern. We choose two wavelength  $\lambda_h = 19$  and  $\lambda_l = 51$  with the unambiguous range pixels that satisfy the Eq. (4):  $LCM(19, 51) = 969 > 912$  (mentioned in Section *Single-shot dual-frequency composite fringe projection profilometry*), to generate the composite fringe image, which is sufficient to overcome phase ambiguity, as the input of the deep convolution neural network. In order to cover the entire dynamic range of the projector  $[0, 255]$ , we set  $a^p + b^p = 255$  in Eq. (1), and the composite pattern along with its cross-section intensity profile for  $\lambda_h = 19$ ,  $\lambda_l = 51$ ,  $a^p = 130$  and  $b^p = 125$  is illustrated in Fig. 2(b).

**Step 2:** Preparation for network training data. According to the principle mentioned in Section *Generate*



**Fig. 2 | Flowchart of our proposed approach.** (a) Part of network training data sets. (b) Hardware system and the cross-section intensity distribution of the designed composite fringe pattern. (c) Test data and prediction results obtained by the training model.



training data, we use two sets of 12-step PS fringe images with the same dual-wavelength  $\lambda_h$  and  $\lambda_l$  of composite pattern to calculate the numerator terms  $M_h$  and the denominator terms  $D_h$  of spectrum-aliasing-free high-frequency wrapped phases and absolute phases as the ground-truth values of the neural network.

**Step 3:** Training data preprocessing. Before feeding the input data and targets into the neural network, data preprocessing is required. Such operation aims at making the raw data more amenable to neural networks, including vectorization and normalization. First, all inputs and targets in a neural network must be tensors of floating-point, this step called data vectorization, and in the experiment, we transform them into *float32* array of shape (*number of images*, 640, 480). Besides, it should be noted that all network inputs and targets need to be converted to a format compatible with TensorFlow. In general, it is unreliable to input relatively large values or heterogeneous data (that is, the size between the input and the target may differ ten or even a thousand times) into a neural network. Thus, data normalization is required. We divide the input images by 255 to convert the previous gray values from 0–255 range to 0–1 range.

**Step 4:** Training the neural network models. After preparing the training data sets, including a large number of unambiguous input data sets and corresponding high-quality ground-truth data sets as shown in Fig. 2(a), we put these specially designed inputs and outputs into the U-Net networks, so that the network will have a more powerful absolute phase retrieval capability. In terms of phase information acquisition, such data-driven-based training network can overcome the problem of poor imaging quality caused by frequency aliasing and has the high-quality phase information extraction function like the traditional PS algorithms; And in terms of phase unwrapping, it can directly recover absolute phase from a single fringe image, so as to reach the physical limit the number of the fringe image required for a single 3D reconstruction and maximize the efficiency of 3D imaging. As shown in Fig. 2(c), we construct two deep convolutional neural networks with the same structure except the final convolution layer, referred to as the U-Net1 and U-Net2, to perform phase information extraction and phase unwrapping tasks, separately. The specific reasons for choosing two networks instead of one will be explained in Section *Network architecture*. Plenty of raw composite fringe images  $I_{cp}^e(x, y)$  are fed into the two deep convolution neural networks, then U-Net1 will be

trained with the corresponding  $M_h$  and  $D_h$  as ground-truth to generate a phase acquisition model, and U-Net2 will be trained with the corresponding absolute phases  $\Phi_h$  as the ground-truth to obtain a phase unwrapping model.

**Step 5:** Prediction for absolute phase. The U-Net1 network is responsible for predicting the numerator terms  $M_{dl}$  and the denominator terms  $D_{dl}$  of a single-frame composite fringe image. Then, taken the output results  $M_{dl}$  and  $D_{dl}$  into the arctangent function, the wrapped phase distribution  $\varphi_{dl}(x, y)$  can be extracted:

$$\varphi_{dl}(x, y) = \arctan \frac{M_{dl}(x, y)}{D_{dl}(x, y)}. \quad (13)$$

Simultaneously, the U-Net2 predicts the “coarse” absolute phase  $\Phi_{coarse}(x, y)$  of the single-frame composite fringe image. Due to the environmental light, large surface reflectivity and discontinuities, it is hard to get high-quality phase information directly. Thus, feeding the wrapped phases  $\varphi_{dl}$  from U-Net1 and the “coarse” absolute phase  $\Phi_{coarse}$  from U-Net2 into Eq. (14) to obtain the fringe order  $k_{dl}(x, y)$ , the high-quality absolute phase  $\Phi_{dl}(x, y)$  can be recovered by Eq. (15).

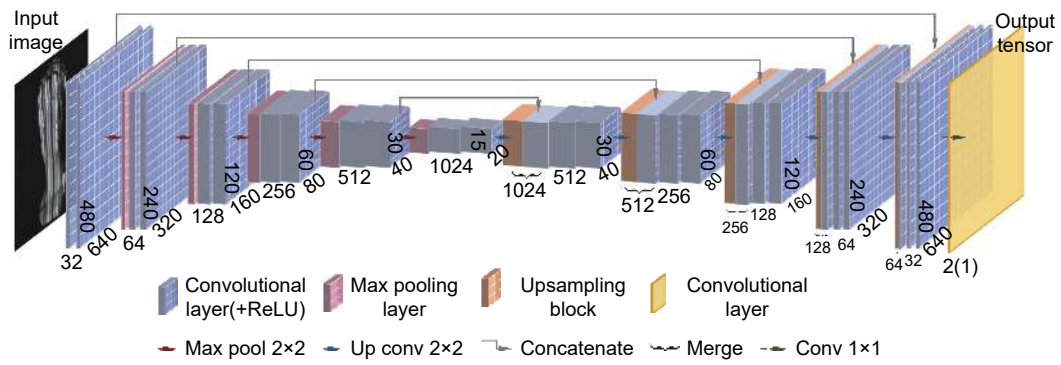
$$k_{dl}(x, y) = \text{Round} \left[ \frac{\Phi_{coarse}(x, y) - \varphi_{dl}(x, y)}{2\pi} \right], \quad (14)$$

$$\Phi_{dl}(x, y) = \varphi_{dl}(x, y) + 2\pi k_{dl}(x, y). \quad (15)$$

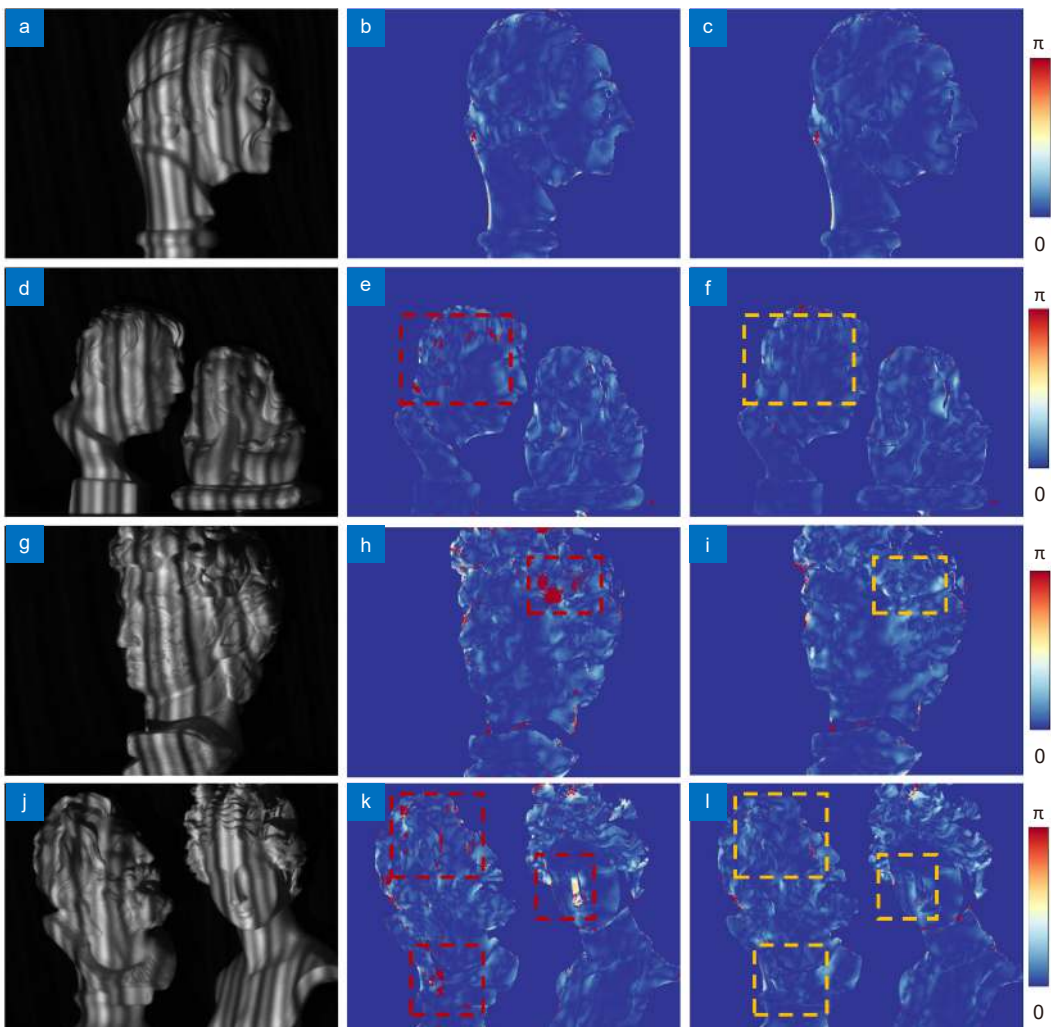
**Step 6:** 3D shape reconstruction. Finally, by utilizing the pre-calibrated parameters of the FPP system<sup>51–53</sup>, 3D information of the objects can be reconstructed.

### Network architecture

Next, we will further discuss the selection strategy and main architecture of the deep learning networks (Fig. 3). For the network architecture selection, we respectively use one U-Net network and two U-Net parallel networks to achieve phase retrieval and phase unwrapping. Figure 4 shows the comparison results of phase prediction using one U-Net network and two U-Net networks, from which we can draw conclusions: Using one U-Net network can predict an absolute phase of the object within the allowable error from different surface complexity, however, since this method directly outputs the absolute phase from the network, the absolute phase predict capability and quality of this end-to-end structure is worse than the result of two U-Net parallel networks. Thus, we use two U-Net parallel networks (marked as U-Net1 and U-Net2) to train the network models.



**Fig. 3 | The U-Net network architecture.**



**Fig. 4 | Comparison between one U-Net network and two U-Net networks (the proposed method).** (a, d, g, j) The raw composite fringe images from four different measurement scenes. (b, e, h, k) The absolute phase result error between deep-learning-predicted value and the ground-truth value by using one U-Net network. (c, f, i, l) The absolute phase result error between deep-learning-predicted value and the ground-truth value by using two U-Net networks.

Taken U-Net1 network as an example to reveal the internal structure of the constructed networks, the input tensors of size  $(H, W, 1)$  are successively processed by a stack of convolutional layers, pooling layers, upsampling

blocks, and concatenate layers. Each of convolutional layer represents a convolution operation, which extracts patches from its input feature map and applies the same transformation to all of these patches, producing an

output feature map. For each convolution layer, the kernel size is  $3 \times 3$  with convolution stride one and zero-padding, and it is activated by the rectified linear unit (ReLU) except for the last  $1 \times 1$  convolution layer. The output of the convolutional layer is a 3D tensor of shape  $(h, w, d)$ , where  $h \times w$  is the size of feature map input,  $d$  is the number of channels also representing filters that encode specific aspects of the input data. The number of channels is controlled by the first argument passed to the convolutional layers which is set to 32 in the proposed U-Net network. The role of pooling layer is to aggressively downsample feature maps, consists of extracting windows from the input feature maps and outputting the max value of each channel. Usually, max pooling layer is done with  $2 \times 2$  windows and stride 2, in order to down-sample the feature maps by a factor of 2. Thus, the size of composite image input  $H \times W$  tend to shrink as it gets deeper in the network. After downsampling the input by five times for better extraction, the upsampling block needs to match the raw input size. Then, copy the convolutional layer and merge it with the upsampling layer into a concatenate layer. Besides, the ultimate goal of the network is to achieve a model that can be generalized, that is, perform well on never-seen-before data. However, overfitting is the central obstacle. The processing of fighting overfitting is regularization. In this network, we use the Dropout which is one of the most effective and most commonly used regularization techniques for neural networks to fight overfitting. The loss function we select in this neural network is mean squared error (MSE), which is used to compare these predictions with the targets and generate a loss value. The optimizer chooses the Adam optimization scheme, which is used to update the network weights with the loss value and achieve better gradient propagation. Finally, the network of layers chained together maps input data to predictions.

## Experiments and results

To verify the performance of the proposed DCFPP method, we construct a monocular FPP system, which consists of a monochrome camera and a digital light processing (DLP) projector. The camera used in the system is a Basler acA640-750  $\mu\text{m}$  one equipped with an 8.5 mm Computar lens, which has 8-bit pixel depth and a maximum frame rate of 750 fps at a full  $640 \times 480$  resolution. The used projector is a LightCrafter 4500 one with a resolution of  $912 \times 1140$  and a projection pattern rate of 120

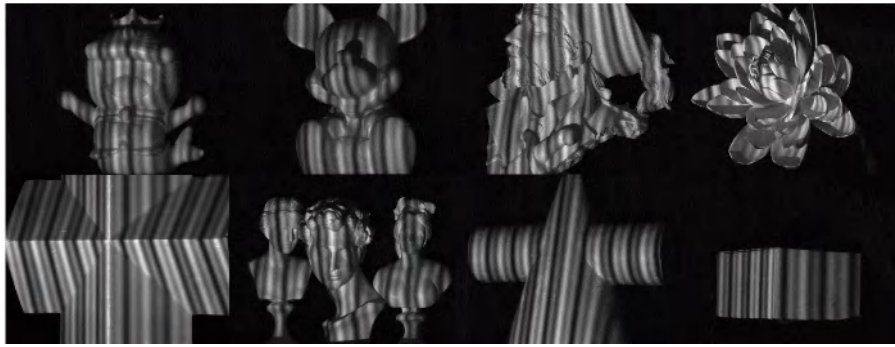
Hz with 8-bit. The field of view (FOV) of the system is about  $210 \text{ mm} \times 160 \text{ mm}$ , and the distance from the camera to the region of interest is approximately 400 mm. The network training experiment is computed on a desktop with Intel Core i7-7800X CPU and a NVIDIA GeForce GTX 1080 Ti GPU, and we use the Python deep learning framework Keras with the TensorFlow platform (developed by Google) to speed up the computation of the training model.

### Training the network model and testing the data

As mentioned earlier, a set of input and output data for training the network includes a dual-frequency composite fringe image  $I_{cp}^c$ , as well as the numerator  $M_h$ , denominator  $D_h$  and the absolute phase  $\Phi_h$ , where  $M_h$  and  $D_h$  are calculated by the 12-step PS method, and  $\Phi_h$  is obtained by the number theory method (refer to Section *Generate training data*). We project 25 fringe patterns each time in one projection period, including 24 PS sinusoidal fringe patterns and one dual-frequency composite pattern. Making full use of the three-color wheel projection mechanism of the DLP projector, three different images can be captured in red, green, and blue channels respectively, and combined into a single RGB image. Therefore, the system projection speed can actually be increased by three times. In this experiment, we set the pattern exposure to 148.5 ms and the pattern period to 150 ms. To maximize the generalization ability of the neural network, we need to obtain more training data. Thus, in the training experiment, a total of 1032 datasets from different scenes are collected including 800 training sets and 232 validation sets. Each of dataset contains one composite dual-wavelength ( $\lambda_h=19$ ,  $\lambda_l=51$ ) fringe image inputs, the ground-truth values numerator  $M_h$  and denominator  $D_h$ , and the ground-truth values the absolute phase  $\Phi$ . [Figure 5](#) shows some typical shooting scenes of the training datasets. The convolutional neural network is executed in 200 epochs, of which the mini-batch (used to compute a single gradient-descent update for the weights of the model) is 2, the initial model will be learned through the above process. In order to further optimize this model, network parameters and structure need to be adjusted. Due to data augmentation, the time for network training takes 8.3 hours on an NVIDIA graphics card.

We collect 60 scenes data that different from the training and the validation sets to test the accuracy of the model. The processing speed of our approach can reach





**Fig. 5 | Part of input training datasets.** The surface shapes contain single complex surface, geometric surface and discontinuous surface, and the materials include plaster, plastic, and paper.

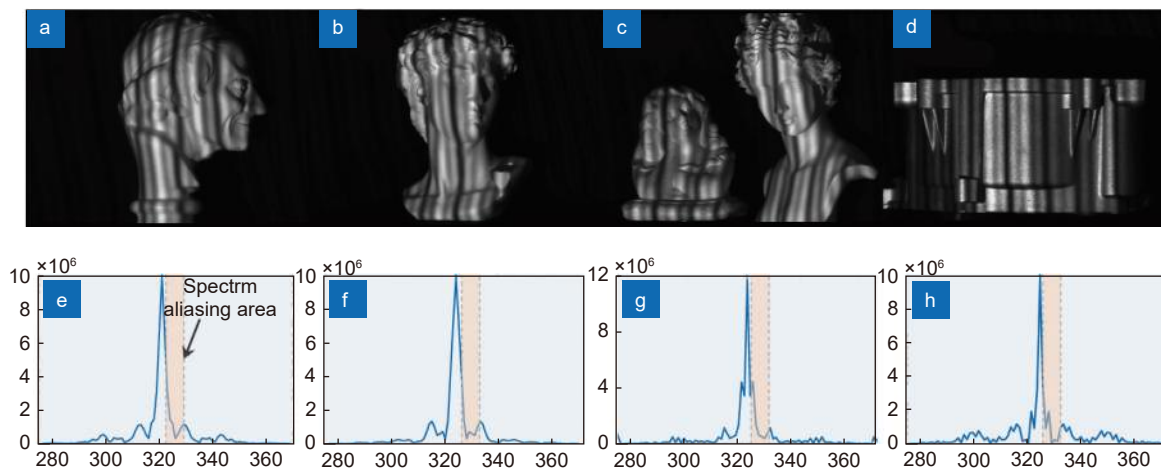
about 15 fps. We can put the captured and processed image data into the trained network model to retrieve the phase information of target object and complete the off-line 3D measurement. Although our system can only complete 3D measurement of complex objects and moving objects in an offline state, our single-frame imaging method provides basic support for real-time online processing.

It should be noted that since our network models are trained from the composite fringe images with dual wavelengths (19 and 51) and the fringe images with different wavelengths at the same position correspond to different fringe orders, the trained model is only valid for composite fringe images with wavelengths 19 and 51. However, as long as the selected frequency/wavelength combination meets the selection conditions mentioned in Section *Single-shot dual-frequency composite fringe projection profilometry* to eliminate fringe pattern ambiguity, the trained model on the composite fringe images with the selected frequency combination can also perform single-frame measurement on the composite image with the same frequency combination.

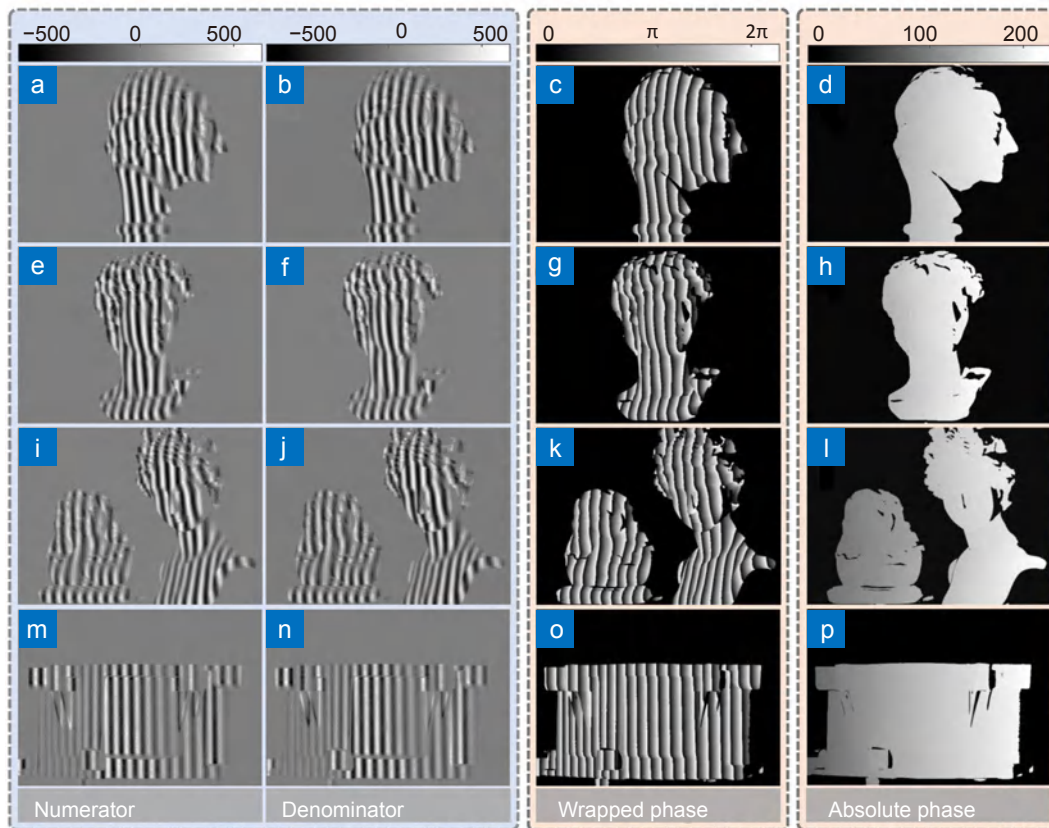
### Qualitative evaluation

To test the proposed approach, we conducted static experiments and dynamic experiments, respectively.

We first measured four static scenarios that our network has never seen before, including a Voltaire statue plaster model, a David plaster model, little girl and women combination models, and a metal workpiece. These scenes involve a single object with continuous complex surface shapes, a combination of multiple objects with isolated surfaces, and workpieces with different surface reflectivity materials. Figure 6(a–d) show the captured composite fringe images  $I_{cp}(x, y)$  with  $\lambda_h=19$  pixel,  $\lambda_l=51$  pixel which are the input of the constructed neural network, and Fig. 6(e–h) are corresponding cross-sections of their spectrum intensities, from which we can see that the spectrum aliasing is so serious that it is difficult to separate and extract effective dual-frequency information through applying the filter window in the frequency domain. The U-Net1 network model predicts the numerator  $M_{dl}$  and denominator  $D_{dl}$  results for each input image, as shown in the first two columns of Fig. 7.



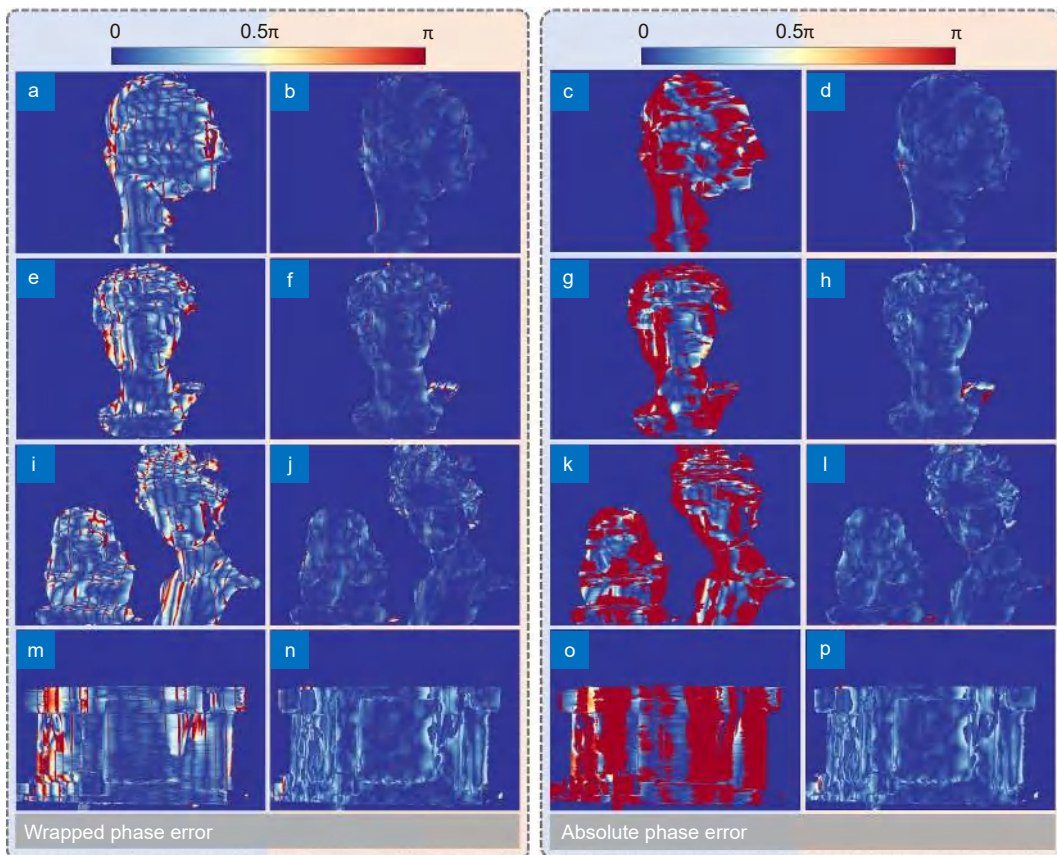
**Fig. 6 | Four sets of network test inputs in static scenarios (a–d) and their corresponding spectral cross-sectional intensity distributions (e–h).**



**Fig. 7 | The prediction results of our proposed method (DCFPP) in four static scenarios. (a, e, i, m) The numerator  $M_{dl}$  results. (b, f, j, n) The denominator  $D_{dl}$  results. (c, g, k, o) The wrapped phase  $\phi_{dl}$  results. (d, h, l, p) The high-quality absolute phase maps  $\Phi_{dl}$ .**

These intermediate results are then fed into the arctangent function (Eq. (13)) to calculate the phase distribution  $\phi_{dl}$ , as shown in the third column of Fig. 7. The U-Net2 training model is responsible for outputting a coarse absolute phase map, and then the high-quality absolute phase  $\Phi_{dl}$  can be calculated through Eq. (14) and Eq. (15), as shown in the last column of Fig. 7. We also compare our proposed DCFPP with the traditional dual-frequency composite FT method in the above four static scenes. Since the 3D information is obtained from the phase data, the 3D measurement accuracy can be reflected by the accuracy of the phase data. In the experiment, we directly perform error analysis on the recovered phase information of the objects. Taking the wrapped phase maps calculated by the 12-step PS method and the absolute phase generated by the traditional number theory method as the ground-truth values, the high-frequency phase errors of our approach and traditional method are shown in Fig. 8, where the first column (Fig. 8(a, e, i, m)) and the third column (Fig. 8(c, g, k, o)) are the errors of traditional method, the second columns (Fig. 8(b, f, j, n)) and the last columns (Fig. 8(d, h, l, p)) show the phase error results of the proposed DCFPP. It

can be seen that, compared with the traditional method, our method can significantly improve the performance of phase extraction and phase unwrapping from a single fringe image. Due to frequency spectrum aliasing between fundamental frequency (the low-frequency) and zero frequency (refer to Fig. 6), the foundational spectrum cannot be filtered out exactly, and the inexact phase information will lead to poor phase imaging quality, thus causing serious phase unwrapping errors. By contrast, our approach eliminates the need to analyze the image spectrum and directly retrieves the high-quality aliasing-free absolute phase by the unambiguous composite fringe input. The comparison results of traditional dual-frequency composite FT methods proved that DCFPP can significantly improve the performance of single fringe phase retrieval and phase unwrapping. For the quantitative analysis of the method, we calculate the mean absolute error (MAE) of the wrapped phase and absolute phase from these four scenes, as shown in the Table 1. For the traditional method, the low-quality wrapped phase leads to serious phase unwrapping errors, so that the calculated absolute phase has larger error values. For the DCFPP, the reason why the absolute phase



**Fig. 8 | Phase error comparison results of traditional dual-frequency composite FT method and the proposed DCFPP method. (a, e, i, m) The wrapped phase error calculated by traditional method. (b, f, j, n) The wrapped phase error predicted by the DCFPP. (c, g, k, o) The absolute phase error of traditional method. (d, h, l, p) The absolute phase error of the DCFPP.**

**Table 1 | MAE of wrapped phase and absolute phase of the traditional dual-frequency composite FT method and the proposed DCFPP method (noted that the “FT method” mentioned in the table refers to the traditional dual-frequency composite FT method).**

MAE (rad)	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
	FT	DCFPP	FT	DCFPP	FT	DCFPP	FT	DCFPP
Wrapped phase	0.259	0.063	0.255	0.089	0.430	0.125	0.923	0.092
Absolute phase	3.314	0.034	2.185	0.071	4.333	0.083	5.707	0.055

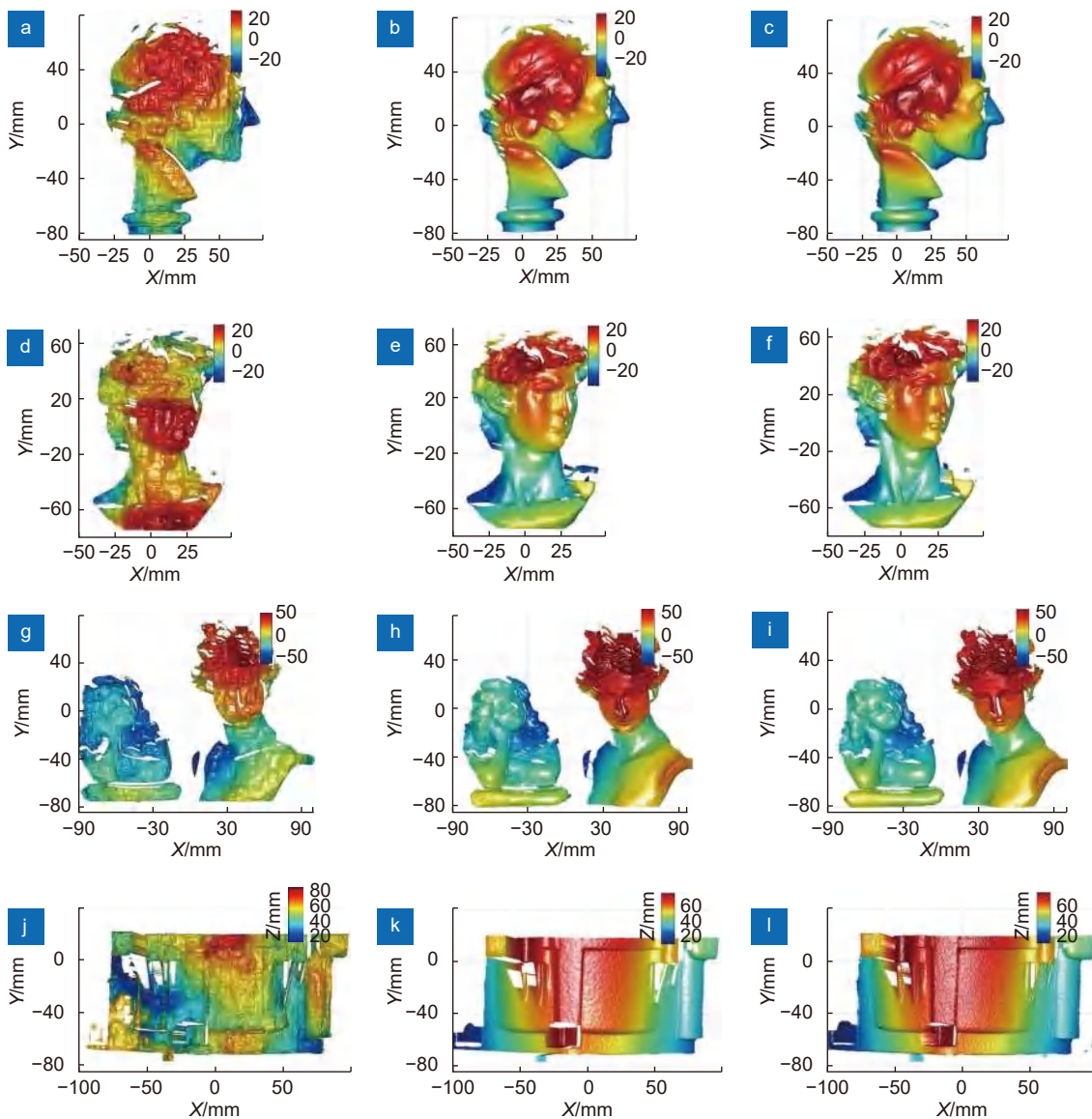
MAE is smaller than the wrapped phase MAE is that  $2\pi$  jump area of the predicted wrapped phase and the cutoff area of the label wrapped phase do not completely coincide. The errors close to  $2\pi$  caused by these very slight misalignments (often only one pixel apart) will be eliminated after phase unwrapping.

Furthermore, through phase-height mapping and the calibration parameters of the camera-projector FPP system, the 3D reconstruction results of the above four scenarios can be obtained. Figure 9 shows the comparison results of the three methods: the end-to-end network, the DCFPP method and the 12-step PS with number-theoretic method (the ground-truth generation method). Fig. 9(a, d, g, j) are the results of the method<sup>40</sup>. In their

end-to-end deep neural network, they use one single-frequency fringe pattern as input and directly output the corresponding depth map. From which we can see that the 3D reconstruction results of the end-to-end network are poor. The low accuracy results further verify the theoretical analysis in Section *Generate training data* that a single-frequency fringe image is insufficient to eliminate the phase/depth ambiguity. Our proposed method (proposed method (Fig. 9(b, e, h, k)) using only one composite image can yield the imaging quality comparable to that obtained by the traditional 12-step PS with number-theoretic method (Fig. 9(c, f, i, l)).

In the second experiment, we measure an object in constant motion to validate the capability of the





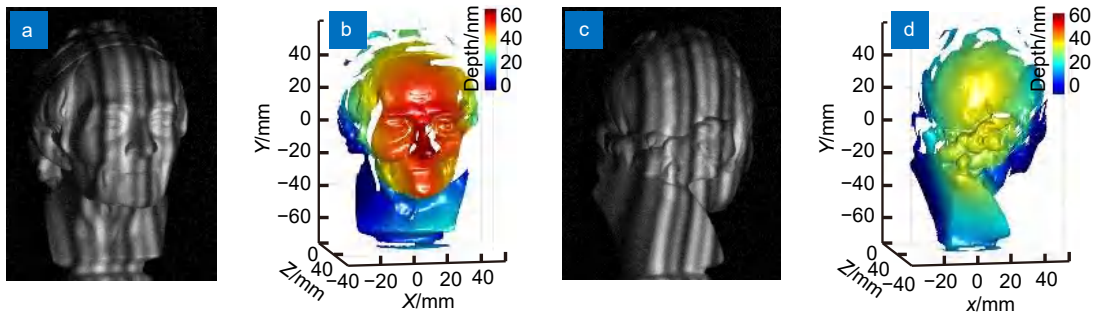
**Fig. 9 | 3D reconstruction results of end-to-end network, the proposed DCFPP and 12-step PS with number-theoretic method in four measurement scenes. (a, d, g, j) 3D reconstruction results of the end-to-end network. (b, e, h, k) 3D reconstruction results of DCFPP. (c, f, i, l) 3D reconstruction results obtained by 12-step PS with number-theoretic method (ground-truth).**

proposed DCFPP approach in the dynamic scenarios. Figure 10 shows the 3D reconstruction results of a rotating Voltaire plaster statue model using DCFPP method in selected moments. During the measurement, a single-frame composite fringe pattern is continuously projected on the surface of the object, and a monochrome camera simultaneously captures the gray fringe image of each frame. In conventional phase-shifting profilometry, motion introduces additional phase shift, which breaks the basic assumptions of phase-shifting profilometry and produces motion ripples in the reconstructed result<sup>6</sup>, while our method uses only one image, which fundamentally overcomes the influence of motion, so there are

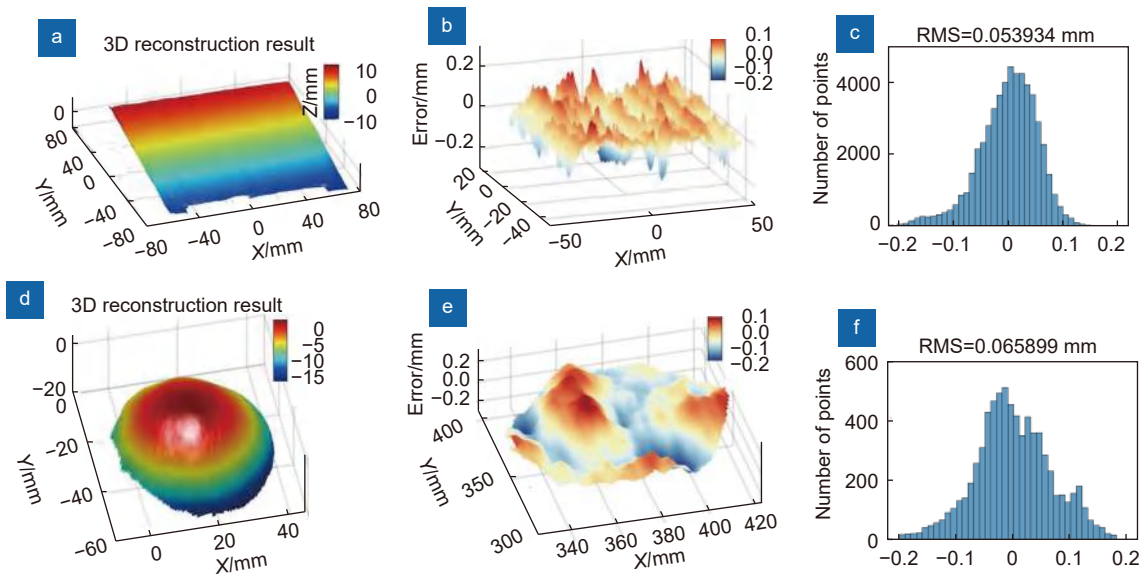
no motion ripples. The whole measurement process of the rotating plaster statue is shown in Fig. 10 (Multimedia view). It can be seen that due to the single-frame nature of DCFPP, the motion-induced artifacts can be avoided in the reconstruction process.

### Quantitative evaluation

At last, to quantitatively evaluate the 3D reconstruction precision of the proposed method, we respectively measured a standard ceramic plate and a standard ceramic sphere with radii  $R = 25.4$  mm. The precision analysis results are shown in Fig. 11, where Fig. 11(a) and 11(d) are the 3D reconstruction results calculated by our



**Fig. 10 | Measurement results of a dynamic scene.** (a, c) The captured composite fringe images at two different moments. (b, d) The corresponding 3D results reconstructed by DCFPP. (Multimedia view: see Supplementary information [Visualization 1](#) for the whole measurement process of the rotating plaster statue)



**Fig. 11 | Precision analysis of standard ceramic plate and a standard ceramic sphere.** (a, d) 3D reconstruction results by DCFPP. (b, e) Error distribution. (c, f) RMS error.

method, [Fig. 11\(c\)](#) and [11\(f\)](#) respectively show the distribution of errors of the plate and the standard ceramic sphere. Specifically, the ground-truth values of both of them are generated by fitting a plane or a sphere using 3D reconstruction data. The root mean square (RMS) error of them are 0.054 mm and 0.065 mm, respectively. This experiment proves that our method can achieve high-quality 3D measurement just using a single fringe image.

## Conclusions

In this study, we present a deep learning-based single-shot 3D measurement technology, which is able to recover the absolute 3D information of complex scenes with large surface discontinuities or isolated objects while projecting only a single composite fringe pattern. By combining the deep learning network with the physical model of FPP, we take a well-designed unambiguous

composite fringe pattern as input, and the phase information without spectrum aliasing as the ground-truth to drive the neural networks to achieve robust, high-quality single-shot absolute phase recovery. Compared with the traditional spatial frequency multiplexing FT method, our DCFPP approach avoids the resulting poor 3D measurement accuracy caused by spectrum aliasing, whose imaging quality is comparable to the performance of traditional 12-step PS method which uses more than 12 fringe patterns.

This paper aims to show that deep learning is an efficient tool for synthesizing temporal and spatial information. It can avoid the spectrum aliasing problem of traditional single-frame phase measurement methods, and assist in achieving robust phase unwrapping for complex scenes with large surface discontinuities or isolated objects from a single fringe image. However, due to the intensity containing varying reflectance that cannot be

correctly mapped to the absolute phase distribution with high-accuracy, it is still difficult to retrieve high-quality absolute phase information in an end-to-end deep learning-based network. In the future, we will explore more advanced network structures and integrate more suitable physical models into deep learning networks to realize higher-speed, higher-accuracy and more robust 3D shape measurement through fewer neural networks or even an end-to-end manner.

## References

- Gorthi SS, Rastogi P. Fringe projection techniques: whither we are. *Opt Lasers Eng* **48**, 133–140 (2010).
- Zhang ZH, Towers CE, Towers DP. Time efficient color fringe projection system for 3D shape and color using optimum 3-frequency selection. *Opt Express* **14**, 6444–6455 (2006).
- Su XY, Zhang QC. Dynamic 3-D shape measurement method: a review. *Opt Lasers Eng* **48**, 191–204 (2010).
- Tao TY, Chen Q, Da J, Feng SJ, Hu Y et al. Real-time 3-D shape measurement with composite phase-shifting fringes and multi-view system. *Opt Express* **24**, 20253–20269 (2016).
- Feng SJ, Zhang L, Zuo C, Tao TY, Chen Q et al. High dynamic range 3D measurements with fringe projection profilometry: a review. *Meas Sci Technol* **29**, 122001 (2018).
- Feng SJ, Zuo C, Tao TY, Hu Y, Zhang ML et al. Robust dynamic 3-D measurements with motion-compensated phase-shifting profilometry. *Opt Lasers Eng* **103**, 127–138 (2018).
- Pan B, Xie HM, Wang ZY, Qian KM, Wang ZY. Study on subset size selection in digital image correlation for speckle patterns. *Opt Express* **16**, 7037–7048 (2008).
- Hu Y, Chen Q, Feng SJ, Zuo C. Microscopic fringe projection profilometry: a review. *Opt Lasers Eng* **135** (2020).
- Tao TY, Chen Q, Feng SJ, Qian JM, Hu Y et al. High-speed real-time 3D shape measurement based on adaptive depth constraint. *Opt Express* **26**, 22440–22456 (2018).
- Qian JM, Feng SJ, Tao TY, Hu Y, Liu K et al. High-resolution real-time 360° 3D model reconstruction of a handheld object with fringe projection profilometry. *Opt Lett* **44**, 5751–5754 (2019).
- Qian JM, Feng SJ, Xu MZ, Tao TY, Shang YH et al. High-resolution real-time 360° 3D surface defect inspection with fringe projection profilometry. *Opt Lasers Eng* **137**, 106382 (2021).
- Zuo C, Chen Q, Gu GH, Feng SJ, Feng FXY et al. High-speed three-dimensional shape measurement for dynamic scenes using bi-frequency tripolar pulse-width-modulation fringe projection. *Opt Lasers Eng* **51**, 953–960 (2013).
- Heist S, Lutzke P, Schmidt I, Dietrich P, Kühmstedt P et al. High-speed three-dimensional shape measurement using GOBO projection. *Opt Lasers Eng* **87**, 90–96 (2016).
- Heist S, Kühmstedt P, Tünnermann A, Notni G. Theoretical considerations on aperiodic sinusoidal fringes in comparison to phase-shifted sinusoidal fringes for high-speed three-dimensional shape measurement. *Appl Opt* **54**, 10541–10551 (2015).
- Takeda M, Mutoh K. Fourier transform profilometry for the automatic measurement of 3-D object shapes. *Appl Opt* **22**, 3977–3982 (1983).
- Su XY, Chen WJ. Fourier transform profilometry: a review. *Opt Lasers Eng* **35**, 263–284 (2001).
- Kemao Q. Two-dimensional windowed Fourier transform for fringe pattern analysis: principles, applications and implementations. *Opt Lasers Eng* **45**, 304–317 (2007).
- Huang L, Kemao Q, Pan B, Asundi AK. Comparison of Fourier transform, windowed Fourier transform, and wavelet transform methods for phase extraction from a single fringe pattern in fringe projection profilometry. *Opt Lasers Eng* **48**, 141–148 (2010).
- Zhang ZH, Jing Z, Wang ZH, Kuang DF. Comparison of Fourier transform, windowed Fourier transform, and wavelet transform methods for phase calculation at discontinuities in fringe projection profilometry. *Opt Lasers Eng* **50**, 1152–1160 (2012).
- Zuo C, Feng SJ, Huang L, Tao TY, Yin W et al. Phase shifting algorithms for fringe projection profilometry: a review. *Opt Lasers Eng* **109**, 23–59 (2018).
- Pan B, Kemao Q, Huang L, Asundi A. Phase error analysis and compensation for nonsinusoidal waveforms in phase-shifting digital fringe projection profilometry. *Opt Lett* **34**, 416–418 (2009).
- Zuo C, Huang L, Zhang ML, Chen Q, Asundi A. Temporal phase unwrapping algorithms for fringe projection profilometry: a comparative review. *Opt Lasers Eng* **85**, 84–103 (2016).
- Liu K, Wang YC, Lau DL, Hao Q, Hassebrook LG. Dual-frequency pattern scheme for high-speed 3-D shape measurement. *Opt Express* **18**, 5229–5244 (2010).
- Zuo C, Tao TY, Feng SJ, Huang L, Asundi A et al. Micro Fourier transform profilometry ( $\mu$ ftp): 3D shape measurement at 10,000 frames per second. *Opt Lasers Eng* **102**, 70–91 (2018).
- Takeda M, Gu Q, Kinoshita M, Takai H, Takahashi Y. Frequency-multiplex Fourier-transform profilometry: a single-shot three-dimensional shape measurement of objects with large height discontinuities and/or surface isolations. *Appl Opt* **36**, 5347–5354 (1997).
- Zhong JG, Zhang YL. Absolute phase-measurement technique based on number theory in multifrequency grating projection profilometry. *Appl Opt* **40**, 492–500 (2001).
- Guan C, Hassebrook LG, Lau DL. Composite structured light pattern for three-dimensional video. *Opt Express* **11**, 406–417 (2003).
- Sansoni G, Redaelli E. A 3D vision system based on one-shot projection and phase demodulation for fast profilometry. *Meas Sci Technol* **16**, 1109–1118 (2005).
- Yue HM, Su XY, Liu YZ. Fourier transform profilometry based on composite structured light pattern. *Opt Laser Technol* **39**, 1170–1175 (2007).
- Chen WJ, Su XY, Cao Y, Xiang LQ, Zhang QC. Fourier transform profilometry based on a fringe pattern with two frequency components. *Optik-Int J Light Electron Opt* **119**, 57–62 (2008).
- Zhang ZH. Review of single-shot 3D shape measurement by phase calculation-based fringe projection techniques. *Opt Lasers Eng* **50**, 1097–1106 (2012).
- García-Isáis C, Ochoa NA. One shot profilometry using a composite fringe pattern. *Opt Lasers Eng* **53**, 25–30 (2014).
- Feng SJ, Chen Q, Gu GH, Tao TY, Zhang L et al. Fringe pattern analysis using deep learning. *Adv Photonics* **1**, 025001 (2019).
- Yin W, Chen Q, Feng SJ, Tao TY, Huang L et al. Temporal phase unwrapping using deep learning. *Sci Rep* **9**, 20175 (2019).
- van der Jeught S, Dirckx JJJ. Deep neural networks for single



- shot structured light profilometry. *Opt Express* **27**, 17091–17101 (2019).
36. Qian JM, Feng SJ, Tao TY, Hu Y, Li YX et al. Deep-learning-enabled geometric constraints and phase unwrapping for single-shot absolute 3D shape measurement. *APL Photonics* **5**, 046105 (2020).
  37. Feng SJ, Zuo C, Yin W, Gu GH, Chen Q. Micro deep learning profilometry for high-speed 3D surface imaging. *Opt Lasers Eng* **121**, 416–427 (2019).
  38. Qian JM, Feng SJ, Li YX, Tao TY, Han J et al. Single-shot absolute 3D shape measurement with deep-learning-based color fringe projection profilometry. *Opt Lett* **45**, 1842–1845 (2020).
  39. Shi JS, Zhu XJ, Wang HY, Song LM, Guo QH. Label enhanced and patch based deep learning for phase retrieval from single frame fringe pattern in fringe projection 3D measurement. *Opt Express* **27**, 28929–28943 (2019).
  40. Nguyen H, Wang YZ, Wang ZY. Single-shot 3D shape reconstruction using structured light and deep convolutional neural networks. *Sensors* **20**, 3718 (2020).
  41. Zheng Y, Wang SD, Li Q, Li BW. Fringe projection profilometry by conducting deep learning from its digital twin. *Opt Express* **28**, 36568–36583 (2020).
  42. Zhang S. Absolute phase retrieval methods for digital fringe projection profilometry: a review. *Opt Lasers Eng* **107**, 28–37 (2018).
  43. Ghiglia DC, Pritt MD. *Two-Dimensional Phase Unwrapping: Theory, Algorithms, and Software* (Wiley-Interscience, New York, 1998).
  44. Chollet F. *Deep Learning with Python* (Manning Publications, Shelter Island, 2018).
  45. Qian JM, Tao TX, Feng SJ, Chen Q, Zuo C. Motion-artifact-free dynamic 3D shape measurement with hybrid Fourier-transform phase-shifting profilometry. *Opt Express* **27**, 2713–2731 (2019).
  46. Lilienblum E, Michaelis B. Optical 3D surface reconstruction by a multi-period phase shift method. *J Comput* **2**, 73–83 (2007).
  47. Pribanić T, Mrvoš S, Salvi J. Efficient multiple phase shift patterns for dense 3D acquisition in structured light scanning. *Im-age Vis Comput* **28**, 1255–1266 (2010).
  48. Ding Y, Xi JT, Yu YG, Chicharo J. Recovering the absolute phase maps of two fringe patterns with selected frequencies. *Opt Lett* **36**, 2518–2520 (2011).
  49. Ding Y, Xi JT, Yu YG, Cheng WQ, Wang S et al. Frequency selection in absolute phase maps recovery with two frequency projection fringes. *Opt Express* **20**, 13238–13251 (2012).
  50. Yin W, Zuo C, Feng SJ, Tao TY, Hu Y et al. High-speed three-dimensional shape measurement using geometry-constraint-based number-theoretical phase unwrapping. *Opt Lasers Eng* **115**, 21–31 (2019).
  51. Zhang Z. A flexible new technique for camera calibration. *IEEE Trans Pattern Anal Mach Intell* **22**, 1330–1334 (2000).
  52. Zhang S, Huang PS. Novel method for structured light system calibration. *Opt Eng* **45**, 083601 (2006).
  53. Huang L, Zhang QC, Asundi A. Camera calibration with active phase target: improvement on feature detection and optimization. *Opt Lett* **38**, 1446–1448 (2013).

## Acknowledgements

This work was supported by National Natural Science Foundation of China (62075096, 62005121, U21B2033), Leading Technology of Jiangsu Basic Research Plan (BK20192003), “333 Engineering” Research Project of Jiangsu Province (BRA2016407), Jiangsu Provincial “One belt and one road” innovation cooperation project (BZ2020007), Fundamental Research Funds for the Central Universities (30921011208, 30919011222, 30920032101), Post-graduate Research & Practice Innovation Program of Jiangsu Province (KYCX21\_0273), and Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense (JSGP202105).

## Competing interests

The authors declare no competing financial interests.

## Supplementary information

Supplementary information for this paper is available at <https://doi.org/10.29026/oea.2022.210021>

RESEARCH

Open Access



# Deep-learning-enabled temporally super-resolved multiplexed fringe projection profilometry: high-speed kHz 3D imaging with low-speed camera

Wenwu Chen<sup>1,2,3</sup>, Shijie Feng<sup>1,2,3\*</sup>, Wei Yin<sup>1,2,3</sup>, Yixuan Li<sup>1,2,3</sup>, Jiaming Qian<sup>1,2,3</sup>, Qian Chen<sup>3\*</sup> and Chao Zuo<sup>1,2,3\*</sup>

\*Correspondence:  
shijiefeng@njjust.edu.cn;  
chenqian@njjust.edu.cn;  
zuochao@njjust.edu.cn

<sup>1</sup> Smart Computational Imaging Laboratory (SCILab), School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, Jiangsu Province, China

<sup>2</sup> Smart Computational Imaging Research Institute (SCIRI) of Nanjing University of Science and Technology, Nanjing 210019, Jiangsu Province, China

<sup>3</sup> Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing University of Science and Technology, Nanjing 210094, Jiangsu Province, China

## Abstract

Recent advances in imaging sensors and digital light projection technology have facilitated rapid progress in 3D optical sensing, enabling 3D surfaces of complex-shaped objects to be captured with high resolution and accuracy. Nevertheless, due to the inherent synchronous pattern projection and image acquisition mechanism, the temporal resolution of conventional structured light or fringe projection profilometry (FPP) based 3D imaging methods is still limited to the native detector frame rates. In this work, we demonstrate a new 3D imaging method, termed deep-learning-enabled multiplexed FPP (DLMFPP), that allows to achieve high-resolution and high-speed 3D imaging at near-one-order of magnitude-higher 3D frame rate with conventional low-speed cameras. By encoding temporal information in one multiplexed fringe pattern, DLMFPP harnesses deep neural networks embedded with Fourier transform, phase-shifting and ensemble learning to decompose the pattern and analyze separate fringes, furnishing a high signal-to-noise ratio and a ready-to-implement solution over conventional computational imaging techniques. We demonstrate this method by measuring different types of transient scenes, including rotating fan blades and bullet fired from a toy gun, at kHz using cameras of around 100 Hz. Experimental results establish that DLMFPP allows slow-scan cameras with their known advantages in terms of cost and spatial resolution to be used for high-speed 3D imaging tasks.

**Keywords:** 3D imaging, Fringe projection profilometry, Multiplex, Deep learning, Temporal super-resolution

## Introduction

Over recent decades, significant advancements in optoelectronics have ignited interests in capturing and documenting instantaneous phenomena. The ability to capture immediate three-dimensional (3D) geometric changes in objects provides invaluable insights into fast events, crucial for diverse fields such as industrial inspection [1], biomedicine [2], and solid mechanics [3]. Among the array of 3D imaging techniques, fringe projection profilometry (FPP) [4] is one of the most promising modalities due to its capacity for high-accuracy and full-field 3D measurements.

To enhance the speed of FPP, efforts have been made to improve the speed of measurement system. Binary defocusing techniques, for instance, have emerged to increase the projection speed of digital light processing (DLP) systems [5, 6]. By projecting binary fringes (1-bit) instead of grayscale patterns (8-bit) in a defocused manner, these techniques have demonstrated the capability to increase projection speeds from a hundred frames per second (fps) to thousands or even tens of thousands fps. Additionally, custom projectors utilizing rotating wheels [7] or LED arrays [8, 9] have also been developed to achieve high-speed pattern projection.

Although system speed has improved, motion can still compromise 3D measurements if numerous patterns are required for dynamic 3D reconstruction [10]. Therefore, researchers have presented methods using a small number of patterns, such as dual-frequency phase-shifting (PS) [11], bi-frequency PS [12], 2+2 PS [9], composite PS [13], and micro Fourier transform profilometry [14]. These approaches utilize each projected pattern for both wrapped phase calculation and absolute phase unwrapping, effectively reducing the number of patterns. Fourier transform profilometry (FTP) employs a single fringe pattern for 3D reconstruction but struggles with complex shapes due to spectrum aliasing [15]. Recent advancements in artificial intelligence have introduced deep neural networks (DNNs) [16, 17] to optical metrology [18]. Properly trained DNNs can retrieve phase [19] and 3D coordinates [20–23] using a single fringe pattern accurately for complex objects, pushing the 3D measurement speed to the upper limit that is the camera's speed for capturing two-dimensional (2D) images.

However, enhancing the camera's speed often comes at a cost, such as the decrease in pixel resolution and the signal-to-noise ratio (SNR) of captured images. Although high-speed cameras capture images at a high frame rate without reducing the resolution, the cost of the system will sharply increase. Moreover, the speed of 3D imaging is inherently hindered by the rate at which 2D images can be captured and processed. Therefore, we are facing a big challenge that is *"can affordable low-speed cameras be used to replace high-speed cameras and achieve high-speed 3D imaging without compromising image resolution"*.

In recent years, we have witnessed the rapid progress of deep learning in computational imaging [24]. Meanwhile, the refresh rate of digital micro-mirror devices (DMDs) has significantly increased, reaching tens of thousands fps, while at an affordable price. This motivated us to combine computational imaging and deep learning to encode temporal information in space and break through the physical limits of camera hardware speed. Inspired by the concept of holographic multiplexing [25], for the first time to our knowledge, we introduce a novel approach termed deep-learning-enabled multiplexed FPP (DLMFPP). DLMFPP enables high-speed 3D imaging, surpassing the camera's acquisition rate by nearly an order of magnitude, while preserving spatial resolution. We employ a series of fringe images with varying tilt angles. When the speed of projector is higher than that of camera, we capture a multiplexed image overlaid with a sequence of fringe patterns. DLMFPP can decode the image into its original sequence by DNNs embedded with Fourier transform (FT), PS [26], and ensemble learning [27]. By harnessing each fringe pattern to record the scene at different time, it achieves up to 9x temporal super-resolution imaging beyond the camera's frame rate. In practice, the DLMFPP method can be implemented on almost any off-the-shelf FPP system, eliminating the need for complicated optical paths



and furnishing a high SNR and ready-to-use solution compared to conventional computational imaging techniques [28–30]. We validate the effectiveness and versatility of DLMFPP through experimental demonstrations on different types of transient scenes, including rotating fan blades and bullet fired from a toy gun, showcasing its ability to achieve high-speed kHz 3D imaging with low-speed cameras operating at around 100 Hz. By transcending the limitations of sensor frame rates, the DLMFPP allows slow-scan cameras to quantitatively study dynamic processes with both high spatial and temporal resolution.

## Methods

The schematic of the DLMFPP approach is demonstrated in Fig. 1. The projector sequentially projects fringe patterns  $I_m^p$  with different directions onto the dynamic scene. The pattern sequence can be represented as

$$I_m^p(x^p, y^p) = a^p + b^p \cos[\varphi_m^p(x^p, y^p)], \quad (1)$$

where  $(x^p, y^p)$  represents the pixel coordinate of projector,  $a^p$  is the mean value,  $b^p$  is the amplitude, and  $m$  denotes the pattern index  $m = 1, 2, 3, \dots, M$  ( $M$  is the total number of the patterns). The phase  $\varphi_m^p$  is assigned as

$$\varphi_m^p(x^p, y^p) = 2\pi \left( f_x^p x^p \cos\theta_m + f_y^p y^p \sin\theta_m \right), \quad (2)$$

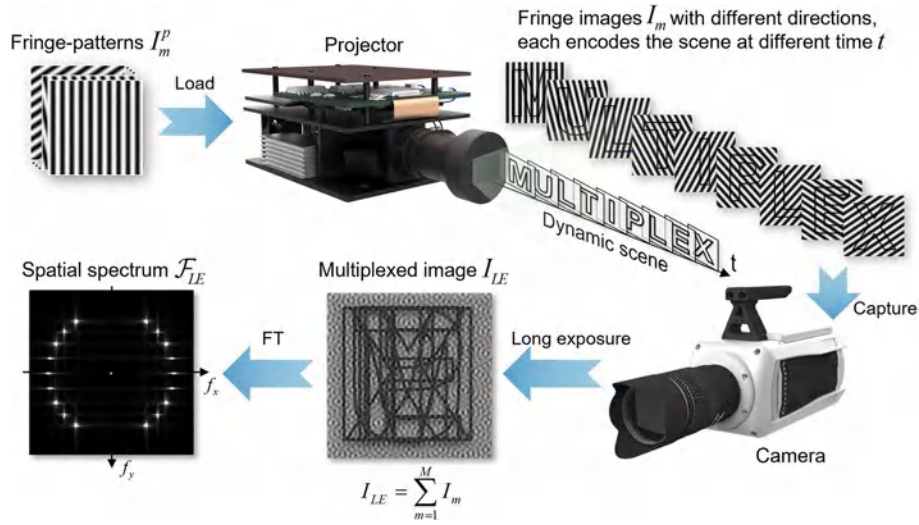
$$\theta_m = (-1)^m \left( \frac{m}{2} + \frac{(-1)^m - 1}{4} \right) \theta, \quad (3)$$

where  $f_x^p$  and  $f_y^p$  are the frequency in  $x^p$ ,  $y^p$  directions, respectively, and  $\theta$  is a scalar characterizing the incline of fringes. After modulated by the object surface, the corresponding fringe images  $I_m$  (shown in Fig. 1) can be expressed as

$$I_m(x, y) = A_m(x, y) + B_m(x, y) \cos[\phi_m(x, y)], \quad (4)$$

where  $(x, y)$  indicates the pixel coordinate of camera,  $A_m$  is the average intensity,  $B_m$  is the modulation, and  $\phi_m$  is the phase to be measured. Letters of “MULTIPLEX” in Fig. 1 represent a dynamic scene, and each  $I_m$  encodes the scene at different time  $t$ . Then, the camera captures a multiplexed image  $I_{LE}$  overlaid by the sequence of  $I_m$  with a long exposure time. After performing FT on  $I_{LE}$ , multiple fundamental frequency components (corresponding to  $I_m$ ) are circularly distributed in the spatial spectrum  $\mathcal{F}_{LE}$ , occupying distinct locations. Specifically, we consider four principles when designing the pattern sequence  $I_m^p$ : (1) the fringe interval in each  $I_m^p$  is kept equal to guarantee the consistent defocusing level when capturing the binary pattern sequence; (2) the zero component in  $\mathcal{F}_{LE}$  should be far away from the fundamental components to avoid spectrum overlap; (3) the fundamental components of these fringe patterns should be distributed in a circular pattern in  $\mathcal{F}_{LE}$ , which minimizes the harm of spectrum leakage; (4) fundamental components near  $f_y$  axis should be excluded as it is hard to employ this kind of near-horizontal fringe pattern to measure 3D shape for a conventional horizontally configured FPP system.

The flowchart of DLMFPP is shown in Fig. 2, where there are two steps to analyze the input multiplexed image. Step 1 is to decompose the multiplexed pattern into a fringe

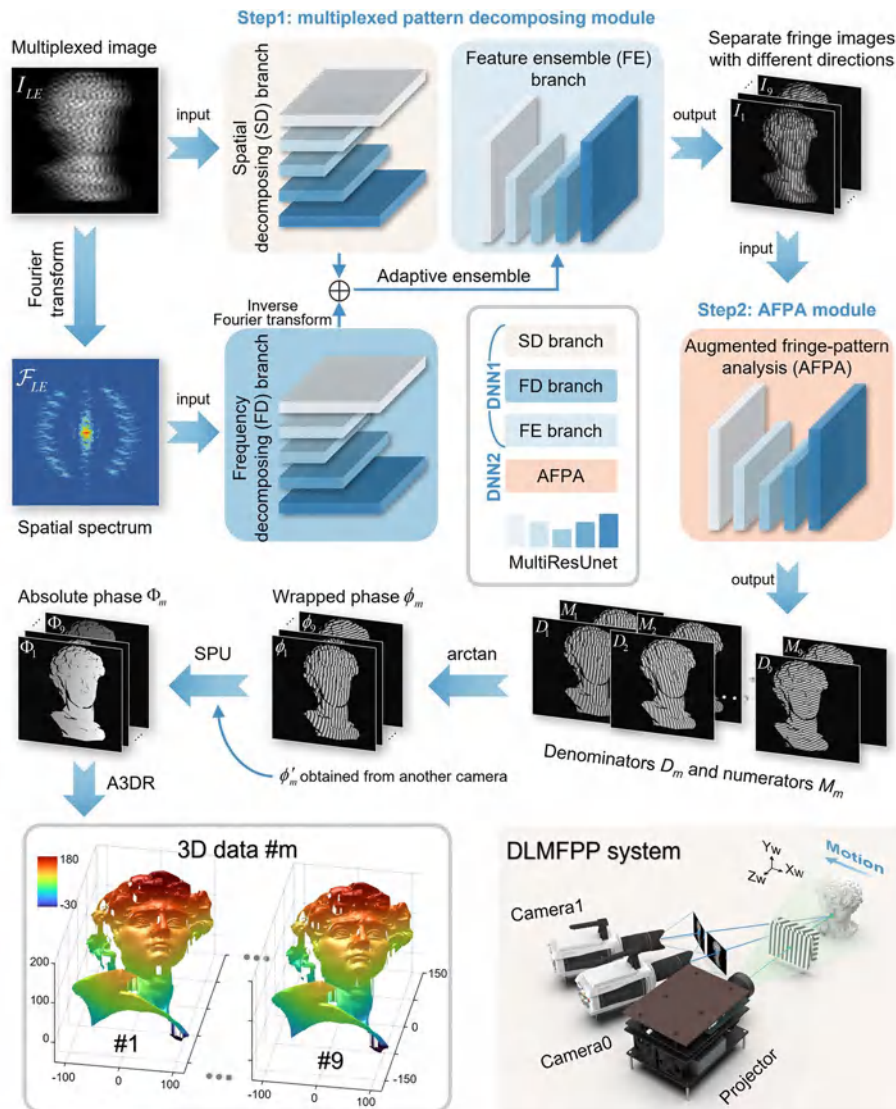


**Fig. 1** Schematic of DLMFPP: The projector sequentially projects fringe patterns  $I_m^p$  [Eq. (1)] onto the dynamic scene, allowing the corresponding modulated fringe images  $I_m$  [Eq. (4)] to encode the scene at different time  $t$ . Then the camera captures a multiplexed image  $I_{LE}$  with a long exposure time, and the spatial spectrum  $\mathcal{F}_{LE}$  (multiple fundamental components corresponding to  $I_m$  are circularly distributed) can be obtained by FT (pattern index  $m = 1, 2, 3, \dots, M$ ,  $M$  is the total number of the patterns). A synthetic scene composed of letters, “MULTIPLY”, is used to illustrate the principle

pattern sequence, each of which corresponds to the measured object at a moment. Step 2 is to analyze the decomposed fringe patterns for phase retrieval. To be specific, inspired by the rationalized deep learning framework [31], we propose a multiplexed pattern decomposing module (DNN1) that comprises three branches. The spatial decomposing (SD) branch is trained to extract the features of the multiplexed image  $I_{LE}$  and decompose it in the spatial domain. The frequency decomposing (FD) branch, which is parallel to the SD branch, incorporates the physical model of FT into the framework to analyze the multiplexed image as follows: (1) it obtains the spatial spectrum  $\mathcal{F}_{LE}$  of  $I_{LE}$  by FT, and feeds its real and imaginary components into the FD branch [32]; (2) the branch then decomposes  $\mathcal{F}_{LE}$  in frequency domain and outputs the real and imaginary parts of the separate spectrums as the branch output; (3) inverse FT (iFT) is performed to obtain separate fringe images. The feature ensemble (FE) branch is engineered to adaptively merge features learned by the SD and FD branches with the idea of ensemble learning [27]. This branch can incorporate features from both spatial and frequency domains and give the final outputs, i.e., separate fringe images  $I_1 - I_9$  in Fig. 2. In Step 2, we design an augmented fringe pattern analysis (AFPA) module (DNN2) embedded with the physical model of PS to retrieve the phase from each fringe image. The module receives each separate fringe image  $I_m$  as input and predicts the corresponding numerator  $M_m$  and denominator  $D_m$ . Then, the wrapped phase  $\phi_m$  in Eq. (4) is demodulated through an arctangent function

$$\phi_m(x, y) = \arctan \frac{cB_m(x, y) \sin[\phi_m(x, y)]}{cB_m(x, y) \cos[\phi_m(x, y)]} = \arctan \frac{M_m(x, y)}{D_m(x, y)}, \quad (5)$$

where  $c$  is a constant determined by the phase demodulation approach, pattern index  $m = 1, 2, 3, \dots, 9$ . After that, the absolute phase  $\Phi_m$  can be acquired with the help of  $\phi'_m$  from another camera via stereo phase unwrapping (SPU) [33], then 3D reconstruction



**Fig. 2** Flowchart of DLMFPP. A multiplexed image  $I_{LE}$  and its spatial spectrum  $\mathcal{F}_{LE}$  are fed into a multiplexed pattern decomposing module (DNN1) comprised of three branches. The DNN1 framework incorporates the physical model of FT and the idea of ensemble learning to decompose  $I_{LE}$  and output separate fringe images  $I_m$ . The AFPA module (DNN2) embedded with the physical model of PS receives each  $I_m$  to predict the corresponding  $M_m$  and  $D_m$ , enabling wrapped phase  $\phi_m$  calculation via Eq. (5). The absolute phase  $\Phi_m$  is then derived by SPU, and 3D data of # $m$  can be reconstructed by the developed A3DR (pattern index  $m = 1, 2, 3, \dots, 9$ ). The insert shows the DLMFPP system configuration, consisting of a projector and two cameras. The projector sequentially projects nine fringe patterns with different directions onto a moving object, then the cameras capture the multiplexed image (shown as  $I_{LE}$ ) with a long exposure time

can be performed. Notably, in a conventional horizontally configured FPP system, the mapping from phase to 3D coordinates is generally designed for vertical fringes. To cope with the case of arbitrarily oriented fringes in this work, we propose the augmented 3D reconstruction (A3DR) method. By creating a unique correspondence value  $x^p \cos\theta_m + (f_y^p/f_x^p)y^p \sin\theta_m$  for every camera pixel coordinate  $(x, y)$ , 3D reconstruction can be performed from Eq. (S13) with pre-calibrated parameters. For further details on system calibration and A3DR, see Supplementary Note 6.

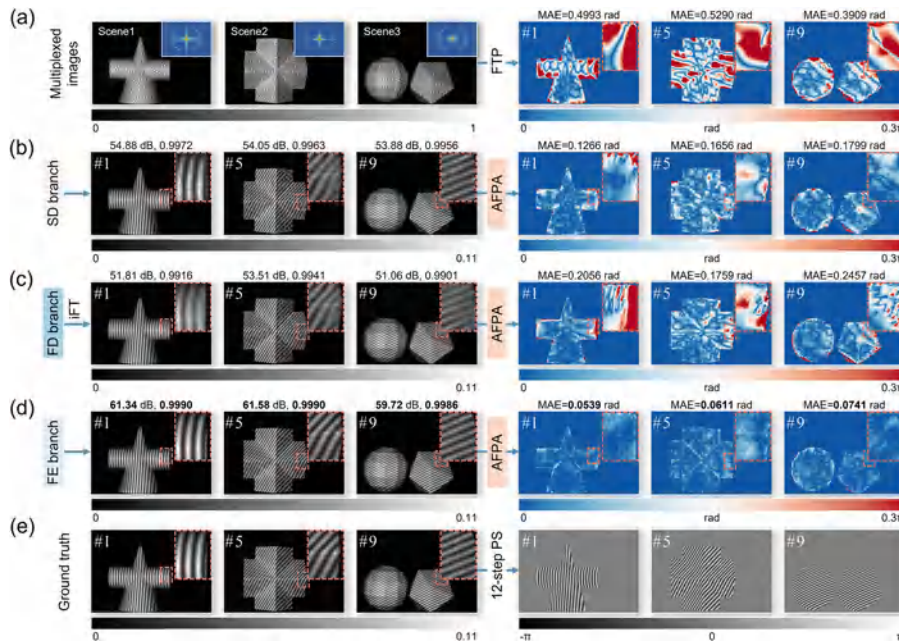
The SD, FD, FE branches and the AFPA module are constructed by MultiResUnet [34], which is a novel architecture that combines MultiRes blocks and residual paths on the well-known U-Net framework [35], owing the advantage to reconcile features from different context size, alleviate the disparity between the encoder-decoder features, save memory and speed up network training (detailed in Supplementary Note 2 and Fig. S2). Network training for multiplexed pattern decomposition and phase retrieval is carried out in a supervised manner, and the process is elaborated in Supplementary Note 4 and Fig. S4. Moreover, for the objective functions of training, the SD and FD branches use joint losses containing data-based and physics-based loss, while the FE branch and the AFPA module use only the data-based loss. The combination of physical and data loss can effectively improve the recovered accuracy and generalization of the DNNs. Details related to the loss functions design are provided in Supplementary Note 5 and Fig. S5. By incorporating FT, PS and ensemble learning, DLMFPP embeds more physical prior knowledge in the network structure and loss functions to provide reliable phase recovery across various scenes and conditions, significantly improving the generalization ability of networks.

We developed the DLMFPP system shown in the insert of Fig. 2, composed by two CMOS cameras (Vision Research Phantom V611) and a customized projection system with an XGA resolution (1024×768) DMD. By functioning in binary (1-bit) mode, the DMD is manipulated to achieve a refresh rate of 1,000 fps. Meanwhile, the cameras are operated at an image resolution (640×440) with pixel depth of 16 bits. The projection system outputs a trigger signal every nine frames, thus the cameras work at a frame rate of ~111.11 Hz. DLP development hardware is used for precisely triggering to ensure signal synchronization between the projector and the cameras. For more information about the system synchronization, see Supplementary Note 1 and Fig. S1. During the training stage, we photographed a variety of objects made of different materials (plastic, plaster, metal, ceramic, etc.) to generate diverse datasets. In this work, 1,200 groups of images were captured, of which 800 groups were used for training and 400 groups for validation. Details of training dataset generation can be found in Supplementary Note 3 and Fig. S3.

## Results

To evaluate the contribution of each branch in DLMFPP, we measured three scenes to conduct an ablation study as shown in Fig. 3. The ground truths of separate fringe images were captured by setting the camera frame rate to 1,000 Hz (same as the DMD refresh rate). Then, the ground truths of phase were obtained by 12-step PS, as in Fig. 3e (detailed in Supplementary Note 3). Figure 3a shows multiplexed images modulated by the scenes (insets show the corresponding Fourier frequency spectrums, locally zoomed in for better visibility) and the phase errors of FTP. We can see substantial phase errors on the sharp edges of the measured surface, and the average mean absolute error (MAE) of these scenes is up to 0.4731 rad. Figure 3b-d show the separate fringe images decomposed by the SD, FD, and FE branches, respectively, and the corresponding phase errors of the reconstructed results demodulated by AFPA. From the fringe images in Fig. 3b, we can observe obvious noise. Meanwhile, blur fringes can be observed around the edges of the object as shown in Fig. 3c, which results in significant phase errors with an average MAE of 0.2091 rad. Contrastingly, in Fig. 3d, the FE branch harnesses the idea

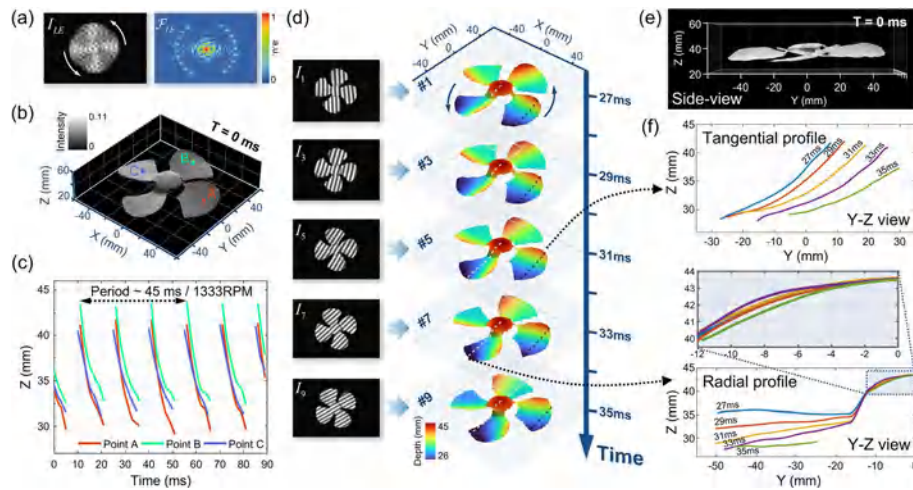




**Fig. 3** Ablation study of DLMFPP: **a** Multiplexed images modulated by 3 different scenes [insets show the corresponding spatial spectrums (locally zoomed in)] and phase errors of FTP; **b-d** separate fringe images decomposed by SD, FD, and FE branches, respectively, evaluated by PSNR and SSIM, and phase errors of the reconstructed results demodulated by AFPA; **e** ground truths of separate fringe images and phase, obtained by setting the camera frame rate same as the DMD refresh rate (1,000 Hz) and 12-step PS ( $m$  represents the  $m$ th pattern index of each scene, and  $m = 1, 2, 3, \dots, 9$ )

of ensemble learning to integrate features from both the spatial and frequency domains, yielding a high-quality restoration of fringe images. The resultant average peak SNR (PSNR) ups to 60.88 dB and the average structural similarity index (SSIM) ups to 0.9989. By feeding these fringe images into AFPA, we can achieve high-accuracy phase recovery with the average MAE of 0.0630 rad.

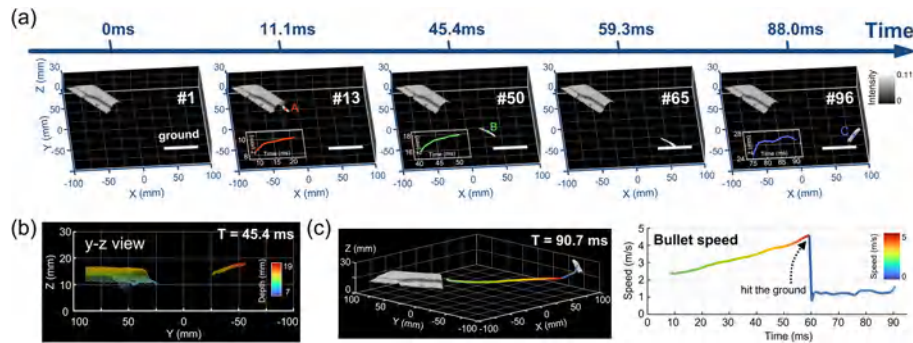
For dynamic 3D measurements of moving objects, we applied DLMFPP to measure a fan with 4 rotating plastic blades. Figure 4a presents a particular frame of the multiplexed image  $I_{LE}$  and corresponding spectrum  $\mathcal{F}_{LE}$  (locally zoomed in for better visibility). Although significant motion blur of the blades is observed in the multiplexed image, the proposed DLMFPP can still successfully reconstruct the 3D shape of the blades, as shown in Fig. 4b and e. It is noted that the motion blur in DLMFPP is not determined by the camera exposure time, but by the projection time, which is near-one-order of magnitude-lower than the exposure time of a single camera frame. This greatly reduced exposure time effectively handles the challenges of motion blur of dynamic scene changes, thus ensuring accurate 3D reconstruction. For more information on the discussion of motion blur in DLMFPP, see Supplementary Note 8 and Fig. S8. Figure 4c plots the displacement of  $z$  at 3 selected point locations within 90 ms [A, B, and C in Fig. 4b], revealing that the rotation period of the fan blades is 45 ms, i.e., the rotation speed is 1,333 rotations per minute (rpm). Figure 4d shows five fringe images ( $I_1, I_3, I_5, I_7$ , and  $I_9$ , corresponding to  $T = 27, 29, 31, 33$ , and  $35$  ms) decoded from the multiplexed image  $I_{LE}$  and the corresponding 3D model reconstructed by the proposed DLMFPP. Moreover, Fig. 4f displays two cross sections of the 3D reconstruction, one of which shows the tangential



**Fig. 4** Measurement of a rotating fan by DLMFPP. **a** The multiplexed image  $I_{LE}$  and corresponding spectrum  $\mathcal{F}_{LE}$  (locally zoomed in). **b** 3D reconstruction of the fan at  $T = 0$  ms. **c** Displacement of  $z$  at 3 selected point locations within 90 ms [A, B, and C in (b)]. **d** Five fringe images ( $I_1, I_3, I_5, I_7,$  and  $I_9$ ), corresponding to  $T = 27, 29, 31, 33,$  and  $35$  ms) decoded from the multiplexed image  $I_{LE}$ , and the corresponding 3D model reconstructed by DLMFPP. **e** Side-view of (b). **f** Two cross sections of the 3D reconstruction, one of which shows the tangential profile (black dot line) and the other the radial profile (white dot line). The local zoomed-in view shows the profile of the centre hub

profile (black dot line) and the other the radial profile (white dot line). The profile of the centre hub is shown in the zoomed-in view. The corresponding 3D movie about the complete process of DLMFPP and 3D reconstruction results of the whole dynamic process of the rotating fan is further provided in Supplementary Movie S1. With this experiment, we can see that DLMFPP accurately retrieved nine 3D images with each multiplexed image  $I_{LE}$ , validating that 1,000 Hz high-speed 3D shape measurement has been achieved with cameras running at  $\sim 111.11$  Hz. Additionally, we applied DLMFPP to image a running fascia gun for a supplementary experiment. It shows that the cyclic movement of the gun head has a period of about 35 ms, which corresponds to a speed of 1,714 rpm of the rotary motor inside the gun. More experimental results are provided in Supplementary Note 9, Fig. S9 and Supplementary Movie S3.

To verify the scalability of our DNNs, we developed another system consisting of two low-speed cameras (Basler acA640-750um) and the same projection unit. The cameras are equipped with zoom lenses that adjust the focal length, aperture size and degree of focus to make the field of view and brightness consistent with the existing datasets. So we can directly utilize the trained DNNs before. The projector operated at the rate of 1,080 fps and the camera at 120 fps. For the dynamic experiment, we measured a one-time transient event: a bullet was fired diagonally downward from a toy gun, and then rebounded from the ground. Representative 3D reconstruction results during the event are presented in Fig. 5a. The bullet began to appear near the muzzle at 11.1 ms. It flew straight forward until 59.3 ms and then hit the ground and rebounded upwards. Three points are selected to demonstrate the performance of DLMFPP [A, B, and C in Fig. 5a]. The displacements in  $z$  direction at selected locations are plotted in insets of Fig. 5a, indicating that DLMFPP has accurately recovered the profile of the fast moving bullet at different moments. Figure 5b shows the side-view ( $y$ - $z$ ) of the 3D reconstruction at  $T = 45.4$  ms, and Fig. 5c shows the trajectory and the variation of the velocity of the bullet



**Fig. 5** Measurement of bullet fired from a toy gun by DLMFPP. **a** 3D reconstruction results at  $T = 0, 11.1, 45.4, 59.3,$  and  $88.0$  ms, with insets presenting displacements in  $z$  direction at A, B, and C locations. **b** The side-view ( $y$ - $z$ ) of the 3D reconstruction at  $T = 45.4$  ms. **c** The 3D reconstruction of the scene at  $T = 90.7$  ms, as well as the trajectory and the variation of the velocity of the bullet during the whole process

during the whole process. The initial speed of the bullet was  $2.4$  m/s at discharge. It accelerated uniformly to  $4.6$  m/s during the flight and then hit the ground with the speed decreased abruptly to  $0.8$  m/s (refer to Supplementary Movie S2 for more details). The experiment demonstrates the scalability of our DNNs for high-speed 3D imaging with low-speed cameras and the capability of DLMFPP to capture one-time transient events.

It should be noted that DLMFPP is the first temporally super-resolved 3D imaging technique proposed in FPP, while previous deep learning-based approaches were developed for single-shot 3D imaging [20–23]. The structure, training process, and loss function design of previous networks cannot meet the necessity for high-accuracy phase recovery and measurement in temporally super-resolved 3D imaging, therefore we proposed DLMFPP to address this challenge. To justify the progressiveness of DLMFPP, in Supplementary Note 7 and Fig. S6, we provide a comparative study and analysis between the proposed DLMFPP and two state-of-the-art deep learning-based approaches. This study demonstrates that DLMFPP solves the dilemma of the state-of-the-art methods in handling regions with large height variations and demodulates high-accuracy phase information from the multiplexed image. DLMFPP achieves the lowest phase error with the average MAE of  $0.0495$  rad, revealing the superior performance achieved from DLMFPP’s advanced network design.

For the 3D imaging speed in DLMFPP, the increase of imaging speed depends on the number of overlapped images in a multiplexed image. The overlapping number is referred to as compression rate (CR). In this work, we employ  $CR = 9$  when the marginal benefit between CR and recovered phase accuracy is highest (detailed in the comparative study of different CRs in Supplementary Note 7 and Fig. S7), allowing DLMFPP to achieve 9x temporal super-resolution. Practically, to trade off temporal resolution and spatial resolution accuracy, the DLMFPP approach is also flexible. If higher phase accuracy is required, CR can be reduced appropriately, and vice versa.

## Discussion and conclusion

In this work, we have introduced a deep-learning-enabled temporally super-resolved 3D measurement approach by multiplexed FPP. By temporally embedding a sequence of fringe patterns with different tilt angles into a single multiplexed image, DLMFPP allows

to achieve high-resolution and high-speed 3D imaging at near-one-order of magnitude-higher 3D frame rate with conventional low-speed cameras. Experimental results demonstrate that kHz 3D imaging can be achieved by using cameras merely running at around 100 Hz without compromising the spatial resolution.

DLMFPP encodes multi-frame temporal information in the spatial dimension, which gives this compressive imaging modality the advantage of cost-effective, low bandwidth/memory requirements, and low power consumption [36]. Moreover, the modality breaks through the limitation of 3D imaging speed imposed by the intrinsic frame rate of the imaging sensor, allowing it to be further used for ultrahigh-speed imaging when combined with high-speed cameras. This new 3D imaging paradigm opens an avenue for the development of high-speed or ultra-high-speed 3D imaging capabilities, thereby pushing the boundaries of current 3D imaging technologies.

Compared to conventional computational imaging techniques [28–30], DLMFPP system eliminates the need for complex optical modulation hardware (e.g., a spatial encoder), avoiding complicated optical paths. Practically, DLMFPP can be implemented on almost any off-the-shelf FPP system. This simple optical path avoids photon losses and makes greater use of optical information, guaranteeing a high SNR in 3D imaging. Moreover, DLMFPP combines the physical models of FT and PS method, and harnesses the idea of ensemble learning to integrate features from both the spatial and frequency domains. This progressive architecture also ensures the high SNR in high-speed 3D imaging with low-speed cameras. From the perspective of space-time-bandwidth product (STBP), the multi-frame modulation mechanism of DLMFPP can rationally harness the spatio-temporal redundancy in fast changing scenes, thereby better utilizing the STBP of sensors compared to conventional single-frame recordings.

Despite promising results in high-speed 3D imaging, DLMFPP still faces challenges. For example, the exclusion of near-horizontal fringe patterns leaves the region near  $f_y$  axis in the multiplexed spatial spectrum unused, which exacerbates the harm of spectrum overlap, affecting the recovered phase quality. Moreover, due to the trade-off between CR and the information capacity of each fringe image, further increasing the multiple of temporal super-resolution results in a loss of final phase quality, and vice versa. It should also be noted that the maximum speed of DLMFPP is still constrained by the projection rate. The speed can be potentially further enhanced by using custom physical grating [7] or LED arrays [8, 9], which will be explored in our future research. Furthermore, there is an untapped potential of DLMFPP, as latest innovations in deep learning can be directly introduced into the method. For example, physics-informed learning can bring domain expertise to improve performance [37–40], and all-optical neural networks operating at the speed of light can accelerate computations [41–43].

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s43074-024-00139-2>.

Supplementary Material 1: Supplementary Information.

Supplementary Material 2: Movie S1.

Supplementary Material 3: Movie S2.

Supplementary Material 4: Movie S3.



**Acknowledgements**

Not applicable.

**Authors' contributions**

C.Z., W.C., and S.F. developed the theoretical description of the method; W.C. performed experiments and analyzed data; Q.C. and C.Z. conceived and supervised the research; All authors contributed to writing the manuscript.

**Funding**

This work was supported by National Key Research and Development Program of China (2022YFB2804603), National Natural Science Foundation of China (62075096, 62005121, U21B2033), Leading Technology of Jiangsu Basic Research Plan (BK20192003), "333 Engineering" Research Project of Jiangsu Province (BRA2016407), Fundamental Research Funds for the Central Universities (30921011208, 30919011222, 30920032101), Fundamental Research Funds for the Central Universities (2023102001, 2024202002).

**Availability of data and materials**

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

**Declarations****Ethics approval and consent to participate**

There is no ethics issue for this paper.

**Consent for publication**

All authors agreed to publish this paper.

**Competing interests**

The authors declare that they have no competing interests.

Received: 7 May 2024 Revised: 8 July 2024 Accepted: 2 August 2024

Published online: 19 August 2024

**References**

- Malamas EN, Petrakis EGM, Zervakis M, Petit L, Legat JD. A survey on industrial vision systems, applications and tools. *Image Vision Comput.* 2003;21(2):171–88.
- Ford KR, Myer GD, Hewett TE. Reliability of landing 3D motion analysis: implications for longitudinal analyses. *Med Sci Sports Exerc.* 2007;39(11):2021.
- Tiwari V, Sutton MA, McNeill SR. Assessment of High Speed Imaging Systems for 2D and 3D Deformation Measurements: Methodology Development and Validation. *Exp Mech.* 2007;47(4):561–79.
- Gorghi SS, Rastogi P. Fringe projection techniques: whither we are? *Optics Lasers Eng.* 2010;48(2):133–40.
- Li B, Wang Y, Dai J, Lohry W, Zhang S. Some recent advances on superfast 3D shape measurement with digital binary defocusing techniques. *Optics Lasers Eng.* 2014;54:236–46.
- Zuo C, Chen Q, Feng S, Feng F, Gu G, Sui X. Optimized pulse width modulation pattern strategy for three-dimensional profilometry with projector defocusing. *Appl Opt.* 2012;51(19):4477–90.
- Heist S, Lutzke P, Schmidt I, Dietrich P, Kühmstedt P, Tünnermann A, et al. High-speed three-dimensional shape measurement using GOBO projection. *Opt Lasers Eng.* 2016;87:90–6.
- Heist S, Mann A, Kühmstedt P, Schreiber P, Notni G. Array projection of aperiodic sinusoidal fringes for high-speed three-dimensional shape measurement. *Opt Eng.* 2014;53(11):112208.
- Caspar S, Honegger M, Rinner S, Lambelet P, Bach C, Ettemeyer A. High speed fringe projection for fast 3D inspection. In: *Optical Measurement Systems for Industrial Inspection VII*. vol. 8082. SPIE; 2011. p. 298–304.
- Feng S, Zuo C, Tao T, Hu Y, Zhang M, Chen Q, et al. Robust dynamic 3-D measurements with motion-compensated phase-shifting profilometry. *Optics Lasers Eng.* 2018;103:127–38.
- Liu K, Wang Y, Lau DL, Hao Q, Hassebrook LG. Dual-frequency pattern scheme for high-speed 3-D shape measurement. *Opt Express.* 2010;18(5):5229–44.
- Zuo C, Chen Q, Gu G, Feng S, Feng F, Li R, et al. High-speed three-dimensional shape measurement for dynamic scenes using bi-frequency tripolar pulse-width-modulation fringe projection. *Optics Lasers Eng.* 2013;51(8):953–60.
- Tao T, Chen Q, Da J, Feng S, Hu Y, Zuo C. Real-time 3-D shape measurement with composite phase-shifting fringes and multi-view system. *Opt Express.* 2016;24(18):20253–69.
- Zuo C, Tao T, Feng S, Huang L, Asundi A, Chen Q. Micro Fourier transform profilometry ( $\mu$ FTP): 3D shape measurement at 10,000 frames per second. *Optics Lasers Eng.* 2018;102:70–91.
- Takeda M, Mutoh K. Fourier transform profilometry for the automatic measurement of 3-D object shapes. *Appl Opt.* 1983;22(24):3977.
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436–44.
- Schmidhuber J. Deep learning in neural networks: An overview. *Neural Netw.* 2015;61:85–117.
- Zuo C, Qian J, Feng S, Yin W, Li Y, Fan P, et al. Deep learning in optical metrology: a review. *Light-Sci Appl.* 2022;11(1):39.
- Feng S, Chen Q, Gu G, Tao T, Zhang L, Hu Y, et al. Fringe pattern analysis using deep learning. *Adv Photon.* 2019;1(02):1.

20. Qian J, Feng S, Li Y, Tao T, Han J, Chen Q, et al. Single-shot absolute 3D shape measurement with deep-learning-based color fringe projection profilometry. *Opt Lett*. 2020;45(7):1842–5.
21. Qian J, Feng S, Tao T, Hu Y, Li Y, Chen Q, et al. Deep-learning-enabled geometric constraints and phase unwrapping for single-shot absolute 3D shape measurement. *Apl Photon*. 2020;5(4):046105.
22. Li Y, Qian J, Feng S, Chen Q, Zuo C. Deep-learning-enabled dual-frequency composite fringe projection profilometry for single-shot absolute 3D shape measurement. *Opto-Electron Adv*. 2022;5(5):210021.
23. Li Y, Qian J, Feng S, Chen Q, Zuo C. Composite fringe projection deep learning profilometry for single-shot absolute 3D shape measurement. *Opt Express*. 2022;30(3):3424–42.
24. Barbastathis G, Ozcan A, Situ G. On the use of deep learning for computational imaging. *Optica*. 2019;6(8):921–43.
25. Shaked NT, Micó V, Trusiak M, Kuś A, Mirsky SK. Off-axis digital holographic multiplexing for rapid wavefront acquisition and processing. *Adv Opt Photon*. 2020;12(3):556.
26. Zuo C, Feng S, Huang L, Tao T, Yin W, Chen Q. Phase shifting algorithms for fringe projection profilometry: A review. *Opt Lasers Eng*. 2018;109:23–59.
27. Feng S, Xiao Y, Yin W, Hu Y, Li Y, Zuo C, et al. Fringe-pattern analysis with ensemble deep learning. *Adv Photon Nexus*. 2023;2(3):036010.
28. Gao L, Liang J, Li C, Wang LV. Single-shot compressed ultrafast photography at one hundred billion frames per second. *Nature*. 2014;516(7529):74–7.
29. Yuan X, Brady DJ, Katsaggelos AK. Snapshot compressive imaging: theory, algorithms, and applications. *IEEE Signal Proc Mag*. 2021;38(2):65–88.
30. He Y, Yao Y, Qi D, He Y, Huang Z, Ding P, et al. Temporal compressive super-resolution microscopy at frame rate of 1200 frames per second and spatial resolution of 100 nm. *Adv Photon*. 2023;5(2):026003.
31. Qiao C, Li D, Liu Y, Zhang S, Liu K, Liu C, et al. Rationalized deep learning super-resolution microscopy for sustained live imaging of rapid subcellular processes. *Nat Biotechnol*. 2023;41(3):367–77.
32. Yin W, Che Y, Li X, Li M, Hu Y, Feng S, et al. Physics-informed deep learning for fringe pattern analysis. *Opto-Electron Adv*. 2024;7(1):230034–1.
33. Weise T, Leibe B, Van Gool L. Fast 3D Scanning with Automatic Motion Compensation. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis: IEEE; 2007. pp. 1–8.
34. Ibtehaz N, Rahman MS. MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Netw*. 2020;121:74–87.
35. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III 18. Springer; 2015. p. 234–241.
36. Zhang Z, Zhang B, Yuan X, Zheng S, Su X, Suo J, et al. From compressive sampling to compressive tasking: retrieving semantics in compressed domain with low bandwidth. *Photonix*. 2022;3(1):19.
37. Kellman MR, Bostan E, Repina NA, Waller L. Physics-based learned design: optimized coded-illumination for quantitative phase imaging. *IEEE Trans Comput Imaging*. 2019;5(3):344–53.
38. Wang F, Bian Y, Wang H, Lyu M, Pedrini G, Osten W, et al. Phase imaging with an untrained neural network. *Light Sci Appl*. 2020;9(1):77.
39. Bostan E, Heckel R, Chen M, Kellman M, Waller L. Deep phase decoder: self-calibrating phase microscopy with an untrained deep neural network. *Optica*. 2020;7(6):559–62.
40. Saba A, Gigli C, Ayoub AB, Psaltis D. Physics-informed neural networks for diffraction tomography. *Adv Photon*. 2022;4(6):066001.
41. Lin X, Rivenson Y, Yardimci NT, Veli M, Luo Y, Jarrahi M, et al. All-optical machine learning using diffractive deep neural networks. *Science*. 2018;361(6406):1004–8.
42. Liu J, Wu Q, Sui X, Chen Q, Gu G, Wang L, et al. Research progress in optical neural networks: theory, applications and developments. *Photonix*. 2021;2:1–39.
43. Luo Y, Zhao Y, Li J, Çetintaş E, Rivenson Y, Jarrahi M, et al. Computational imaging without a computer: seeing through random diffusers at the speed of light. *ELight*. 2022;2(1):4.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Contents lists available at ScienceDirect

## Optics and Lasers in Engineering

journal homepage: [www.elsevier.com/locate/optlaseng](http://www.elsevier.com/locate/optlaseng)

## Micro deep learning profilometry for high-speed 3D surface imaging

Shijie Feng<sup>a,b,c</sup>, Chao Zuo<sup>a,b,c,\*</sup>, Wei Yin<sup>a,b,c</sup>, Guohua Gu<sup>a,b</sup>, Qian Chen<sup>a,b,\*</sup><sup>a</sup> School of Electronic and Optical Engineering, Nanjing University of Science and Technology, No. 200 Xiaolingwei Street, Nanjing, Jiangsu Province 210094, China<sup>b</sup> Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing, Jiangsu Province 210094, China<sup>c</sup> Smart Computational Imaging Laboratory (SCILab), Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China

## ARTICLE INFO

## Keywords:

Deep learning  
3D surface imaging  
Structured light

## ABSTRACT

How to obtain object information as rich as possible, with the highest possible speed and accuracy from recorded optical signals, has been a crucial issue to the pursuit of powerful imaging technologies. Nowadays, the speed of ultra-fast photography can exceed one quadrillion. However, it can record only two-dimensional images which lack the depth information, greatly limiting our ability to perceive and to understand the complex real-world objects. Inspired by recent successes of deep learning methods in computer vision, we present a novel high-speed three-dimensional (3D) surface imaging approach named micro deep learning profilometry ( $\mu$ DLP) using the structured light illumination. With a properly trained deep neural network, the phase information is predicted from a single fringe image and then can be converted into the 3D shape. Our experiments demonstrate that  $\mu$ DLP can faithfully retrieve the geometry of dynamic objects at 20,000 frames per second. Moreover, comparative results show that  $\mu$ DLP has superior performance in terms of the phase accuracy, reconstruction efficiency, and the ease of implementation over widely used Fourier-transform-based fast 3D imaging techniques, verifying that  $\mu$ DLP is a powerful high-speed 3D surface imaging approach.

## 1. Introduction

It is usually said that the first instance of what we would call high-speed photography nowadays was to settle the hot dispute “is there a moment in a horse’s gait when all four hooves are off the ground at once?” in 1872 [1]. Eadweard Muybridge, a pioneer in the field of motion study, developed an imaging system that involved 12 cameras triggered by the legs of the horse through tripwires, successfully capturing photos on photographic glass plates at the shutter speed of near 2000 frames per second (fps) [2]. After that, the major development for high-speed photography came, as with scientific purposes, in the wake of the researches on nuclear weapons during the cold war. With applications of rotating mirror technologies, streak cameras, and rotating prism cameras [3], the imaging speed soared up to 100 million fps, i.e., Mfps. In the late nineteenth century, the high-speed imaging underwent a further advancement owing to the great breakthrough in electronic semiconductor devices, leading to film-based cameras replaced gradually by CCD or CMOS based cameras [4]. Nowadays, with the assistance of laser, e.g., the femtosecond laser pulse [5], the imaging speed can even exceed one quadrillion, i.e.,  $10^{15}$  fps. Benefiting from the ever-increasing power of the high-speed photography, many transient events, which happen at femtosecond to nanosecond time scale and reflect significant fundamental mechanisms, can be analyzed in-depth [6–11].

However, most high-speed cameras or imaging systems can record only two-dimensional (2D) images which lack the depth information. This fundamental restriction greatly limits our ability to perceive and to understand the complex real-world objects. The past several decades have witnessed tremendous development in three-dimensional (3D) imaging technologies in many fields including biomechanics [12], geomaterials [4], industrial manufacturing [13–15], driven by the rapid advances in sensors, optical engineering and computer vision [16–21]. In general, optical 3D surface imaging techniques can be classified into two categories: the passive approaches and the active ones. Stereo vision techniques, as the representative passive methods, capture inherent surface textures from two or more viewpoints and calculate 3D shapes through triangulation [22]. However, they are susceptible to uniform or periodic textures. Compared with the passive sensing, active methods encode test objects with predesigned signals, thus reducing the dependence of the object textures and increasing the accuracy of 3D reconstructions. Time-of-flight (ToF) techniques emit a modulated light ray onto test objects and collect the light scattered back. The distance is then estimated via multiplying the speed of light by the time delay of the light pulse [23]. As the 3D reconstruction of ToF is not based on triangulation, the system can be made very compactly for applications where portable equipment is preferred. Microsoft Kinect 2 exploits this technique for real-time 3D imaging and finds applications for

\* Corresponding authors.

E-mail addresses: [shijiefeng@njust.edu.cn](mailto:shijiefeng@njust.edu.cn) (S. Feng), [zuochao@njust.edu.cn](mailto:zuochao@njust.edu.cn) (C. Zuo), [chenqian@njust.edu.cn](mailto:chenqian@njust.edu.cn) (Q. Chen).<https://doi.org/10.1016/j.optlaseng.2019.04.020>

Received 14 February 2019; Received in revised form 22 April 2019; Accepted 23 April 2019

Available online 15 May 2019

0143-8166/© 2019 Published by Elsevier Ltd.

human-computer interactions [24]. But, the depth precision of ToF is generally not high for short-range inspections as light travels too fast. As another extensively used active methods, the structured light techniques illuminate test scenes with 2D spatially varying intensity pattern. The 3D shape is extracted based on the information from the distortion of captured structured light patterns. Because of the advantages of favorable flexibility and versatility, 3D surface imaging based on the structured light illumination is receiving increasing attention, and becoming more and more important. The commercial success of these techniques includes Microsoft Kinect 1 [25], Intel RealSense [26], Apple iPhone X [27], and OPPO Find X [28]. Owing to advances of intelligent manufacturing, pilotless vehicle, and cloud imaging, the desire to developing real-time ( $\sim 30$  fps) or high-speed ( $> 10,000$  fps) 3D imaging techniques has never been more apparent [29,30].

Rapid developments in high-frame-rate imaging sensors and digital projection technology are providing new avenues for the generation of powerful high-speed 3D surface imaging systems. Compared with high-speed cameras running at tens of thousands fps or even faster, however, projectors normally operate at a much lower rate that is often around 120 fps when gray-scale patterns are projected. Therefore, the defocusing techniques are developed, with which quasi-sinusoidal fringe patterns can be projected at the maximum allowed frame rate (typically more than 1000 fps) with binary dithering techniques and lens-defocused digital light processing projectors [31,32]. Once the limitation of the system hardware is overcome, the major concern focuses on the imaging theory, for which the key is to reduce the number of images required for a single 3D reconstruction. Intuitively, spatial-multiplexing or one-shot techniques, e.g., Fourier transform based profilometry (FT) [33–36], windowed Fourier transform technique (WFT) [37], wavelet transform technique [38], and intensity-correlation-based methods [39,40], are very suitable for scanning moving objects. As the codification can be condensed into a single pattern, these methods have ideal efficiency for high-speed 3D surface imaging. However, their spatial resolution and depth accuracy are not high for discontinuities, e.g., object edges, due to the inherent hypothesis of the continuity and the smoothness for local areas in these methods.

For high-accuracy 3D surface imaging, researchers typically prefer time-multiplexing or multi-shot techniques that can benefit from abundant information collected temporally. Some techniques project many patterns of random intensity to implement active high-speed stereovision 3D measurements [41,42]. However, the 3D reconstructions tend to compromise for rapidly moving objects since a relatively long sequence of images (usually  $> 9$  frames) is required to extract a single 3D frame. In contrast, the phase-shifting profilometry (PSP) [43], which is one of the most widely used multi-shot approaches, can produce accurate 3D reconstructions by projecting a small-scale set of phase-shifting fringe images (minimum three images). Nevertheless, it is still sensitive to motion even with the minimum images. The reason is the object motion violates the nominal phase shifts of the raw fringe patterns, leading to artificial ripples on reconstructed surfaces [44]. Besides, the motivation to remove the phase ambiguity due to the periodic nature of sinusoids is also a challenge for time-critical PSP applications, which can easily double or even triple the size of the image sequence [45].

To reduce the size of the image sequence (captured in the time domain) while collecting comparable amount of information, some researchers suggest strengthening the encoding capability in the space domain. To reduce the images for phase unwrapping, one can have more than one viewpoints, e.g., using more cameras to capture structured-light patterns. Benefiting from the geometric constraint, the methods can discriminate the fringe order without capturing extra images [46–49]. But, the weakness is that the structure of the imaging system would become complex. Also, the cost would increase significantly because of the use of additional high-speed cameras. Alternatively, without resorting to more viewpoints, the spacial coding strategy can also be introduced into the time-multiplexing techniques by condensing two images into a single one or reusing the existing patterns with more than one

purpose [50–54]. These approaches can remove the phase ambiguity without greatly increasing the projected images, but would suffer in the process of phase unwrapping when the projected fringe is very dense [55]. Recently, micro FTP ( $\mu$ FTP) was developed to measure 3D profiles for transient scenes at 10,000 fps [32]. Although the dynamic 3D shapes can be recovered from dense fringe patterns, several uniform images (i.e., pure white images) have to be projected along with the structured-light patterns for robust phase retrieval. Thus, the size of overall image sequence is still relatively large, making the 3D imaging sensitive to fast moving objects.

In this work, we present a novel micro deep learning profilometry ( $\mu$ DLP), which enables high-quality 3D shape reconstructions for transient scenes. The micro means small values for both the frequency variations and periods of fringe patterns, allowing highly-accurate phase measurement and high resistance to the global illumination. Deep learning is a powerful machine learning technique that has shown great success in numerous imaging and computer vision applications [56–61]. Thanks to the strength of machine learning, the proposed method shows superiority in three aspects to the state-of-art methods. The first one is the high efficiency. The phase information can be extracted from a single image via a properly trained neural network. Compared with  $\mu$ FTP, it only uses half of the images to obtain a 3D image. Then, the second advantage is the high-quality phase measurement. As indicated by our experiments, the phase error of  $\mu$ DLP is only one-third of those of FT and WFT and is almost half of that of  $\mu$ FTP. Further, with only three images our method can nearly reproduce the ground-truth 3D result that is calculated with the multi-shot phase-shifting method that uses 36 images. Last, the proposed method is easy to use. Different from Fourier-transform-based methods in which the phase measurement deeply relies on the fine tuning of parameters, e.g., the window size in FT, the sigma, the sampling intervals, and the frequency threshold in WFT, the presented  $\mu$ DLP is fully automatic once the neural network has been trained, which means the exhaustive search for the optimal parameters can be avoided. Experiments demonstrate that  $\mu$ DLP is a powerful high-speed 3D surface imaging approach that can reconstruct high-accuracy 3D shapes for transient scenes at 20,000 fps.

## 2. Theory

### 2.1. Phase retrieval through a deep neural network

In  $\mu$ DLP, the fringe image is captured with a system of structured light illumination, which consists of a projector and a camera typically. According to the schematic shown in Fig. 1, the projector emits a fringe image onto the measured object to encode the illuminated surface. The camera captures the image from a different viewpoint, from which the

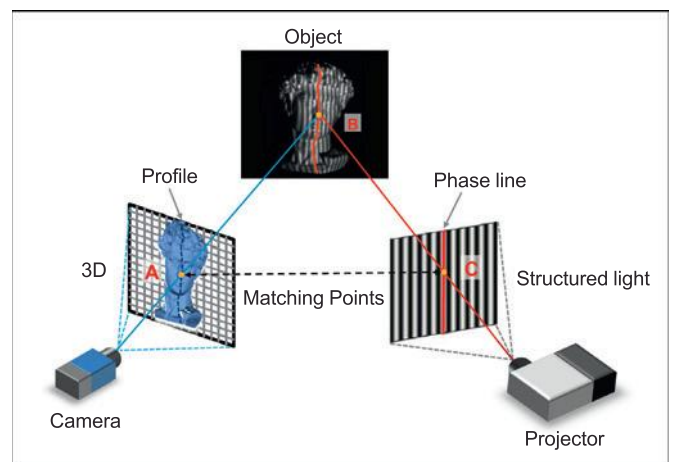
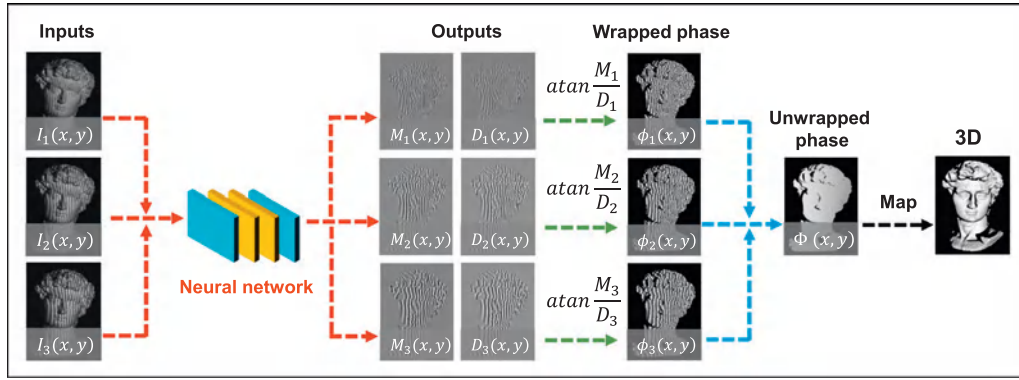


Fig. 1. Schematic of 3D surface imaging by structured light illumination.





**Fig. 2.** Schematic of the proposed  $\mu$ DLP. With a few fringe patterns  $I_1(x, y)$ ,  $I_2(x, y)$ , and  $I_3(x, y)$ , the neural network predicts the numerator  $M_t(x, y)$  and the denominator  $D_t(x, y)$  for each input fringe image. These intermediate results are then fed into the arctangent function to calculate the phase distribution  $\phi_t(x, y)$ . After phase unwrapping, an unwrapped absolute phase map  $\Phi(x, y)$  is obtained and is further converted into the 3D reconstruction.

stripes are observed with distortion due to the depth variation of the object. The phase is then calculated from the captured fringe image, which works as a cue to compute the 3D information.

During the image projection,  $\mu$ DLP exploits several fringe patterns with slightly different wavelengths or fringe pitches  $\{\lambda_1, \lambda_2, \dots, \lambda_T\}$ . For rapid projection, the sinusoidal patterns are generated in the binary mode and are projected by a defocused projector [62,63]. The wavelengths of projected patterns are carefully chosen by considering: First, the selected  $\lambda$  is supposed to be small enough, i.e., the frequency should be sufficiently high for high-quality phase retrieval. Second, the least common multiple (LCM) of the wavelengths should be larger than the horizontal or vertical resolution of the projector so that the phase ambiguity can be removed properly. In this work, we project vertical fringes, which means  $LCM(\lambda_1, \lambda_2, \dots, \lambda_T)$  should be greater than the width of projection plane. With the determined wavelengths, the intensity of projected patterns can be written as

$$I_t^p(x^p, y^p) = a + b \cos\left(\frac{2\pi x^p}{\lambda_t}\right) \quad (1)$$

where  $(x^p, y^p)$  is the pixel coordinate of the projector, and  $t = 1, 2, \dots, T$ . Parameters  $a$  and  $b$  are the mean value and the amplitude, respectively.

Then, the generated patterns are projected and captured sequentially. The intensity of captured images can be represented as

$$I_t(x, y) = A(x, y) + B(x, y) \cos \phi_t(x, y) \quad (2)$$

where  $(x, y)$  is the pixel coordinate of the camera,  $A(x, y)$  the background intensity,  $B(x, y)$  the modulation, and  $\phi_t(x, y)$  the phase to be recovered. In most phase measurement techniques, the wrapped phase map is often retrieved from an inverse trigonometric function:

$$\phi_t(x, y) = \arctan \frac{M_t(x, y)}{D_t(x, y)} = \arctan \frac{cB(x, y) \sin \phi_t(x, y)}{cB(x, y) \cos \phi_t(x, y)} \quad (3)$$

where  $M_t(x, y)$  and  $D_t(x, y)$  denote the numerator and the denominator of the arctan function, respectively.  $c$  is a constant that depends on the phase demodulation algorithm, e.g.,  $c = 0.5$  for FT and  $c = \frac{N}{2}$  for  $N$ -step PSP.

To realize the process of phase retrieval with machine learning, we construct a deep convolutional neural network. As mentioned above, we prefer small sets of fringe images for high-speed 3D surface imaging. However, one or more assistant phase maps are required for robust phase unwrapping of dense fringe pattern [45]. Thus, we have a balance by totally employing three fringe patterns (i.e.,  $T = 3$ ) for 3D imaging that can produce three phase maps, one of which is used for 3D reconstruction and the rest for reliable phase unwrapping. Fig. 2 demonstrates the schematic of the proposed method. The neural network is trained to predict the numerator  $M_t(x, y)$  and the denominator  $D_t(x, y)$  for each input image  $I_t(x, y)$ . Each pair of numerator and denominator  $\{M_t(x,$

$y), D_t(x, y)\}$  is then fed into the arctangent function (Eq. (3)) to obtain the wrapped phase map  $\phi_t(x, y)$ . Next, an unwrapped phase distribution  $\Phi(x, y)$  is obtained by the temporal phase unwrapping algorithm based on projection distance minimization. Finally, the 3D surface is calculated from the absolute phase map with calibrated mapping parameters between the camera and the projector.

Note that we presented a machine-learning-based fringe analysis method [56] that employs two neural networks to calculate the phase information. For applications of transient 3D measurements, some improvements have been made in this work. First,  $\mu$ DLP uses only one network for the phase retrieval, thus easing the learning process and saving the time cost of the training process. To compensate the influence of the absence of the background intensity, a more powerful three-scale data processing architecture is developed here to perceive the surface details and learn the phase extraction. Moreover, the neural network in  $\mu$ DLP can learn fringe patterns of different frequencies simultaneously and output the intermediate results for corresponding fringe patterns, which improves the measurement efficiency of the phase and 3D contours.

Fig. 3 shows the internal structure of the neural network in  $\mu$ DLP. The labeled dimension of each layer or block indicates the size of the output data. The inputs of the network are the fringe images  $\{I_1(x, y), I_2(x, y), I_3(x, y)\}$ . The size of each input image is  $W \times H$  pixels, where  $W$  is the width and  $H$  is the height. Three data-flow paths are constructed to process the input images at different scales. In the first path which keeps the original size of input data, the fringe images are successively processed by a convolutional layer, a group of residual blocks and another convolutional layer.  $C$  is the number of filters used in the convolutional layer and equals the number of channels of output data. Each filter is used to extract a feature map (channel) for the output tensor. The same input data also undergoes similar but more sophisticated procedures in the second and the third paths where the data are first down-sampled by  $\times 2$  and  $\times 4$  for high-level perceptions and then upsampled to match the original dimensions. Eventually, the results of each data-flow path are concatenated to produce the final outputs that feature three pairs of  $\{M(x, y), D(x, y)\}$  corresponding to every input image  $I_t(x, y)$ . With the design of multi-scale data-flow paths, geometric details that the input images contain can be perceived precisely, ensuring the estimation of high-quality phase information. Note that it is difficult to output the wrapped phase directly with the input of the fringe image, since the sharp discontinuity at the  $2\pi$  jump is hard to learn by the neural network. Therefore, for high-accurate phase estimations, the deep neural network is trained to calculate the intermediate results that vary continuously in space, i.e., the numerator and the denominator. Further details about the architecture of the network are provided in Appendix A.

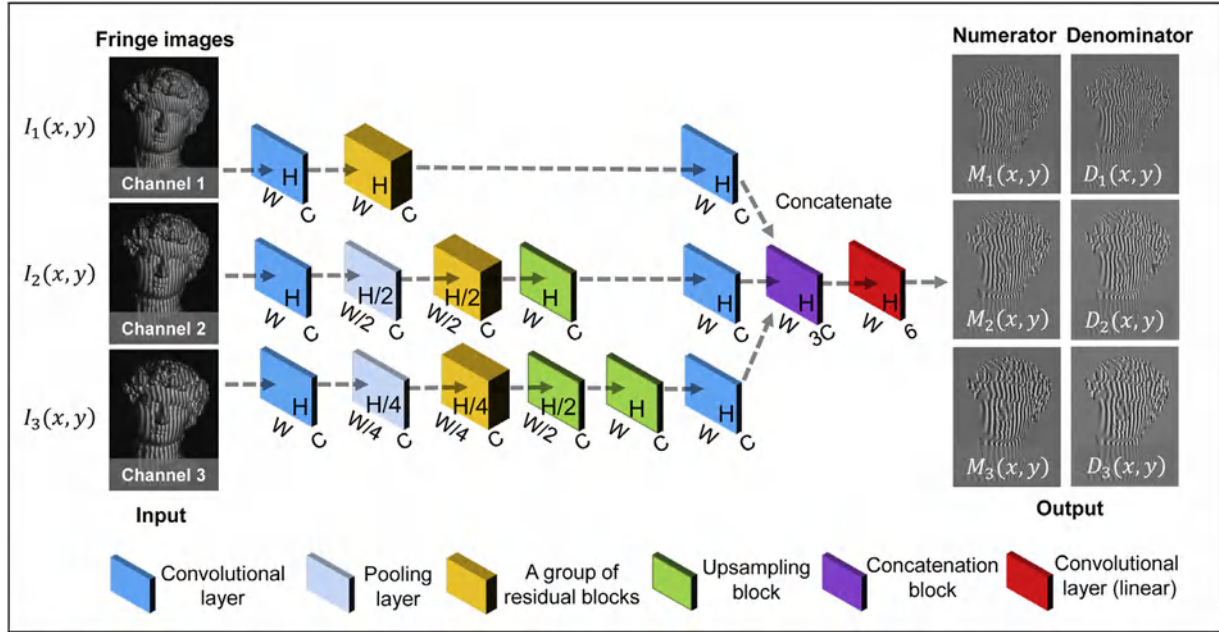


Fig. 3. Architecture of the proposed multi-scale deep neural network. The input data have three channels containing the three fringe images. The neural network has three data-flow paths that involve different kinds of layers/blocks, which can process the input data at different scales and extract useful information with downsampling rates of  $\times 1$ ,  $\times 2$  and  $\times 4$ , respectively. The outputs of the network are three pairs of numerator and denominator that correspond to each fringe pattern.

### 2.2. Phase unwrapping and 3D reconstruction

After feeding the estimated pair of numerator and denominator into Eq. (3),  $\mu$ DLP calculates wrapped phase maps  $\phi_t(x, y)$  for each input fringe image. To remove the phase discontinuity of  $\phi_t(x, y)$ , we use the temporal phase unwrapping approach based on the projection distance minimization [32]. Given a vector of wrapped phase  $\varphi = (\phi_1, \phi_2, \dots, \phi_T)^{Trs}$  of the pixel  $(x, y)$ , where  $Trs$  means the transposition, the vector of corresponding unwrapped phase  $\Phi = (\Phi_1, \Phi_2, \dots, \Phi_T)^{Trs}$  can be expressed as

$$\Phi = \varphi + 2\pi\mathbf{k} \quad (4)$$

where  $\mathbf{k} = (k_1, k_2, \dots, k_T)^{Trs}$  is the vector of integer fringe order that we calculate for phase unwrapping. By taking the wavelengths into account, we have the following relationship

$$\Phi_1 \lambda_1 = \Phi_2 \lambda_2 = \dots = \Phi_T \lambda_T \quad (5)$$

Eq. (5) reveals that the unwrapped phase  $\Phi = (\Phi_1, \Phi_2, \dots, \Phi_T)^{Trs}$  of each pixel forms a line in space  $R^T$ . Therefore, with the mentioned constraint that  $LCM(\lambda_1, \lambda_2, \dots, \lambda_T) > W^p$  where  $W^p$  is the width of projection plane in pixel, there will be a unique qualified fringe order vector  $\mathbf{k}$  that corresponds to the measurement range. In theory, the unwrapped phase  $\Phi$  of each pixel would align perfectly along the line expressed by Eq. (5). However, the unwrapped phase often scatters around the line due to the effects of random noise and non-sinusoidal fringe intensity in reality. Therefore, the distance between each candidate unwrapped phase and its projection onto this line is calculated. The desired  $\Phi = (\Phi_1, \Phi_2, \dots, \Phi_T)^{Trs}$  is determined when the distance is minimized.

As a group of unwrapped phase maps is obtained after phase unwrapping, one of them is selected as  $\Phi(x, y)$  for the 3D reconstruction. In the perspective of the camera, given the point  $(x^w, y^w, z^w)$  of test object is imaged by pixel  $(x, y)$ , we have the following projection relationship in

homogeneous coordinates

$$s^c \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = P^c \begin{pmatrix} x^w \\ y^w \\ z^w \\ 1 \end{pmatrix} = \begin{pmatrix} p_{11}^c & p_{12}^c & p_{13}^c & p_{14}^c \\ p_{21}^c & p_{22}^c & p_{23}^c & p_{24}^c \\ p_{31}^c & p_{32}^c & p_{33}^c & p_{34}^c \end{pmatrix} \begin{pmatrix} x^w \\ y^w \\ z^w \\ 1 \end{pmatrix} \quad (6)$$

where  $s^c$  is a scaling factor,  $P^c$  is the projection matrix of camera that is the product of the extrinsic parameter matrix and the intrinsic parameter matrix of the camera. In the other perspective of projector, there is a similar process when the projector is considered as an inverse camera

$$s^p \begin{pmatrix} x^p \\ y^p \\ 1 \end{pmatrix} = P^p \begin{pmatrix} x^w \\ y^w \\ z^w \\ 1 \end{pmatrix} = \begin{pmatrix} p_{11}^p & p_{12}^p & p_{13}^p & p_{14}^p \\ p_{21}^p & p_{22}^p & p_{23}^p & p_{24}^p \\ p_{31}^p & p_{32}^p & p_{33}^p & p_{34}^p \end{pmatrix} \begin{pmatrix} x^w \\ y^w \\ z^w \\ 1 \end{pmatrix} \quad (7)$$

where  $s^p$  is a scaling factor,  $P^p$  is the projection matrix of projector that is the product of the extrinsic parameter matrix and the intrinsic parameter matrix of the projector. Given the unwrapped phase of this pixel is  $\Phi$ , the relationship between the camera pixel and its corresponding projector pixel can be expressed by

$$\Phi(x, y) = \frac{2\pi}{\lambda} x^p \quad (8)$$

Thus, the 3D coordinate can be calculated by combing Eqs. (6) and (7), giving

$$\begin{pmatrix} x^w \\ y^w \\ z^w \end{pmatrix} = \begin{pmatrix} p_{11}^c - p_{31}^c x & p_{12}^c - p_{32}^c x & p_{13}^c - p_{33}^c x \\ p_{21}^c - p_{31}^c y & p_{22}^c - p_{32}^c y & p_{23}^c - p_{32}^c y \\ p_{11}^c - p_{31}^c x^p & p_{12}^c - p_{32}^c x^p & p_{13}^c - p_{33}^c x^p \end{pmatrix}^{-1} \begin{pmatrix} p_{34}^c x - p_{14}^c \\ p_{34}^c y - p_{24}^c \\ p_{34}^c x^p - p_{14}^c \end{pmatrix} \quad (9)$$

The projection matrices of the camera and the projector can be obtained with the system calibration [46]. Note that gigabyte-scale image data are often recorded in applications of high-speed imaging. Although the 3D reconstruction can be carried out off-line, the time cost would be still very high. To increase the calculation speed, we suggest Eq. (9) to be implemented with a graphics processing unit [64] or several look-up tables [65], which can greatly save the time cost of the 3D reconstruction.

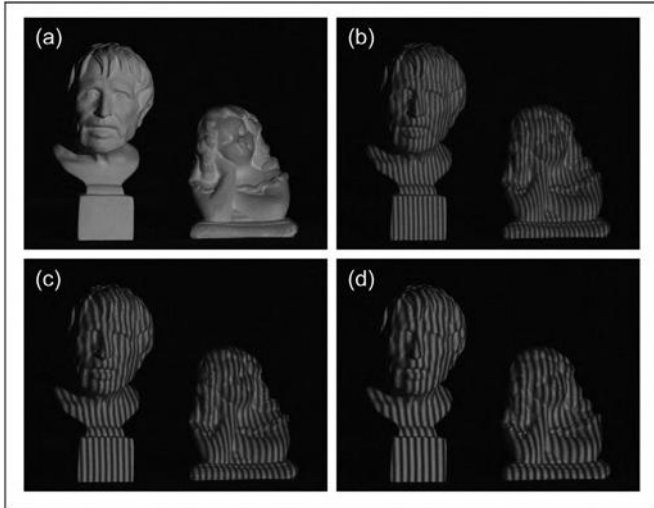


Fig. 4. Testing the trained network using a scene which is not present in the training phase. (a) The measured scene; (b) fringe image  $I_1(x, y)$  with  $\lambda_1 = 9$ ; (c) fringe image  $I_2(x, y)$  with  $\lambda_2 = 11$ ; (d) fringe image  $I_3(x, y)$  with  $\lambda_3 = 13$ .

### 3. Experiments

To validate the proposed method, we built a structured light illumination system that consisted of a projector (DLP 4100, Texas Instruments) with resolution of  $1024 \times 768$  and a high-speed camera (V611, Vision Research Phantom) with resolution of  $640 \times 440$  and with pixel depth of 8 bits. The camera equipped with a lens of 24 mm focal length. The distance between the test object and the imaging system was about 1.5 m. The wavelengths of projected images were selected as  $\{\lambda_1 = 9, \lambda_2 = 11, \lambda_3 = 13\}$ , which provided unambiguous 3D reconstructions for the whole projection range (i.e.,  $LCM(9, 11, 13) = 1287 > 1024$ ).

The implementation of  $\mu$ DLP has two steps: training and testing. In the training stage, the training data were collected from different scenes. Analogous to traditional approaches of structured light illumination that require fringes with enough signal-to-noise ratio or without saturated pixels,  $\mu$ DLP also prefers the training objects without very dark or shiny surfaces. Otherwise, the training process would be damaged, since it is

hard to obtain reliable ground truth data for these objects. Here, our training data set was collected from 45 scenes. With the 12-step phase-shifting method, we captured 1620 different fringe patterns and their corresponding ground-truth data for each wavelength (see Appendix B for more details on the collection of the training data). The neural network was implemented using TensorFlow framework (Google) and was computed on a GTX Titan graphics card (NVIDIA). To monitor during training the accuracy of the neural network on the data that it has never seen before, we created a validation set including 120 fringe images from 10 validation scenes which were separate from the training scenarios. With 120 epochs of training, the training loss and the validation loss of the network converged. And there is not overfitting to our training dataset. We provide further details of the training results in Appendix A.

#### 3.1. The performance of $\mu$ DLP for static scene

To test the performance of the trained neural network, we measured a static scenario that includes two isolated plaster models, as shown in Fig. 4(a). Note that our neural network never sees these models in the training stage. Fig. 4(b)–(d) are the captured fringe images  $I_1(x, y)$ ,  $I_2(x, y)$ , and  $I_3(x, y)$ , respectively. With these images, the trained neural network predicted the numerator and the denominator for each of the input fringe image. The results are shown in the first two columns of Fig. 5. The estimated numerators and denominators were then fed into Eq. (3) to calculate the wrapped phase maps that are shown in the third column of Fig. 5. Finally, we calculated the unwrapped phase distributions that are displayed in the last column of Fig. 5. As we can see, the discontinuity have been removed completely for all of the wrapped phase.

We chose one of the unwrapped phase maps, i.e.,  $\Phi_2(x, y)$ , to investigate the quality of the phase estimated by  $\mu$ DLP. In the investigation, 12-step phase-shifting method was used to calculate a reference phase map which was unwrapped in the same way. Moreover, we also applied FT, WFT, and  $\mu$ FTP for comparison. Fig. 6 shows the phase error of each method. We can see the errors of WFT and FT are more significant than those of  $\mu$ FTP and  $\mu$ DLP. Further,  $\mu$ DLP shows better performance than  $\mu$ FTP due to less phase errors observed at the object edges. To compare the error maps in detail, we studied two recovered areas of complex surfaces, as can be seen in Fig. 7. The selected regions are the hair of the left model and the face of the right one. These two regions of interest (ROI) have rich details, which can be used to evaluate the capability of han-

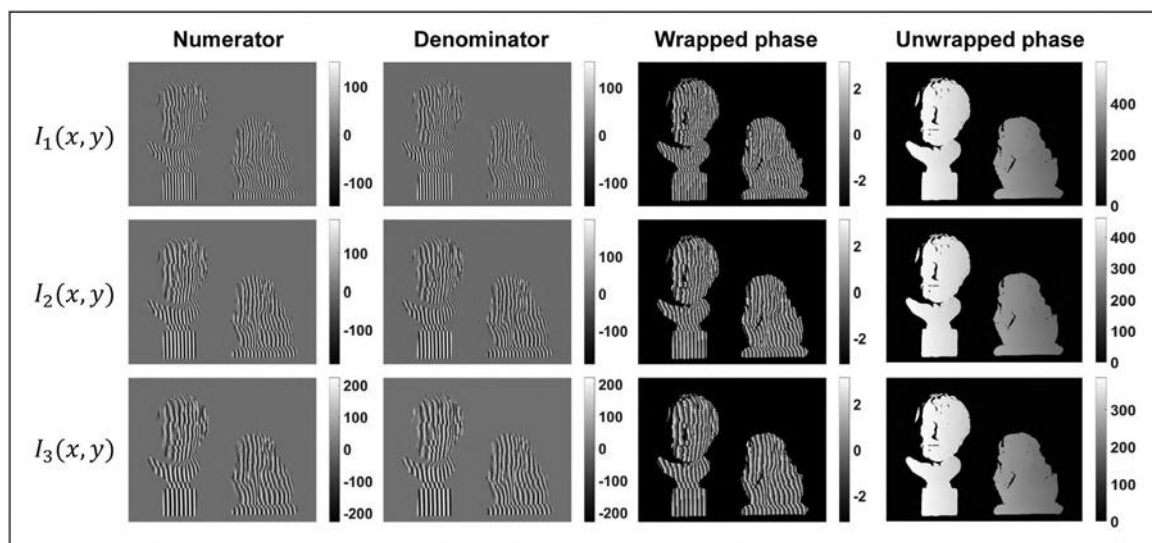


Fig. 5. Predicted results of the trained neural network. Each row shows the estimated numerator, denominator, wrapped phase, and unwrapped phase for each fringe image.



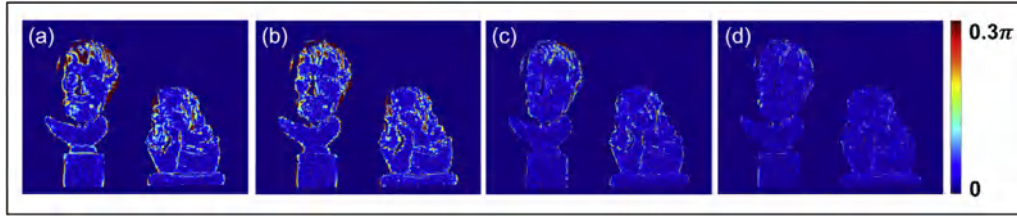


Fig. 6. Comparison of the phase error distribution for methods: (a) WFT, (b) FT, (c)  $\mu$ FTP, and (d)  $\mu$ DLP.

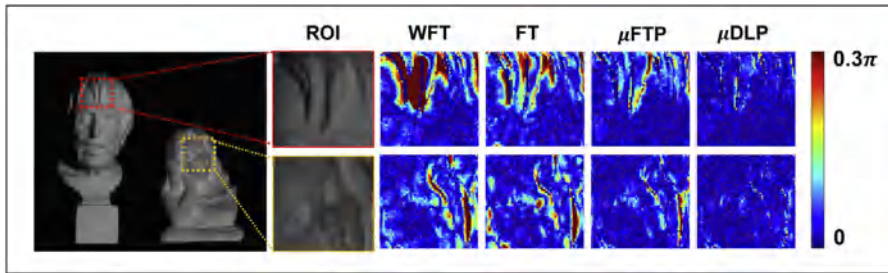


Fig. 7. Comparison of the phase error of two ROI. The first ROI is selected from the hair of the left model, and the second is picked from the face of the right model. The zoom-in phase error of different approaches are demonstrated for each region.

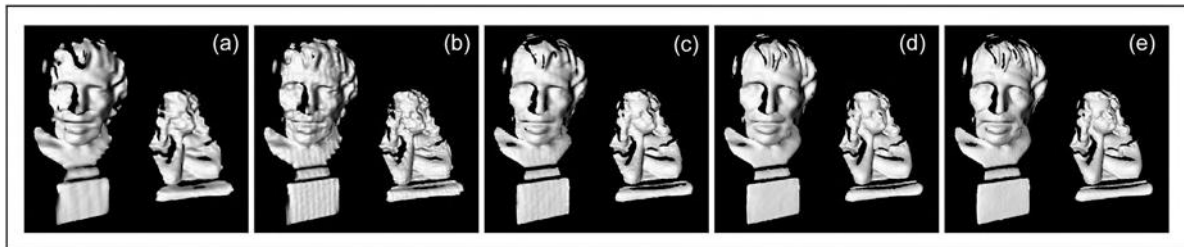


Fig. 8. 3D reconstructions of the methods: (a) WFT, (b) FT, (c)  $\mu$ FTP, (d)  $\mu$ DLP, and (e) 12-step phase-shifting method (ground truth).

Table 1

Quantitative comparison of the proposed  $\mu$ DLP with WFT, FT and  $\mu$ FTP in terms of MAE of unwrapped phase and the required number of images.

	WFT	FT	$\mu$ FTP	$\mu$ DLP
MAE (rad)	0.36	0.26	0.13	0.077
Images	3	3	6	3

dling profiles with fine structures. In Fig. 7, we can observe WFT has the largest phase error, especially for the region of hair. By contrast, FT performed better than WFT as there are less errors at the reconstructed hair. But, it still failed to accurately retrieve the phase of the facial contour of the right model. In contrast to WFT and FT,  $\mu$ FTP shows increased but yet not high enough accuracy for these areas. As to  $\mu$ DLP, it has the least phase errors for both the hair of the left model and the details of the face of the right one. For quantitative evaluation, the mean absolute error (MAE) of unwrapped phase and the number of used images for the phase retrieval are shown in Table 1. Although the same images are used, the error of  $\mu$ DLP is smaller than one-third of those of WFT and FT. Compared with  $\mu$ FTP,  $\mu$ DLP only exploited half of the patterns while improved the phase accuracy by almost 50%.

Further, we converted the unwrapped phase maps into 3D rendered geometries, as shown in Fig. 8. Also, several ROI were selected for the detailed comparison. Fig. 9 shows the enlarged views of reconstructions of the face and the pedestal of the left model, and the face and the arms of the right model. From the result of WFT, the general profiles of these regions have been recovered but with significant loss of details compared with the reference that was reconstructed by 12-step phase-

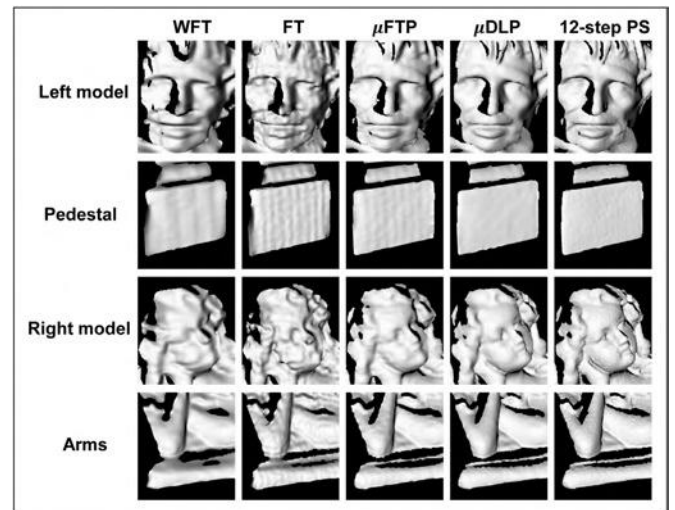
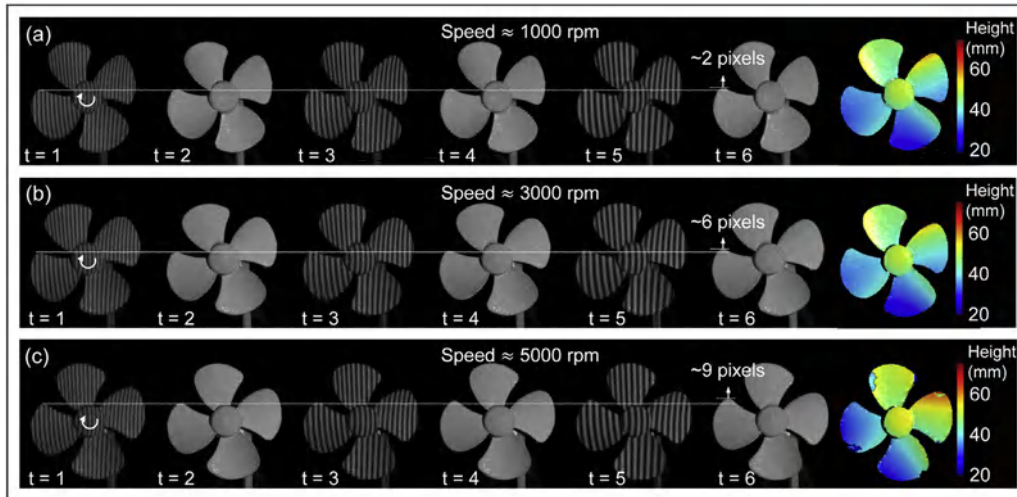


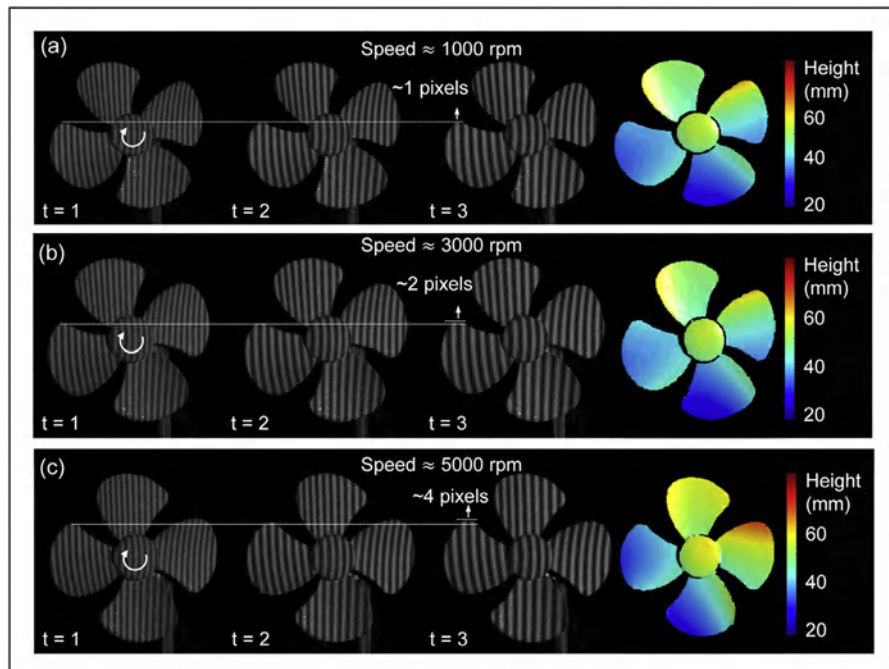
Fig. 9. Amplified views of the 3D reconstructions of four ROI: The face of the left model, the pedestal of the left model, the face of the right model, and the arms of the right model.

shifting method. From the 3D reconstruction of FT, the result features many grainy distortions that are mainly due to the inevitable spectra leakage and overlapping in the frequency domain. In contrast to WFT and FT,  $\mu$ FTP successfully retrieved some fine structures, e.g., the nose and the mouse of the right model. But, it still failed to preserve a few sharp edges. Finally, from the result of our method, we can see the deep-learning based approach yielded the highest-quality 3D reconstruction,





**Fig. 10.** 3D surface imaging of an electric fan rotating at different speeds by  $\mu$ FTP. (a)–(c) Images captured at 1000 rpm, 3000 rpm, and 5000 rpm with their corresponding 3D reconstructions.



**Fig. 11.** 3D surface imaging of an electric fan rotating at different speeds by the proposed  $\mu$ DLP. (a)–(c) Images captured at 1000 rpm, 3000 rpm, and 5000 rpm with their corresponding 3D reconstructions.

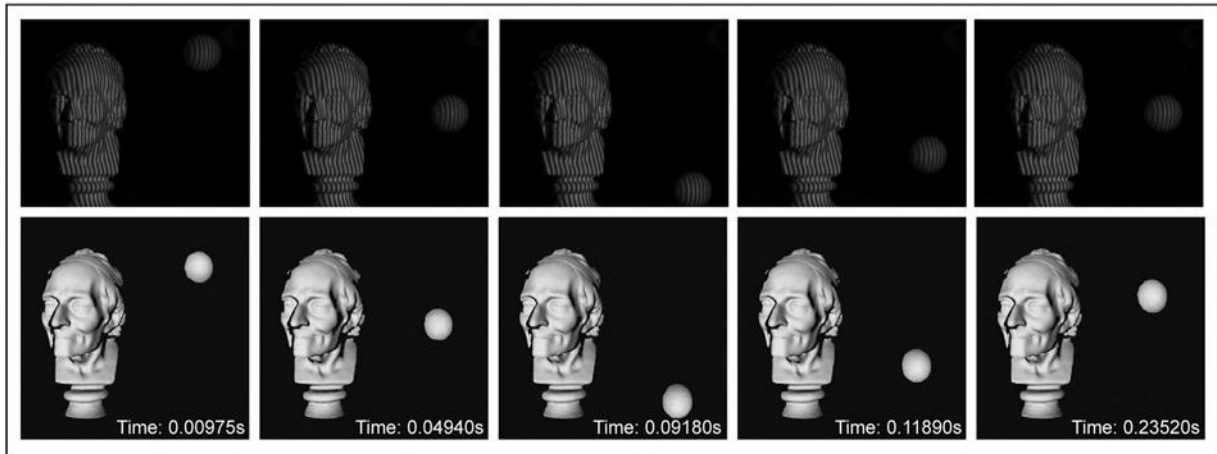
which almost reproduced the reference 3D model. It is worthwhile to mention that only three images were used in our method while  $12 \times 3$  images were employed by the 12-step phase-shifting method. This experiment verifies that  $\mu$ DLP can produce high-fidelity phase measurements and 3D reconstructions, and is superior to the state-of-art high-speed 3D surface imaging approaches regarding the accuracy and efficiency.

### 3.2. The performance of $\mu$ DLP for dynamic scene

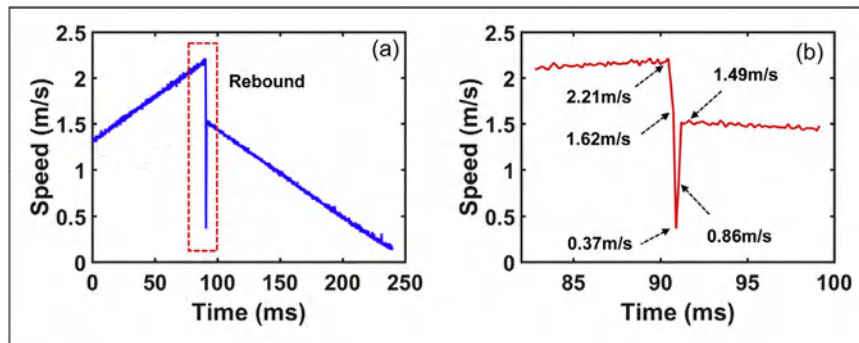
We measured an electric fan rotating at a high speed to show  $\mu$ DLP's performance of handling fast rotating objects. The radius of the fan is about 50 mm. For comparison, we also used the  $\mu$ FTP to test the same scene. By tuning the input current (from 0.3A to 5A), we let the fan rotate from 1000 rotations per minute (rpm) to 5000 rpm. Fig. 10 shows the images captured by  $\mu$ FTP and the corresponding surface reconstructions when the fan rotated at about 1000 rpm, 3000 rpm, and 5000 rpm, respectively. During the tests, the fan rotated clockwise, and the sys-

tem kept capturing the images at 20,000 fps for both approaches. As the phase information was extracted from a pair of images (a fringe image and a plain image) in  $\mu$ FTP, it reconstructed the 3D surface at 10,000 fps. In Fig. 10, we can observe that within a period of 3D reconstruction the left blade shifted upward about two and six pixels respectively with the rotating speed of 1000 rpm and 3000 rpm. Under these conditions,  $\mu$ FTP successfully measured the contour of the blades. However, when the fan accelerated to 5000 rpm, several areas were retrieved with many errors as can be observed from Fig. 10(c). The reason lies in the fact that  $\mu$ FTP exploited six images to reconstruct a single 3D frame. When the speed reached up to 5000 rpm, the left blade moved  $\sim 9$  pixels during the capture of the six images. Because of the long period of the image capture, the 3D reconstruction becomes fragile for the object motion.

By contrast,  $\mu$ DLP can reconstruct 3D shapes at 20,000 fps with the fact that the height-related phase was measured from a single fringe image. Fig. 11 shows the captured images and the corresponding recovered 3D results of  $\mu$ DLP. Although the speed increased to 5000 rpm, our



**Fig. 12.** Measurement of a dynamic scene that includes a static model and a falling table tennis, which are also not present in the training process. The first row shows captured fringe images at five different moments, and the second the corresponding 3D reconstructions obtained through  $\mu$ DLP.



**Fig. 13.** Investigation of the speed of the table tennis. (a) The speed of the table tennis during the fall; (b) the amplified view of the red box in (a) showing the change of speed at the moment when the sphere hit the ground.

method can still measure the surface robustly. As fewer images were used by  $\mu$ DLP, the motion caused a shift of merely about 4 pixels as can be seen in Fig 11 (c), which did not affect the 3D reconstruction. From this experiment, thanks to the powerful computational capability of machine learning, the number of images can decrease significantly, which is favorable for overcoming the influence of object motion and dealing with fast moving objects.

Then, another dynamic scene was measured to further validate  $\mu$ DLP's capability of handling transient events. The scene consisted of a static plaster model and a falling table tennis. During the measurement, the fringe patterns were projected repeatedly onto the scene and the camera was synchronized with the projector at 20,000 fps. The first row of Fig. 12 shows the captured fringe images  $I_2(x, y)$  at five different moments. We can see in this transient process the table tennis gradually fell to the lowest point, and then bounced after hitting the ground. The dynamic process was retrieved by  $\mu$ DLP and is shown in Visualization 1. The 3D images corresponding to the selected moments are displayed in the second row of Fig. 12. We can observe that both the static model and the dropping sphere have been faithfully reconstructed with the deep-learning based technique.

Further, we analyzed the velocity of the falling table tennis using the retrieved geometry. First, the 3D point cloud of the table tennis was fitted to the function of sphere. Then, we estimated the center of the sphere, and calculated the speed by computing the displacement of the center between successive 3D frames. The velocity of the sphere during this transient event is shown in Fig. 13(a). As the measurement just started after the fall, the table tennis had an initial velocity which is about 1.36 m/s. As time went on, it moved faster due to the acceleration of gravity. When the velocity reached the maximum, the sphere hit the ground. Fig. 13(b) shows the speed of the sphere before and after the rebound. We can see the table tennis had the maximum ve-

locity of 2.21 m/s before the hitting the ground. The speed began to decrease sharply right away after the hit. Within about one millisecond, the velocity reduced to 1.62 m/s and 0.37 m/s. Then, the speed went up instantly to 1.49 m/s due to the elastic potential energy. We can see the speed at this moment is smaller than the previous maximum velocity. The reason could be the fact that some of the energy was consumed to overcome the damping effect during the energy conversion. Next, the table tennis gradually raised but with a diminishing speed until it reached a point where the velocity came close to zero. From the overall process, we can see it happened in less than 0.25 s. Although the time period is very short,  $\mu$ DLP reconstructed the 3D shape of the falling sphere accurately and analyzed the velocity successfully with the geometry information. This experiment demonstrates that  $\mu$ DLP can not only reconstruct 3D shapes of the dynamic objects but also be applied to the study of some key physical quantities of the transient events.

### 3.3. Quantitative evaluation of 3D reconstruction accuracy

Last but not least, we measured a pair of gauge spheres made from ceramic to demonstrate the accuracy of 3D reconstruction quantitatively. The shape of the gauge spheres have been calibrated by a coordinate measurement machine. Fig. 14(a) shows the tested spheres whose radii are 25.398 mm and 25.403 mm, respectively, and their center-to-center distance is 100.069 mm. With the proposed method, we computed the 3D point cloud and fitted the 3D points into the sphere model. The reconstructed result is shown in Fig. 14(b), where the "jet" colormap is used to represent data values of reconstruction errors. The radii of reconstructed spheres are 25.449 mm and 25.470 mm, with the deviations of 0.051 mm and 0.067 mm respectively. The measured center-to-center distance is 100.134 mm with the error of 0.065 mm. Further, Figs. 14(c)

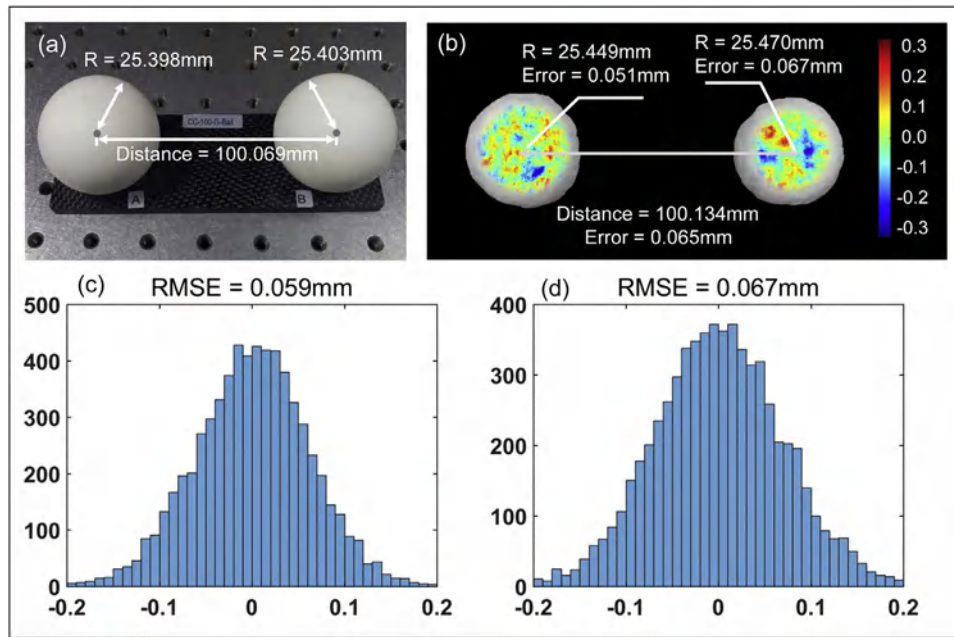


Fig. 14. Quantitative analysis of the reconstruction accuracy of  $\mu$ DLP. (a) Measured objects: a pair of gauge spheres; (b) 3D reconstruction with accuracy analysis; (c) histogram of the 3D error of sphere A; (d) histogram of the 3D error of sphere B.

and 14(d) show that the root-mean-square error (RMSE) of the spheres are 0.059 mm and 0.067 mm respectively. Since the measured shapes are very close to the ground truth, this experiment validates that our method can provide reliable phase information as well as high-accuracy 3D measurements.

#### 4. Conclusion

In this work, we present a novel high-speed 3D surface imaging approach  $\mu$ DLP that can reconstruct dense and precise 3D shapes of transient events. Different from most of fast 3D imaging techniques using structured light illumination,  $\mu$ DLP can extract phase information from a single fringe image through a properly trained deep neural network. With only several fringe images of slightly different wavelengths, unambiguous high-quality 3D reconstructions can be obtained.

$\mu$ DLP has three major advantages over the existing high-speed 3D imaging techniques. The first one is the high-accuracy phase retrieval. From our experiment, the phase error of  $\mu$ DLP is smaller than one-third of those of FT and WFT, and is almost half of that of  $\mu$ FTP. Moreover,  $\mu$ DLP can preserve details for fine structures or edges of test objects, resulting in the 3D reconstruction that is even comparable to that of 12-step phase-shifting method. Next, the second advantage of  $\mu$ DLP is the high efficiency. According to experimental results,  $\mu$ DLP leveraged only half of the patterns of  $\mu$ FTP but achieved nearly doubled phase precision. Also,  $\mu$ DLP used only three images to produce a high-quality 3D reconstruction that is close to that of 12-step phase-shifting method, by which, however, 36 fringe images were employed. Last,  $\mu$ DLP is easy to implement. Unlike the approaches based on Fourier transform, the performance of which heavily relies on tuning parameters, e.g., the window size for FT, the sigma, the sampling intervals, and the frequency threshold for WFT,  $\mu$ DLP is fully automatic and does not require a manual parameter search to optimize its performance once the neural network has been trained. Owing to these merits,  $\mu$ DLP can faithfully reconstruct 3D shapes of fast moving objects at 20,000 fps as demonstrated by the experimental result. The rate of 3D reconstruction can be further increased once more powerful equipment is in use. We believe the proposed  $\mu$ DLP could narrow the gap between the high-speed 3D imaging and the high-rate 2D photography, providing new insights for extensive studies and applications.

#### Funding

National Natural Science Foundation of China (61705105, 61722506, 11574152), National Key R&D Program of China (2017YFF0106403), Final Assembly “13th Five-Year Plan” Advanced Research Project of China (30102070102), Equipment Advanced Research Fund of China (61404150202), The Key Research and Development Program of Jiangsu Province (BE2017162), Outstanding Youth Foundation of Jiangsu Province (BK20170034), National Defense Science and Technology Foundation of China (0106173), “Six Talent Peaks” project of Jiangsu Province (2015-DZXX-009), “333 Engineering” Research Project of Jiangsu Province (BRA2016407, BRA2015294), Fundamental Research Funds for the Central Universities (30917011204, 30916011322, 30919011222).

#### Appendix A. Architecture and training of the neural network

The input fringe patterns are handled by three different data-flow paths, as demonstrated in Fig. 3. In the first path which keeps the original size of input data, the fringe images are successively processed by a convolutional layer, a group of residual blocks and another convolutional layer. Meanwhile, the same input data undergoes similar but more sophisticated procedures in the second and the third paths where the data are first downsampled by  $\times 2$  and  $\times 4$  for high-level perceptions and then upsampled to match the original dimensions. The downsampling is achieved through a max-pooling layer [66]. For each channel of the input, the pooling layer finds the maximum value in a  $2 \times 2$  or  $4 \times 4$  neighborhood. It then replaces the pixels in the  $2 \times 2$  or  $4 \times 4$  window with the found pixel of the maximum value. Therefore, the size of output is reduced by half/quarter for both the height and the width.

In the convolutional layers, the kernel size is  $3 \times 3$  and the convolution stride is one. Zero-padding is used to control the spatial size of the output data, so that the input and output height and width are the same. The output of the convolutional layer is a three-dimensional (3D) tensor of shape  $(H, W, C)$ , where  $H$  and  $W$  are the height and width in pixels of the input fringe pattern.  $C$  is the number of filters used in the convolutional layer and equals the number of channels of output data. Each filter is used to extract a feature map (channel) for the output tensor. Therefore with more filters, the convolutional network can perceive more details of measured surfaces. But the cost is that the network will

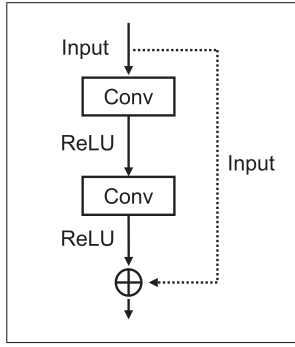


Fig. A1. Architecture of the residual block.

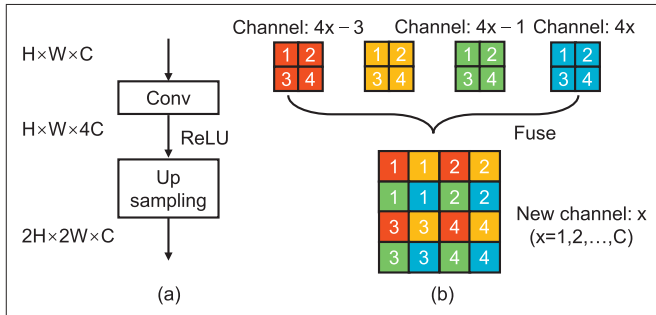


Fig. A2. (a) Architecture of the upsampling block; (b) diagram of the upsampling process.

consume more time during training. Thus, we have  $C = 50$  filters in the work to achieve a balance. Except for the last convolutional layer which is activated linearly, the rest ones use the rectified linear unit (ReLU) as activation function, i.e.,  $ReLU(x) = \max(0, x)$ . Compared with other activation functions, e.g., sigmoid function [67], it has been demonstrated to enable better training of deeper networks [68].

In our network, we also used residual blocks whose architecture is shown in Fig. A1. The residual framework is composed of 2 sets of convolutional layer (Conv) activated by ReLU stacked one above the other [69]. It creates a shortcut between the input and output and can solve the degradation of accuracy as the network becomes deeper, thus easing the training process. To match the dimension of the original image, we upsample the output data from residual blocks using the upsampling block as shown in Fig. A2(a). The data first passes through a convolutional layer with ReLU activation. We then use quadruple filters to extract features from the input for providing rich information for the following upsampling, whose schematic is shown in Fig. A2(b). For the upsampled channel  $x$ , it is generated by original channels from  $4x - 3$  to  $4x$ , thus allowing the output data with  $\times 2$  spatial resolution. Next, the outputs of these three data flow paths are concatenated into a tensor with triple channels. Finally, the last convolutional layer yields a six-channel output datum which consists of three pairs of numerator  $M(x, y)$  and denominator  $D(x, y)$ . The reason why we have the last convolutional layer to be linear is that the neural network is trained to predict the numerator and the denominator which can be negative.

To train the network, we minimize the mean-squared-errors of the output numerators and the output denominators with respect to the ground truth, which are obtained using the 12-step phase-shifting algorithm. The parameters of the network, i.e., the weights, bias and convolutional kernels, are trained using the backpropagation [70]. Thus, the loss function is computed as

$$Loss(\theta) = \frac{1}{H \times W} \sum_{t=1}^3 \left( \|Y_t^M(\theta) - G_t^M\|^2 + \|Y_t^D(\theta) - G_t^D\|^2 \right) \quad (A1)$$

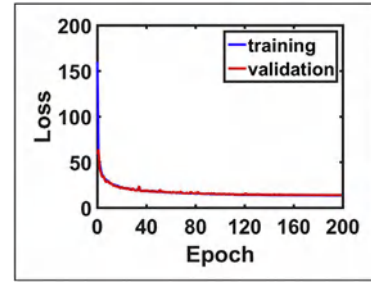


Fig. A3. Loss curve of the training and validation set for the neural network.

where  $G_t^M$  and  $G_t^D$  are the ground-truth numerator and denominator for the input fringe image  $I_t$ .  $Y_t^M(\theta)$  and  $Y_t^D(\theta)$  the numerator and denominator predicted by the network with the parameter space  $\theta$  that includes the weights, bias and convolutional kernels.

During the training, the network uses the score of loss function as a feedback signal to adjust the parameters in  $\theta$  by a little bit, in a direction that would lower the loss score. To this end, the adaptive moment estimation (ADAM) is used in our networks to tune the parameters for finding the minimum of the loss function [71]. In the implementation of ADAM, we start the training with a learning rate of  $10^{-4}$ . We drop it by a factor of 2 if the validation loss has stopped improving for 10 epochs, which helps the loss function get out of local minima during training. To characterize the training, we plot the progression of the training and validation loss over training epochs, i.e., the number of iterations in the backpropagation over all of the dataset. Fig. A3 shows the loss curves converge after 120 epochs. From both curves, we can see there is not overfitting to our training dataset. As to the time cost, the training over 200 epochs took 3.16 hours.

## Appendix B. Collection of training data

Prior to practical measurements, the developed neural network needs a training process in which the network learns to retrieve the phase. To obtain the ground-truth data used to train the neural network, we exploit the  $N$ -step phase-shifting method as it allows precise phase measurements. With this method, the captured phase-shifted fringe patterns with different wavelengths can be written as

$$I_n^t(x, y) = A(x, y) + B(x, y) \cos [\phi_t(x, y) - \delta_n] \quad (B1)$$

where  $n = 0, 1, \dots, N - 1$  indicates the step of phase shift, and  $t = 1, 2, 3$  implies the used wavelengths.  $\delta_n$  is the phase shift that equals  $\frac{2\pi n}{N}$ . With the least square method, the ground-truth phase can be calculated by

$$\phi_t(x, y) = \arctan \frac{\sum_{n=0}^{N-1} I_n^t(x, y) \sin \delta_n}{\sum_{n=0}^{N-1} I_n^t(x, y) \cos \delta_n} \quad (B2)$$

According to Eq. (B2), the numerator and the denominator can be expressed as

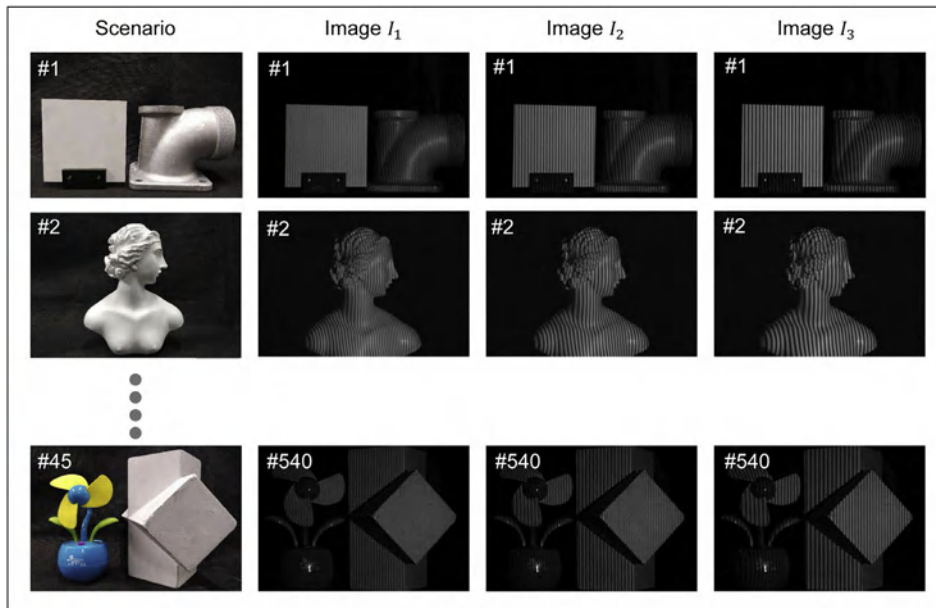
$$M_t(x, y) = \sum_{n=0}^{N-1} I_n^t(x, y) \sin \delta_n \quad (B3)$$

$$D_t(x, y) = \sum_{n=0}^{N-1} I_n^t(x, y) \cos \delta_n \quad (B4)$$

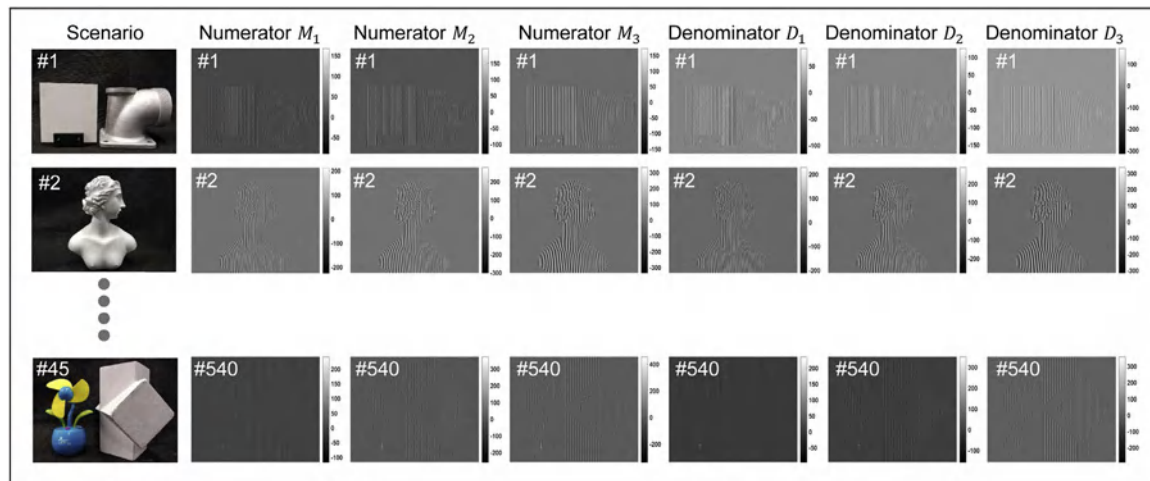
Equations (B3) and (B4) are used to calculate the ground-truth numerator and denominator that are exploited to train the neural network.

In our experiments, three sets of 12-step phase-shifting fringe patterns with wavelengths  $\{\lambda_1 = 9, \lambda_2 = 11, \lambda_3 = 13\}$  were generated according to Eq. (B1). These patterns were then projected onto different measured objects. The camera captured the reflected fringe patterns simultaneously at a different viewpoint and transferred them to our computer. In our experiment, we collected the training data from 45 different scenes including simple and complex objects. For each scene, we recorded  $12 \times 3$  phase-shifting fringe patterns. Thus, 1620 fringe images were collected for all of the scenes. The captured training data are





**Fig. B1.** The collected training data. The first column shows different tested scenarios. For each of them, we captured three sets of 12 phase-shifting fringe patterns and totally obtained 540 training input images for fringe images with three different wavelengths, as demonstrated in the second to the fourth column.



**Fig. B2.** Ground truth of the collected training data. The first column shows the tested scenarios. Within each set of fringe patterns of the same wavelengths, we calculated the ground-truth numerator and denominator by the 12-step phase-shifting algorithm. The second to the fourth columns displays the ground-truth numerator computed through Eq. (B3). The fifth to the seventh column shows the ground-truth denominator obtained through Eq. (B4).

demonstrated in Fig. B1. The first column shows the measured scenes. The second to the fourth column shows the captured fringe images with different wavelengths, respectively. Within each set of fringe patterns of the same wavelength, we calculated the corresponding ground-truth data by the 12-step phase-shifting algorithm. The results are shown in Fig. B2, where the second to the fourth column displays the ground-truth numerator, and the fifth to the seventh column shows the ground-truth denominator. It is noted that before being fed into the networks, the raw fringe images  $\{I_1(x, y), I_2(x, y), I_3(x, y)\}$  were divided by 255 for normalization, which can make the learning process easier for the network. Moreover, for a preferable selection of training objects, one is suggested choosing objects without very dark or shiny surfaces to ensure captured fringe images with enough signal-to-noise ratio or without saturated points.

### Supplementary material

Supplementary material associated with this article can be found, in the online version, at [10.1016/j.optlaseng.2019.04.020](https://doi.org/10.1016/j.optlaseng.2019.04.020).

### References

- [1] Nicoletto G, Post D, Smith C. Moire interferometry for high sensitivity measurements in fracture mechanics. in: SESA/JSME Jt Conf Exp Mech; 1982. Oahu-Maui, HI
- [2] Muybridge E.. The horse in motion. library of congress prints and photographs division. 2017.
- [3] Stamper J, McLean E, Obenschain S, Ripin B, Thompson J, Luessen L. Fast electrical and optical measurements. New York: Martinus Nijhoff; 1986. 691
- [4] Xing H, Zhang Q, Braithwaite CH, Pan B, Zhao J. High-speed photography and digital optical measurement techniques for geomaterials: fundamentals and applications. Rock Mech Rock Eng 2017;50(6):1611–59.
- [5] Liang J, Wang LV. Single-shot ultrafast optical imaging. Optica 2018;5(9):1113–27.
- [6] Goda K, Tsia K, Jalali B. Serial time-encoded amplified imaging for real-time observation of fast dynamic phenomena. Nature 2009;458(7242):1145.
- [7] Nakagawa K, Iwasaki A, Oishi Y, Horisaki R, Tsukamoto A, Nakamura A, Hirotsawa K, Liao H, Ushida T, K Goda ea. Sequentially timed all-optical mapping photography (stamp). Nat Photonics 2014;8(9):695.
- [8] Velten A, Lawson E, Bardagjy A, Bawendi M, Raskar R. Slow art with a trillion frames per second camera. In: in: ACM SIGGRAPH 2011 Talks. ACM; 2011. p. 44.
- [9] Beurg M, Fettiplace R, Nam JH, Ricci AJ. Localization of inner hair cell mechanotransducer channels using high-speed calcium imaging. Nat Neurosci 2009;12(5):553.
- [10] Gorkhover T, Schorb S, Coffee R, Adolph M, Foucar L, Rupp D, Aquila A, Bozek JD, Epp SW, Erk B. Femtosecond and nanometre visualization of structural dynamics in superheated nanoparticles. Nat Photonics 2016;10(2):93.

- [11] Gao L, Liang J, Li C, Wang LV. Single-shot compressed ultrafast photography at one hundred billion frames per second. *Nature* 2014;516(7529):74.
- [12] Ford KR, Myer GD, Hewett TE. Reliability of landing 3d motion analysis: implications for longitudinal analyses. *Med Sci Sports Exerc* 2007;39(11):2021–8.
- [13] Jiang H, Zhao H, Li X. High dynamic range fringe acquisition: a novel 3-d scanning technique for high-reflective surfaces. *Opt Lasers Eng* 2012;50(10):1484–93.
- [14] Malamas EN, Petrakis EG, Zervakis M, Petit L, Legat JD. A survey on industrial vision systems. *ApplTools ImageVision Comput* 2003;21(2):171–88.
- [15] Pan B, Qian K, Xie H, Asundi A. Two-dimensional digital image correlation for in-plane displacement and strain measurement: a review. *Meas Sci Technol* 2009;20(6):062001.
- [16] Cai Z, Liu X, Tang Q, Peng X, Gao BZ. Light field 3d measurement using unfocused plenoptic cameras. *Opt Lett* 2018;43(15):3746–9.
- [17] Chen F, Brown GM, Song M. Overview of three-dimensional shape measurement using optical methods. *Opt Eng* 2000;39(1):10–22.
- [18] Geng J. Structured-light 3d surface imaging: a tutorial. *Adv Opt Photonics* 2011;3(2):128–60.
- [19] Xiong Z, Zhang Y, Wu F, Zeng W. Computational depth sensing: toward high-performance commodity depth cameras. *IEEE Signal Process Mag* 2017;34(3):55–68.
- [20] Chen C, Gao N, Wang X, Zhang Z, Gao F, Jiang X. Generic exponential fringe model for alleviating phase error in phase measuring profilometry. *Opt Lasers Eng* 2018;110:179–85.
- [21] Huang L, Idir M, Zuo C, Asundi A. Review of phase measuring deflectometry. *Opt Lasers Eng* 2018;107:247–57.
- [22] Scharstein D, Szeliski R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int J Comput Vis* 2002;47(1–3):7–42.
- [23] Cui Y, Schuon S, Chan D, Thrun S, Theobalt C. 3D shape scanning with a time-of-flight camera. In: *Computer Vision and Pattern Recognition (CVPR)*, 2010 IEEE Conference on. IEEE; 2010. p. 1173–80.
- [24] Gonzalez-Jorge H, Rodríguez-González P, Martínez-Sánchez J, González-Aguilera D, Arias P, Gestó M, Díaz-Vilarino L. Metrological comparison between kinect i and kinect ii sensors. *Measurement* 2015;70:21–6.
- [25] Smisek J, Jancosek M, Pajdla T. 3D With kinect. London: Springer London; 2013. 3–25
- [26] Song S, Lichtenberg SP, Xiao J. Sun rgb-d: a rgb-d scene understanding benchmark suite. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2015.
- [27] Apple iphone x. <https://www.apple.com/iphone/>>, Accessed: 24 December 2018.
- [28] Oppo find x. [https://www.oppo.com/en/smartphone-find\\_x/](https://www.oppo.com/en/smartphone-find_x/), Accessed: 24 December 2018.
- [29] Jeught SvD, Dirckx JJ. Real-time structured light profilometry: a review. *Opt Lasers Eng* 2016;87:18–31.
- [30] Zhang S. High-speed 3d shape measurement with structured light methods: a review. *Opt Lasers Eng* 2018;106:119–31.
- [31] Lei S, Zhang S. Flexible 3-d shape measurement using projector defocusing. *Opt Lett* 2009;34(20):3080–2.
- [32] Zuo C, Tao T, Feng S, Huang L, Asundi A, Chen Q. Micro fourier transform profilometry ( $\mu$ ftp): 3d shape measurement at 10,000 frames per second. *Opt Lasers Eng* 2018;102:70–91.
- [33] Su X, Chen W. Fourier transform profilometry: a review. *Opt Lasers Eng* 2001;35(5):263–84.
- [34] Takeda M, Mutoh K. Fourier transform profilometry for the automatic measurement of 3-d object shapes. *Appl Opt* 1983;22(24):3977–82.
- [35] Zhang Q, Su X. High-speed optical measurement for the drumhead vibration. *Opt Express* 2005;13(8):3110–16.
- [36] Huang L, Kemao Q, Pan B, Asundi AK. Comparison of fourier transform, windowed fourier transform, and wavelet transform methods for phase extraction from a single fringe pattern in fringe projection profilometry. *Opt Lasers Eng* 2010;48(2):141–8.
- [37] Kemao Q. Two-dimensional windowed fourier transform for fringe pattern analysis: principles, applications and implementations. *Opt Lasers Eng* 2007;45(2):304–17.
- [38] Zhong J, Weng J. Spatial carrier-fringe pattern analysis by means of wavelet transform: wavelet transform profilometry. *Appl Opt* 2004;43(26):4993–8.
- [39] Morita H, Yajima K, Sakata S. Reconstruction of surfaces of 3-d objects by m-array pattern projection method. In: *1988 Second International Conference on Computer Vision*. IEEE; 1988. p. 468–73.
- [40] Zhang Z. Review of single-shot 3d shape measurement by phase calculation-based fringe projection techniques. *Opt Lasers Eng* 2012;50(8):1097–106.
- [41] Heist S, Lutzke P, Schmidt I, Dietrich P. P. kühmstedt, a. tünnermann, g. notni, high-speed three-dimensional shape measurement using gobo projection. *Opt Lasers Eng* 2016;87:90–6.
- [42] Schaffer M, Grosse M, Harendt B, Kowarschik R. High-speed three-dimensional shape measurements of objects with laser speckles and acousto-optical deflection. *Opt Lett* 2011;36(16):3097–9.
- [43] Zuo C, Feng S, Huang L, Tao T, Yin W, Chen Q. Phase shifting algorithms for fringe projection profilometry: a review. *Opt Lasers Eng* 2018;109:23–59.
- [44] Feng S, Zuo C, Tao T, Hu Y, Zhang M, Chen Q, Gu G. Robust dynamic 3-d measurements with motion-compensated phase-shifting profilometry. *Opt Lasers Eng* 2018;103:127–38.
- [45] Zuo C, Huang L, Zhang M, Chen Q, Asundi A. Temporal phase unwrapping algorithms for fringe projection profilometry: a comparative review. *Opt Lasers Eng* 2016;85:84–103.
- [46] Tao T, Chen Q, Da J, Feng S, Hu Y, Zuo C. Real-time 3-d shape measurement with composite phase-shifting fringes and multi-view system. *Opt Express* 2016;24(18):20253–69.
- [47] An Y, Hyun JS, Zhang S. Pixel-wise absolute phase unwrapping using geometric constraints of structured light system. *Opt Express* 2016;24(16):18445–59.
- [48] Hyun JS, Chiu GTC, Zhang S. High-speed and high-accuracy 3d surface measurement using a mechanical projector. *Opt Express* 2018;26(2):1474–87.
- [49] Li Z, Zhong K, Li YF, Zhou X, Shi Y. Multiview phase shifting: a full-resolution and high-speed 3d measurement framework for arbitrary shape dynamic objects. *Opt Lett* 2013;38(9):1389–91.
- [50] Zuo C, Chen Q, Gu G, Feng S, Feng F. High-speed three-dimensional profilometry for multiple objects with complex shapes. *Opt Express* 2012;20(17):19493–510.
- [51] Zuo C, Chen Q, Gu G, Feng S, Feng F, Li R, Shen G. High-speed three-dimensional shape measurement for dynamic scenes using bi-frequency tripolar pulse-width-modulation fringe projection. *Opt Lasers Eng* 2013;51(8):953–60.
- [52] Zhang Y, Xiong Z, Wu F. Unambiguous 3d measurement from speckle-embedded fringe. *Appl Opt* 2013;52(32):7797–805.
- [53] Wang Y, Liu K, Hao Q, Lau DL, Hassebrook LG. Period coded phase shifting strategy for real-time 3-d structured light illumination. *IEEE Trans Image Process* 2011;20(11):3001–13.
- [54] Zhang Z, Towers CE, Towers DP. Time efficient color fringe projection system for 3d shape and color using optimum 3-frequency selection. *Opt Express* 2006;14(14):6444–55.
- [55] Tao T, Chen Q, Feng S, Hu Y, Da J, Zuo C. High-precision real-time 3d shape measurement using a bi-frequency scheme and multi-view system. *Appl Opt* 2017;56(13):3646–53.
- [56] Feng S, Chen Q, Gu G, Tao T, Zhang L, Hu Y, Yin W, Zuo C. Fringe pattern analysis using deep learning. *Adv Photonics* 2019;1(2):025001.
- [57] Nah S, Kim TH, Lee KM. Deep multi-scale convolutional neural network for dynamic scene deblurring. In: *CVPR*, vol. 1; 2017. p. 3.
- [58] Rivenson Y, Göröcs Z, Günaydin H, Zhang Y, Wang H, Ozcan A. Deep learning microscopy. *Optica* 2017;4(11):1437–43.
- [59] Rivenson Y, Zhang Y, Günaydin H, Teng D, Ozcan A. Phase recovery and holographic image reconstruction using deep learning in neural networks. *Light* 2018;7(2):17141.
- [60] Schmidhuber J. Deep learning in neural networks: an overview. *Neural Netw* 2015;61:85–117.
- [61] Sinha A, Lee J, Li S, Barbastathis G. Lensless computational imaging through deep learning. *Optica* 2017;4(9):1117–25.
- [62] Floyd RW. An adaptive algorithm for spatial gray-scale. In: *Proc Soc Inf Disp* 1976;17:75–7.
- [63] Zuo C, Chen Q, Feng S, Feng F, Gu G, Sui X. Optimized pulse width modulation pattern strategy for three-dimensional profilometry with projector defocusing. *Appl Opt* 2012;51(19):4477–90.
- [64] Feng S, Chen Q, Zuo C. Graphics processing unit assisted real-time three-dimensional measurement using speckle-embedded fringe. *Appl Opt* 2015;54(22):6865–73.
- [65] Liu K, Wang Y, Lau DL, Hao Q, Hassebrook LG. Dual-frequency pattern scheme for high-speed 3-d shape measurement. *Opt Express* 2010;18(5):5229–44.
- [66] Masci J, Meier U, Cireşan D, Schmidhuber J. Stacked convolutional auto-encoders for hierarchical feature extraction. In: *International Conference on Artificial Neural Networks*. Springer; 2011. p. 52–9.
- [67] Cybenko G. Approximation by superpositions of a sigmoidal function, mathematics of control. *SignalsSyst* 1989;2(4):303–14.
- [68] Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10. USA: Omnipress; 2010. p. 807–14.*
- [69] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. 770–778*
- [70] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521(7553):436.
- [71] Kingma D.P., Ba J. Adam: a method for stochastic optimization. *CoRR abs/1412.6980*.

# Deep-learning-enabled geometric constraints and phase unwrapping for single-shot absolute 3D shape measurement

Cite as: APL Photonics 5, 046105 (2020); <https://doi.org/10.1063/5.0003217>

Submitted: 31 January 2020 . Accepted: 31 March 2020 . Published Online: 14 April 2020

Jiaming Qian , Shijie Feng , Tianyang Tao , Yan Hu , Yixuan Li, Qian Chen , and Chao Zuo 



View Online



Export Citation



CrossMark

APL Photonics The Future Luminary Award

Journal  
Impact Factor  
**4.383**

LEARN MORE!

# Deep-learning-enabled geometric constraints and phase unwrapping for single-shot absolute 3D shape measurement

Cite as: APL Photon. 5, 046105 (2020); doi: 10.1063/5.0003217

Submitted: 31 January 2020 • Accepted: 31 March 2020 •

Published Online: 14 April 2020



Jiaming Qian,<sup>1,2,3,a)</sup>  Shijie Feng,<sup>1,2,3,b)</sup>  Tianyang Tao,<sup>1,2,3</sup>  Yan Hu,<sup>1,2,3</sup>  Yixuan Li,<sup>1,2,3</sup> Qian Chen,<sup>1,2,c)</sup>  and Chao Zuo<sup>1,2,3,d)</sup> 

## AFFILIATIONS

<sup>1</sup>School of Electronic and Optical Engineering, Nanjing University of Science and Technology, No. 200 Xiaolingwei Street, Nanjing, Jiangsu Province 210094, China

<sup>2</sup>Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China

<sup>3</sup>Smart Computational Imaging (SCI) Laboratory, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China

<sup>a)</sup>Electronic mail: [jiaming\\_qian@njust.edu.cn](mailto:jiaming_qian@njust.edu.cn)

<sup>b)</sup>Electronic mail: [ShijieFeng@njust.edu.cn](mailto:ShijieFeng@njust.edu.cn)

<sup>c)</sup>Electronic mail: [chenqian@njust.edu.cn](mailto:chenqian@njust.edu.cn)

<sup>d)</sup>Author to whom correspondence should be addressed: [zuochao@njust.edu.cn](mailto:zuochao@njust.edu.cn)

## ABSTRACT

Fringe projection profilometry (FPP) has become a more prevalently adopted technique in intelligent manufacturing, defect detection, and some other important applications. In FPP, efficiently recovering the absolute phase has always been a great challenge. The stereo phase unwrapping (SPU) technologies based on geometric constraints can eliminate phase ambiguity without projecting any additional patterns, which maximizes the efficiency of the retrieval of the absolute phase. Inspired by recent successes of deep learning for phase analysis, we demonstrate that deep learning can be an effective tool that organically unifies phase retrieval, geometric constraints, and phase unwrapping into a comprehensive framework. Driven by extensive training datasets, the neural network can gradually “learn” to transfer one high-frequency fringe pattern into the “physically meaningful” and “most likely” absolute phase, instead of “step by step” as in conventional approaches. Based on the properly trained framework, high-quality phase retrieval and robust phase ambiguity removal can be achieved only on a single-frame projection. Experimental results demonstrate that compared with traditional SPU, our method can more efficiently and stably unwrap the phase of dense fringe images in a larger measurement volume with fewer camera views. Limitations about the proposed approach are also discussed. We believe that the proposed approach represents an important step forward in high-speed, high-accuracy, motion-artifacts-free absolute 3D shape measurement for complicated objects from a single fringe pattern.

© 2020 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1063/5.0003217>

## I. INTRODUCTION

Optical non-contact three-dimensional (3D) shape measurement techniques have been widely applied for many aspects, such as intelligent manufacturing, reverse engineering, and heritage digitalization.<sup>1</sup> The fringe projection profilometry (FPP)<sup>2</sup> is one of

the most popular optical 3D imaging techniques due to its simple hardware configuration, flexibility in implementation, and high measurement accuracy.

With the development of imaging and projection devices, it becomes possible to realize the high speed 3D shape measurement based on FPP.<sup>3–7</sup> Meanwhile, the acquisition of high-quality 3D



information in high-speed scenarios is increasingly crucial to many applications, such as online quality inspection, stress deformation analysis, and rapid reverse molding.<sup>8,9</sup> To achieve 3D measurement in high-speed scenarios, efforts are usually carried out by reducing the number of images required per reconstruction to improve the measurement efficiency. The ideal way is to obtain 3D data in a single frame. Recently, we have realized high-accuracy phase acquisition from a single fringe pattern by using deep learning.<sup>10,11</sup> However, these works just obtain a single-shot wrapped phase. To realize 3D measurement, phase unwrapping is required, which is one of the operations in FPP that affects the measurement efficiency the most. The most commonly used phase unwrapping methods are temporal phase unwrapping (TPU) algorithms,<sup>12,13</sup> which recover the absolute phase with the assistance of Gray-code patterns or multi-wavelength fringes. However, the requirement of additional patterns decreases the measurement efficiency. The stereo phase unwrapping (SPU)<sup>14</sup> method based on geometric constraints can solve the phase ambiguity problem through the spatial relationships between multiple cameras and one projector without projecting any auxiliary patterns. Although requiring more cameras (at least two) than traditional methods, SPU, indeed, maximizes the efficiency of FPP. However, conventional SPU is generally insufficient to robustly unwrap the phase of dense fringe images, while increasing the frequency of fringes is essential to the measurement accuracy. To solve this trade-off, some auxiliary algorithms are proposed, which usually focus on four directions. (1) The first direction utilizes spatial phase unwrapping methods<sup>15</sup> to reduce phase unwrapping errors of SPU.<sup>14,16</sup> As the disadvantages of spatial phase unwrapping, these methods cannot handle discontinuous or disjointed phases. (2) The second direction enhances the robustness of SPU by embedding the auxiliary information in the fringe patterns.<sup>17,18</sup> Since the assistance based on the intensity information is provided, the sensitivity of intensity to ambient light noise and large surface reflectivity variations of objects will cause them to fail. (3) The third aspect is to increase the number of perspectives and recover the absolute phase through more geometric constraints.<sup>19</sup> This method is more adaptive for the complex scene measurement but comes at the increased cost. Besides, simply increasing the number of views is insufficient to unwrap the phase of dense fringe images, which needs to be combined with (4) the depth constraint strategy.<sup>20–22</sup> However, the conventional depth constraint strategy can only unwrap the phase in a narrow depth range, and setting a suitable depth constraint range is also difficult. The adaptive depth constraint (ADC)<sup>5,23</sup> strategy can enlarge the measurement volume and automatically select the depth constraint range but only if the correct absolute phase can be obtained for the first measurement. In addition, since the stability of SPU relies on the similarity of the phase information of matching points in different perspectives,<sup>19</sup> on the one hand, SPU requires high-quality system calibration and is more difficult to implement algorithmically than other phase unwrapping methods, such as TPU; on the other hand, it has high demand for the quality of the wrapped phase so that the wrapped phase in SPU is usually acquired by the phase-shifting (PS) algorithm,<sup>24</sup> which is a multi-frame phase acquisition method with a high spatial resolution and high measurement accuracy. However, the use of multiple fringe patterns reduces the measurement efficiency of SPU. The other commonly used phase acquisition technologies are Fourier transform (FT) methods<sup>25,26</sup> with single-shot nature, which are not suitable for SPU due to

the poor imaging quality around discontinuities and isolated areas in the phase map.

From the above discussion, it is not difficult to know that although SPU is the best suitable for 3D measurement in high-speed scenes, it still has some defects, such as limited measurement volume, inability to robustly achieve phase unwrapping of high-frequency fringe images, loss of measurement efficiency due to reliance on multi-frame phase acquisition methods, complexity of algorithm implementation, and so on. Inspired by successes of deep learning in FPP<sup>10,11,27,28</sup> and the advance of geometric constraints, on the basis of our previous deep-learning-based works, we further push deep learning into phase unwrapping and incorporate geometric constraints into the neural network. In our work, geometric constraints are implicit in the neural network rather than directly using calibration parameters, which simplifies the entire process of phase unwrapping and avoids the complex adjustment of various parameters. With extensive data training, the network can “learn” to obtain the “physically meaningful” absolute phase from the single-frame projection without the conventional “step-by-step” calculation. Compared with traditional SPU, our approach more robustly unwraps the phase of the higher frequency with fewer perspectives in a larger range. In addition, the limitations of the proposed approach are also analyzed in the Sec. IV.

## II. PRINCIPLE

### A. Phase retrieval and unwrapping with PS and SPU

As shown in Fig. 1, a typical SPU-based system consists of one projector and two cameras. The fringe images are projected by the projector, then modulated by the object, and finally captured by two cameras. For the  $N$ -step PS algorithm, the fringe patterns captured by camera 1 can be expressed as:

$$I_n(u^c, v^c) = A(u^c, v^c) + B(u^c, v^c) \cos(\Phi(u^c, v^c) + 2\pi n/N), \quad (1)$$

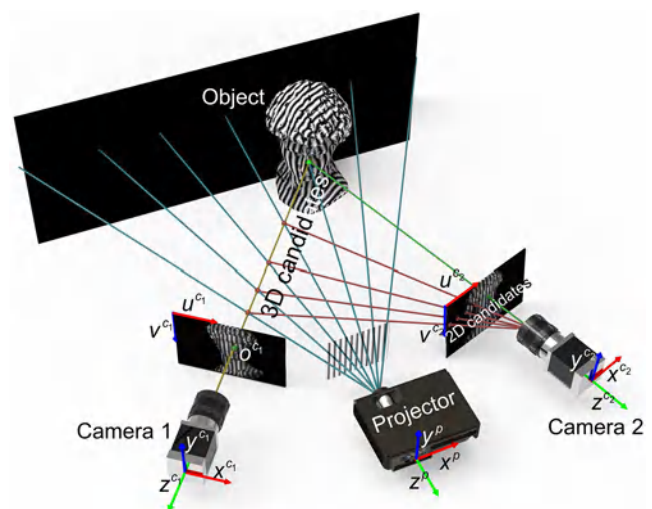


FIG. 1. The principle of SPU.

where  $I_n$  represents the  $(n + 1)$ th captured image,  $n = 0, 1, \dots, N - 1$ ,  $(u^c, v^c)$  is the camera pixel coordinate,  $A$  is the average intensity map,  $B$  is the amplitude intensity map,  $\Phi$  is the absolute phase map, and  $2\pi n/N$  is the phase shift. With the least square method,<sup>29</sup> the wrapped phase  $\varphi$  can be obtained as

$$\varphi = \arctan \frac{M}{D} = \arctan \frac{\sum_{n=0}^{N-1} I_n \sin(2\pi n/N)}{\sum_{n=0}^{N-1} I_n \cos(2\pi n/N)}, \quad (2)$$

where  $(u^c, v^c)$  is omitted for convenience, and  $M$  and  $D$  represent the numerator and denominator of the arctangent function, respectively. The absolute and wrapped phases satisfy the following relation:

$$\Phi = \varphi + 2k\pi, \quad (3)$$

where  $k$  is the fringe order,  $k \in [0, K - 1]$ , and  $K$  denotes the number of the used fringes. The fringe order  $k$  can be obtained by using SPU based on geometric constraints. For an arbitrary point  $o^{c_1}$  in camera 1, there are  $K$  possible fringe orders corresponding to  $K$  absolute phases with which  $K$  3D candidate points can be reconstructed by the calibration parameters between camera 1 and the projector. The retrieved 3D candidates can be projected into camera 2 to obtain the corresponding 2D candidates. Among these 2D candidates, there must be a correct matching point that has a more similar wrapped phase to  $o^{c_1}$  than other candidates.

With this feature, the matching point can be determined through the phase similarity check, and then the phase ambiguity of  $o^{c_1}$  can be eliminated. However, due to calibration errors and ambient light interference, some wrong 2D candidates may have a more similar phase value to  $o^{c_1}$  than the correct matching point. Furthermore, the higher the frequency of the used fringes, the more candidates there are, and the more likely such a situation will happen. Therefore, in order to alleviate this issue, a multi-step PS algorithm with a higher measurement accuracy and robustness toward ambient illumination is preferred, and high-frequency fringe patterns are not recommended.

To enhance the stability of SPU, the common methods adopted are to either increase the number of views or apply the depth constraint strategy. The former, at increased hardware costs, further projects 2D candidates of camera 2 into the third or even the fourth camera for the phase similarity check to exclude more wrong 2D candidates. The latter, at the cost of increased algorithm complexity, can eliminate some wrong 3D candidates outside the depth constraint range in advance. However, the conventional depth constraint algorithm is only effective in a narrow volume. Generally, the SPU with at least three cameras assisted with ADC (the most advanced and complex depth constraint algorithm) can achieve robust phase unwrapping on the premise that the correct absolute phase is obtained for the first measurement.<sup>5,23</sup> However, complex systems and algorithms make such a strategy difficult to implement.

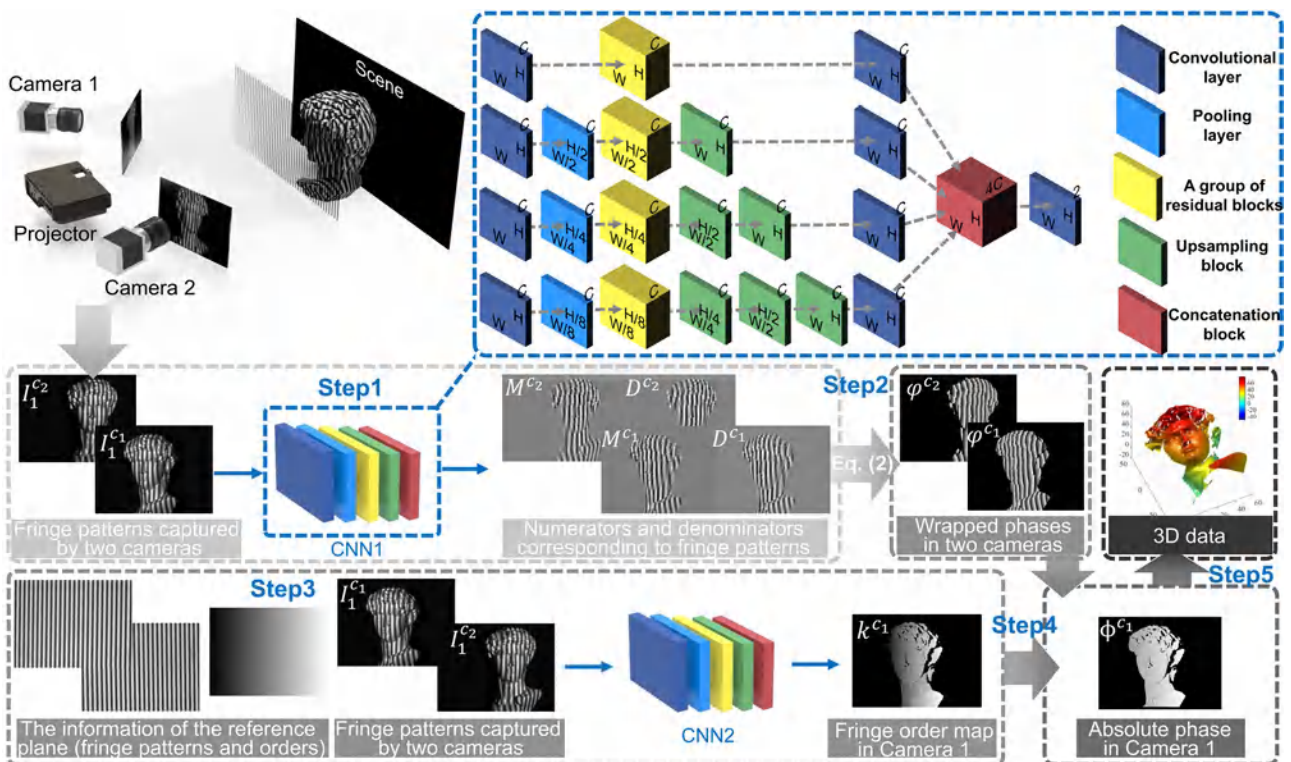


FIG. 2. The flowchart of our method.

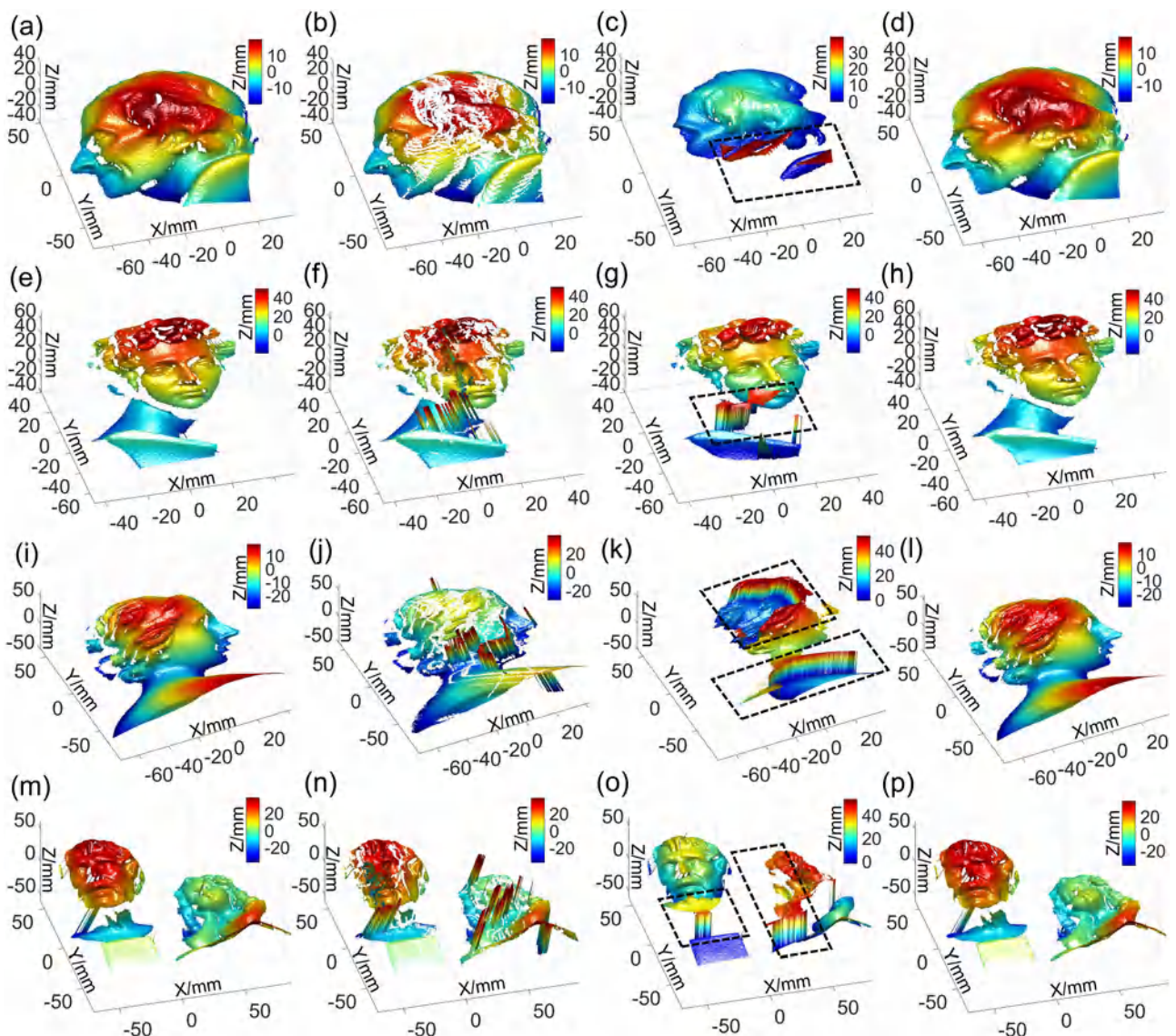


## B. Phase retrieval and unwrapping with deep learning

The ideal SPU should be to use only two cameras and a single frame projector to achieve robust phase unwrapping of dense fringe images in a large measurement volume without any complicated auxiliary algorithms. To this end, inspired by recent successes of deep learning techniques in phase analysis, we combine deep neural networks and SPU to develop a deep-learning-enabled geometric constraints and phase unwrapping method. The flowchart of our approach is shown in Fig. 2. We construct two four-path convolutional neural networks (CNN1 and CNN2) with the

same structure (except for different inputs and outputs) to learn to obtain the high-quality phase information and unwrap the wrapped phase. The detailed architectures of the networks are provided in Appendix A.

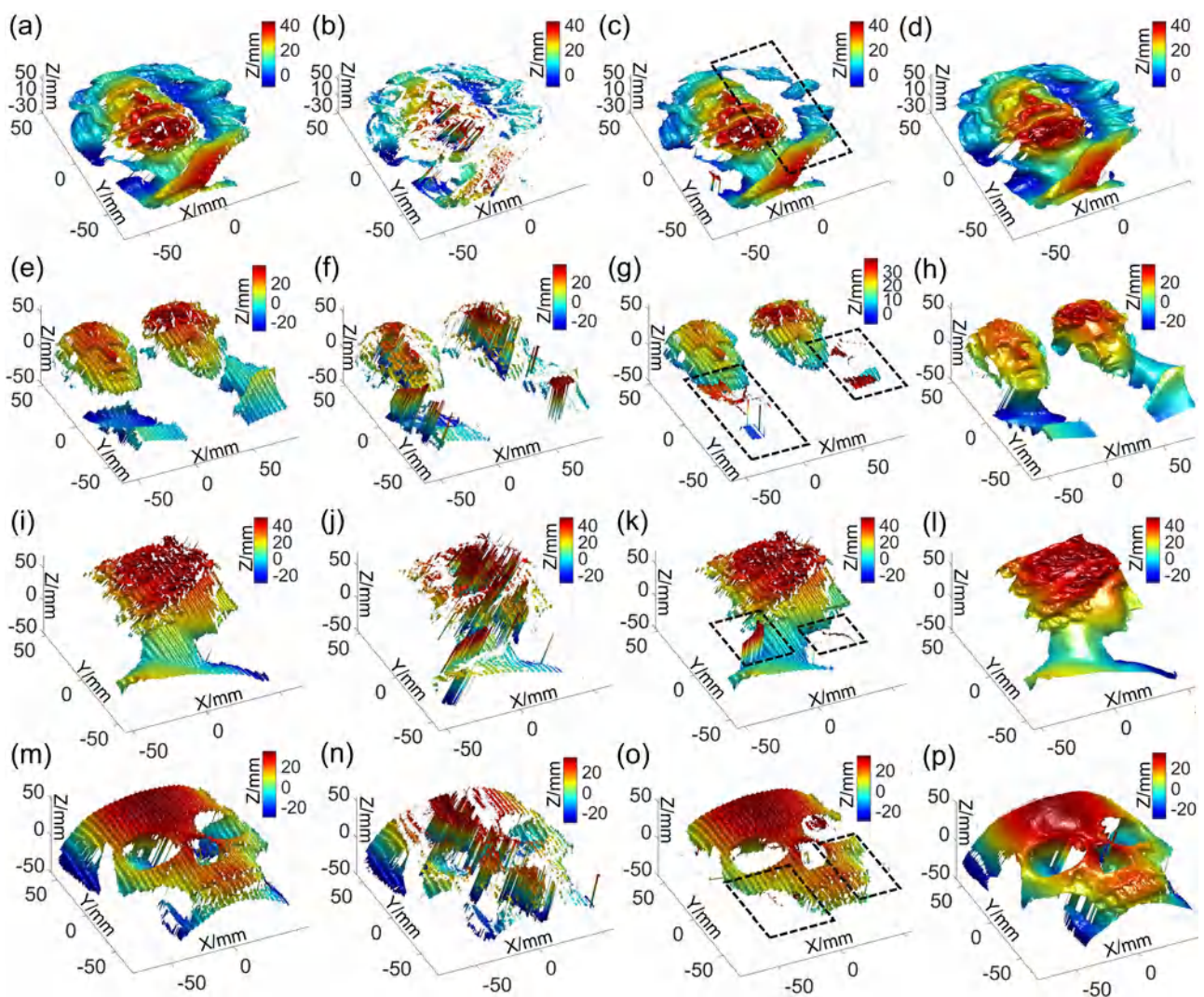
Next, we will discuss our algorithm steps. Step 1: To achieve high-quality wrapped phase information retrieval, the physical model of the conventional PS algorithm is considered. We separately input the single-frame fringe images captured by camera 1 and camera 2 into CNN1 and the outputs are the numerators  $M$  and denominators  $D$  of the arctangent function corresponding to the two fringe patterns instead of directly linked wrapped phases, since such a strategy bypasses the difficulties associated with reproducing abrupt



**FIG. 3.** Measurement results of four static scenes. (a), (e), (i), and (m) The results measured by the first method (taken as the ground-truth data). (b), (f), (j), and (n) The results measured by the second method. (c), (g), (k), and (o) The results measured by the third method. (d), (h), (l), and (p) The results measured by our method.

$2\pi$  phase wraps to provide a high-quality phase estimate.<sup>10</sup> Step 2: After predicting the numerator and denominator terms, high-accuracy wrapped phase maps of camera 1 and camera 2 can be obtained according to Eq. (2). Step 3: To realize the phase unwrapping, enlightened by the geometry-constraint-based SPU described in Sec. II A, which can remove phase ambiguity through spatial relationships between multiple perspectives, the fringe patterns of two perspectives are fed into CNN2. Meanwhile, we integrate the idea of assisting phase unwrapping with the reference plane information<sup>30</sup> to our network and add the data of a reference plane to the inputs to allow CNN2 to more effectively acquire the fringe orders of the measured object. Thus, the raw fringe patterns captured by two cameras, as well as the reference information (containing two

fringe images of the reference plane captured by two cameras, and the fringe order map of the reference plane in the perspective of camera 1) are fed into CNN2. It is worth mentioning that the reference plane information is obtained in advance and subsequent experiments do not need to obtain it repeatedly, which means there is just one extra reference information for the whole setup necessary. The output of CNN2 is the fringe order map of the measured object in camera 1. Step 4: Through the wrapped phases and the fringe orders obtained by the previous steps, high-quality unwrapped phase can be recovered by Eq. (3). Step 5: After acquiring the high-accuracy absolute phase, the 3D reconstruction can be carried out with the calibration parameters<sup>31</sup> between the two cameras (see Appendix B for details).



**FIG. 4.** Measurement results of four dynamic scenes. (a), (e), (i), and (m) The results measured by the first method. (b), (f), (j), and (n) The results measured by the second method. (c), (g), (k), and (o) The results measured by the third method. (d), (h), (l), and (p) The results measured by our method (Multimedia view: <https://doi.org/10.1063/5.0003217.1>; <https://doi.org/10.1063/5.0003217.2>; <https://doi.org/10.1063/5.0003217.3>; <https://doi.org/10.1063/5.0003217.4>, see Visualization 1, Visualization 2, Visualization 3, and Visualization 4 for the whole process of the first scene).



### III. EXPERIMENTS

To verify the effectiveness of the proposed approach, we construct a dual-camera system, which includes a LightCrafter 4500Pro ( $912 \times 1140$  resolution) and two Basler acA640-750  $\mu\text{m}$  cameras ( $640 \times 480$  resolution). 48-period PS fringe patterns are used in our experiments. The size of the measuring field is about  $240 \text{ mm} \times 200 \text{ mm}$ .

To train our networks, we collect training datasets from 1001 different scenarios. With training of hundreds of epochs, the training and validation loss of the networks converge without overfitting. We provide further details of collection of training data and the training process of the neural network in Appendix C.

#### A. Qualitative evaluation

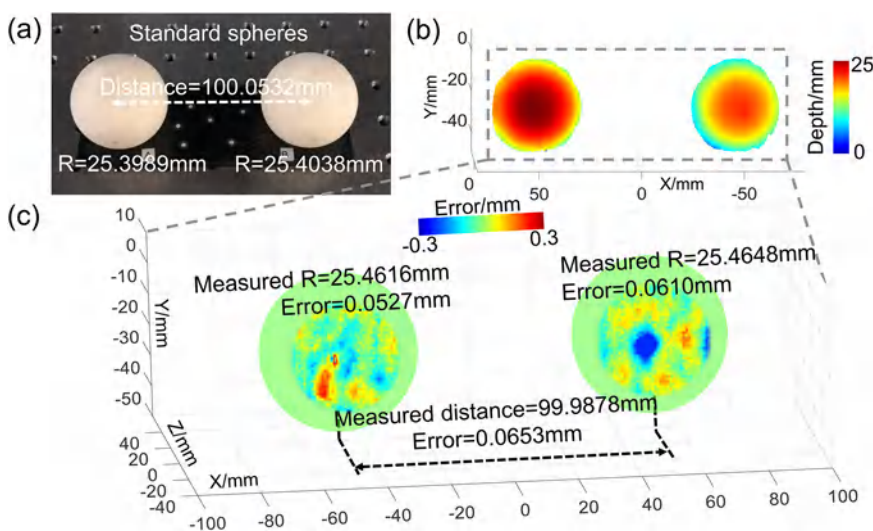
To test the effectiveness of our approach, we firstly measure four static scenarios, containing single or multiple isolated objects with complex shapes, which are not in the training and verification datasets. We use four methods to measure these scenes. The first method is to use PS to obtain the wrapped phase and use triple-camera SPU and ADC to obtain the absolute phase (the results obtained by which are taken as the ground-truth data); the second method is to use PS to obtain the wrapped phase and use dual-camera SPU and the conventional depth constraint strategy to obtain the absolute phase; the third method is to use PS to obtain the wrapped phase and directly use the reference phase to unwrap the phase; the fourth method is our approach. The measurement results are shown in Fig. 3. It can be seen from the results of the second method that the conventional dual-camera SPU and depth constraints are insufficient to unwrap the phase of high-frequency fringes. The parts marked by the black dotted boxes in Fig. 3 show the phase unwrapping errors of the third method, from which we can see that the reference plane can only unwrap the wrapped phase in a limited range, which is between  $-\pi$  and  $\pi$  of the absolute phase

of the reference plane, while with our approach, the ambiguity of the wrapped phase can be accurately eliminated in a large depth range. In addition, our deep-learning-assisted approach can yield high-quality reconstruction results, almost of the same quality as those obtained by conventional PS, triple-camera SPU, and ADC methods.

We also test four continuously moving scenarios to demonstrate the superiority of our approach in the dynamic target measurement (note that all our training and validation datasets are collected in static scenes). The measurement results are shown in Fig. 4 (Multimedia view). It can be seen from the left three columns of Fig. 4 (Multimedia view) that the multi-frame imaging characteristics of the PS algorithm lead to obvious motion-induced artifacts in the reconstruction results when encountering moving objects. In addition, due to the sensitivity to phase errors, the results acquired by SPU obviously perform worse. Because of the single-shot nature of our approach, the measurement can be performed uninterruptedly without being affected by motion artifacts for dynamic scenarios, as shown in the right most column of Fig. 4 (Multimedia view).

#### B. Quantitative evaluation

To quantitatively estimate the reconstruction accuracy of our approach, we measure two standard spheres, whose radii are 25.3989 mm and 25.4038 mm, respectively, and the center-to-center distance is 100.0532 mm with the uncertainty of  $1.1 \mu\text{m}$ . Their errors are  $1.8 \mu\text{m}$  and  $3.5 \mu\text{m}$ , respectively. The measurement result is shown in Fig. 5(b). We perform sphere fitting to measured results of two spheres, and their errors are shown in Fig. 5(c). The radii of the reconstructed spheres are 25.4616 mm and 25.4648 mm with deviations of  $52.7 \mu\text{m}$  and  $61.0 \mu\text{m}$ , respectively. The measured center distance is 99.9878 mm with an error of  $65.3 \mu\text{m}$ . This experiment validates that our method can provide high-quality 3D measurements with fewer cameras, fewer projection images, and simpler algorithms.



**FIG. 5.** Quantitative analysis of our method. (a) The measured standard spheres. (b) 3D reconstruction result of our method. (c) The error distribution of the measured standard spheres.

#### IV. CONCLUSIONS AND DISCUSSIONS

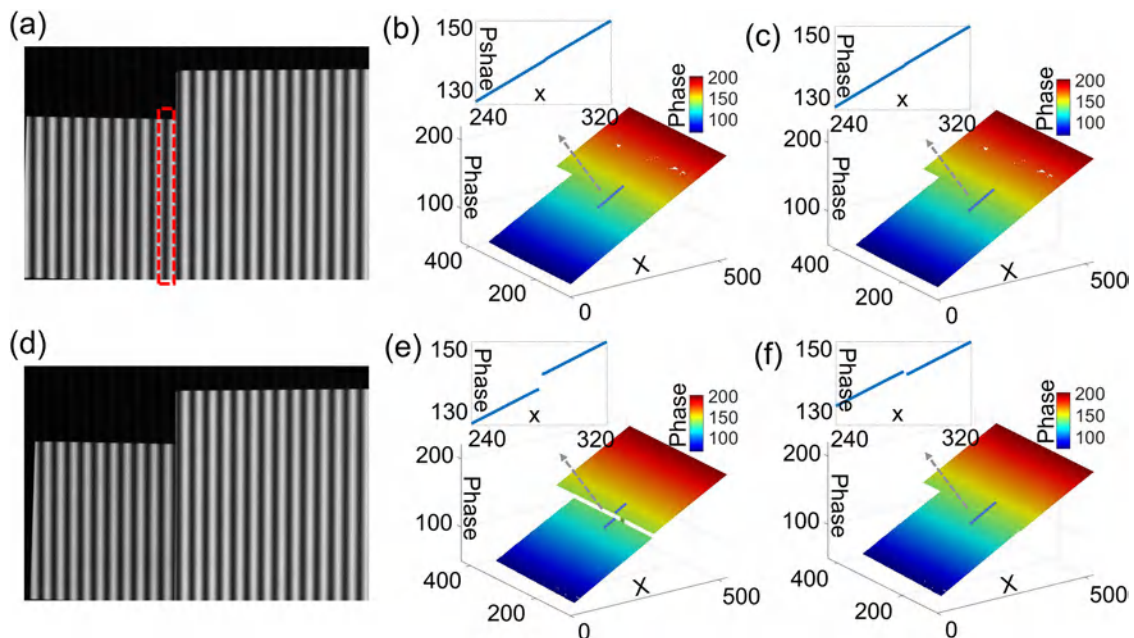
In this work, we present a deep-learning-enabled geometric constraints and phase unwrapping approach for the single-shot absolute 3D shape measurement. Our approach avoids the shortcomings of many traditional methods, such as the trade-off of efficiency and the accuracy of the conventional phase retrieval method and the trade-off of SPU in the phase unwrapping robustness, large measurement range, and the use of high-frequency fringe patterns. On the premise of the single-frame projection, our method can solve the phase ambiguity problem of dense fringe images in a larger measurement range with less perspective information and simpler algorithms. We believe that the proposed approach provides an important guidance for high-accuracy, motion-artifacts-free absolute 3D shape measurement for complicated objects in high-speed scenarios.

For traditional methods, one usually proceeds step by step based on prior knowledge. For example, for SPU, first find 3D candidates, second use depth constraints to remove unreliable candidate points, third project to another perspective, and finally, perform the phase similarity check. Due to the step-by-step process, all information, such as spatial information and temporal information, is not effectively utilized. The comprehensive utilization of all valid data requires strong and professional prior knowledge, which is very difficult to complete. However, deep learning can make it. Through data training and learning, these problems can be

effectively integrated into a comprehensive framework. In our work, this framework is a very organic one, which incorporates phase acquisition, geometric constraints, and phase unwrapping. These methods in the framework are no longer reproduced step by step as traditionally but are organically integrated together. However, since the data sources of our method are 2D images, when the image itself is ambiguous, deep learning is by no means always reliable. For example, when the large depth discontinuity of the object results in missing order and continuity artifact from the camera view (Fig. 6), such inherent ambiguity in the captured fringe pattern cannot be resolved by deep learning techniques without additional auxiliary information, such as fringe images of different frequencies. In the future, we will further integrate the physical model into FPP based on deep learning and construct FPP driven by data and physics.

#### ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (Grant Nos. 61722506, 61705105, and 11574152), National Key R&D Program of China (Grant No. 2017YFF0106403), Outstanding Youth Foundation of Jiangsu Province (Grant No. BK20170034), Fundamental Research Funds for the Central Universities (Grant Nos. 30917011204 and 30919011222), and Leading Technology of Jiangsu Basic Research Plan (Grant No. BK20192003).



**FIG. 6.** Analysis of the limitations of our method. (a) Image of two flat plates captured by camera 1 (no ambiguity in the 2D image). (b) Absolute phase of two plates in (a). (c) The result of two plates in (a) obtained by our method. (d) Image of two flat plates captured by camera 1 (the depth discontinuity of the objects results in missing order [the fringe in the red dotted box in (a)] and continuity artifact from the camera view). (e) Absolute phase of two plates in (d). (f) The result of two plates in (d) obtained by our method.

## APPENDIX A: ARCHITECTURE OF THE NEURAL NETWORKS

We take CNN1 as example to reveal the internal structure of the constructed networks, as shown in the upper right part of Fig. 2. A 3D tensor with size  $(H, W, C_0)$  is used as the input of the network, where  $(H, W)$  is the size of the input images, and  $C_0$  represents the number the input images. For each convolutional layer, the kernel size is  $3 \times 3$  with convolution stride one, zero-padding is used to control the spatial size of the output, and the output is a 3D tensor of shape  $(H, W, C)$ , where  $C = 64$  represents the number of filters used in each convolutional layer. In the first path of CNN1, the input is processed by a convolutional layer, followed by a group of residual blocks (containing four residual blocks) and another convolutional layer. Each residual block consists of two sets of convolutional layer activated by rectified linear unit (ReLU) stacked one above the other,<sup>32</sup> which can solve the degradation of accuracy as the network becomes deeper and ease the training process. In the other three paths, the data are down-sampled by the pooling layers by two, four, and eight times, respectively, for better

feature extraction, and then up-sampled by the upsampling blocks to match the original size. The outputs of four paths are concatenated into a tensor with quad channels. Finally, two channels are generated in the last convolutional layer (one channel is generated in CNN2). Except for the last convolutional layer, which is activated linearly, the rest use the ReLU as activation function. The mean-squared-errors of the outputs with respect to the ground truth are used as the loss function, and the adaptive moment estimation<sup>33</sup> is utilized to tune the parameters for finding the minimum of the loss function.

## APPENDIX B: SYSTEM CALIBRATION AND 3D RECONSTRUCTION

After acquiring the high-accuracy absolute phase, the matching points of two cameras can be uniquely identified. Then, the 3D reconstruction can be carried out with the pre-calibration parameters between the two cameras. The reason why we utilize two cameras for reconstruction instead of one camera and one projector is that the multi-camera system can automatically cancel

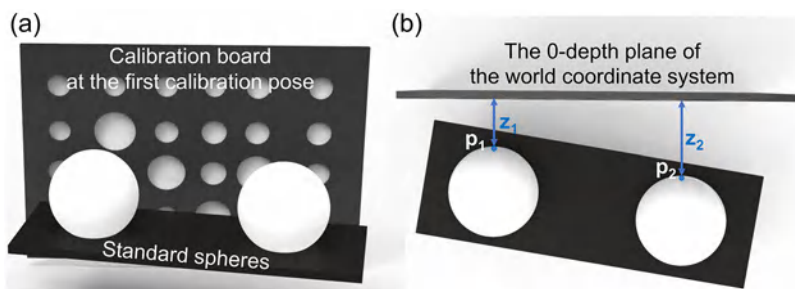


FIG. 7. The relative position of the standard spheres and the calibration board at the first calibration pose.

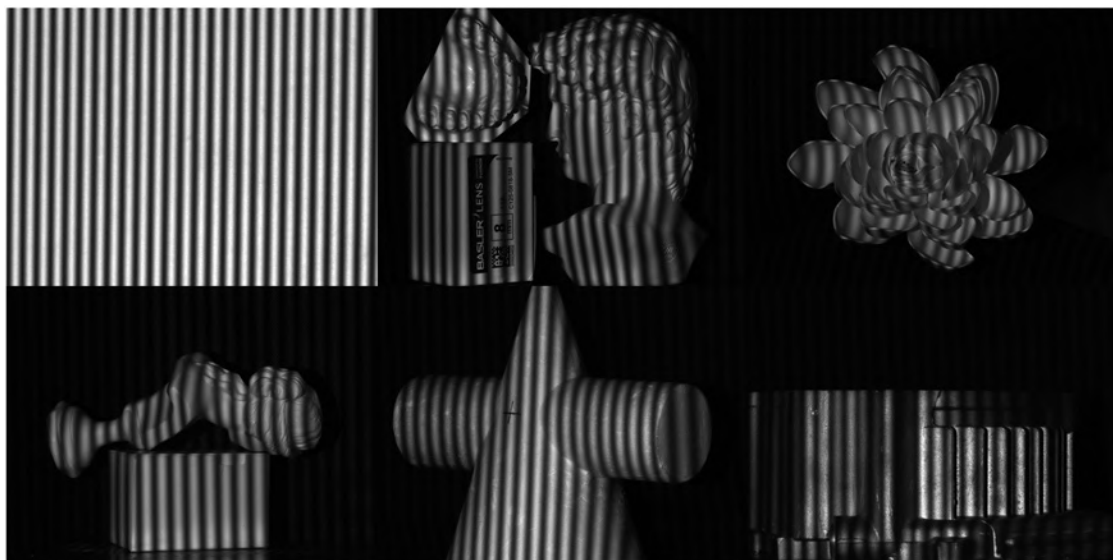
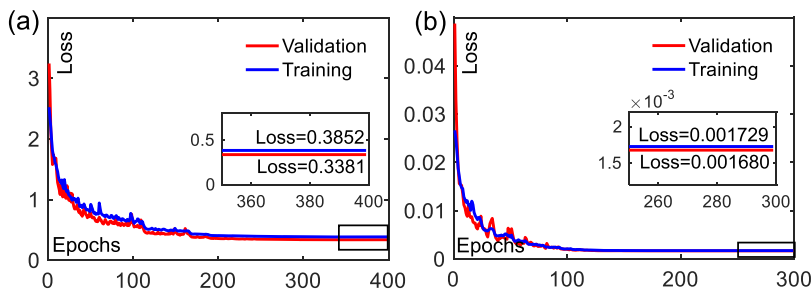


FIG. 8. Six representative scenarios from total 1001 training datasets.



**FIG. 9.** Loss curves of the training and validation sets for (a) CNN1 and (b) CNN2.

nonlinearity errors.<sup>34</sup> The calibration parameters, which contain the intrinsic, extrinsic, and distortion parameters of the cameras are calibrated based on the MATLAB Calibration toolbox and optimized with bundle adjustment.<sup>31,35</sup>

The reconstructed 3D coordinates are in the world coordinate system, the 0-depth plane of which corresponds to the position of the first calibration pose. For example, when the relationship between the position of a pair of standard spheres and the first calibration pose (the 0-depth plane of the world coordinate system) is as shown in Fig. 7, where Fig. 7(a) is the front view of the standard spheres and the calibration board in the first calibration pose and Fig. 7(b) is their top view, the depths of points  $p_1$  and  $p_2$  are  $z_1$  and  $z_2$ , respectively.

### APPENDIX C: TRAINING THE NEURAL NETWORKS

To collect the training datasets, different types of simple and complex objects are arbitrarily combined and rotated  $360^\circ$  to generate 1001 diverse scenes. Figure 8 shows six representative scenarios from total 1001 training datasets, the first of which is the reference plane. Considering the following comparative experiments (verifying that our approach using only two perspectives can perform better than SPU using three cameras in dynamic scenes), we collect data from three views, each set of which consists of 3-step PS fringe patterns captured by three cameras. Within each set of data, we calculate the ground-truth numerator  $M$  and denominator  $D$  by the 3-step PS algorithm and obtain the fringe order maps by using triple-camera SPU and ADC (note that the fringe orders can also be acquired through only a single camera, by projecting multiple fringe patterns of different frequencies and using TPU). Before being fed into the networks, the fringe images are divided by 255 for normalization, and the fringe order maps are divided by the number of the used fringes (48) for normalization, which make the learning process easier for the network. When training the CNNs, 800 sets of data are used for training and 200 sets are used for verification. The training and verification datasets have been uploaded to the figshare (DOI:10.6084/m9.figshare.11926809; <https://figshare.com/s/f150a36191045e0c1bef>).

The constructed neural networks are computed on a GTX Titan graphics card (NVIDIA). Figure 9 shows the loss curve distributions of the CNNs. For CNN1, the loss curves converge after about 200 epochs, and the training of 400 epochs takes 25.56 hours; for CNN2, the loss curves converge after 120 epochs, the training of 300 epochs takes 19.25 h. It is noted that the loss scales of the two networks are different because their outputs are not in the same scale: the numerator  $M$  and denominator  $D$  can reach hundred, while the fringe orders  $k$  are normalized.

### REFERENCES

- J. Salvi, S. Fernandez, T. Pribanic, and X. Llado, "A state of the art in structured light patterns for surface profilometry," *Pattern Recognit.* **43**, 2666–2680 (2010).
- S. S. Gorthi and P. Rastogi, "Fringe projection techniques: Whither we are?," *Opt. Lasers Eng.* **48**, 133–140 (2010).
- S. Zhang, D. Van Der Weide, and J. Oliver, "Superfast phase-shifting method for 3-D shape measurement," *Opt. Express* **18**, 9684–9689 (2010).
- C. Zuo, Q. Chen, G. Gu, S. Feng, F. Feng, R. Li, and G. Shen, "High-speed three-dimensional shape measurement for dynamic scenes using bi-frequency tripolar pulse-width-modulation fringe projection," *Opt. Lasers Eng.* **51**, 953–960 (2013).
- T. Tao, Q. Chen, S. Feng, J. Qian, Y. Hu, L. Huang, and C. Zuo, "High-speed real-time 3D shape measurement based on adaptive depth constraint," *Opt. Express* **26**, 22440–22456 (2018).
- S. Feng, C. Zuo, T. Tao, Y. Hu, M. Zhang, Q. Chen, and G. Gu, "Robust dynamic 3-D measurements with motion-compensated phase-shifting profilometry," *Opt. Lasers Eng.* **103**, 127–138 (2018).
- S. Feng, L. Zhang, C. Zuo, T. Tao, Q. Chen, and G. Gu, "High dynamic range 3D measurements with fringe projection profilometry: A review," *Meas. Sci. Technol.* **29**, 122001 (2018).
- C. Zuo, T. Tao, S. Feng, L. Huang, A. Asundi, and Q. Chen, "Micro Fourier Transform Profilometry ( $\mu$ FTP): 3D shape measurement at 10,000 frames per second," *Opt. Lasers Eng.* **102**, 70–91 (2018).
- J. Qian, S. Feng, T. Tao, Y. Hu, K. Liu, S. Wu, Q. Chen, and C. Zuo, "High-resolution real-time  $360^\circ$  3D model reconstruction of a handheld object with fringe projection profilometry," *Opt. Lett.* **44**, 5751–5754 (2019).
- S. Feng, Q. Chen, G. Gu, T. Tao, L. Zhang, Y. Hu, W. Yin, and C. Zuo, "Fringe pattern analysis using deep learning," *Adv. Photonics* **1**, 025001 (2019).
- S. Feng, C. Zuo, W. Yin, G. Gu, and Q. Chen, "Micro deep learning profilometry for high-speed 3D surface imaging," *Opt. Lasers Eng.* **121**, 416–427 (2019).
- C. Zuo, L. Huang, M. Zhang, Q. Chen, and A. Asundi, "Temporal phase unwrapping algorithms for fringe projection profilometry: A comparative review," *Opt. Lasers Eng.* **85**, 84–103 (2016).
- Z. Zhang, C. E. Towers, and D. P. Towers, "Time efficient color fringe projection system for 3D shape and color using optimum 3-frequency selection," *Opt. Express* **14**, 6444–6455 (2006).
- T. Weise, B. Leibe, and L. Van Gool, "Fast 3D scanning with automatic motion compensation," in *2007 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2007), pp. 1–8.
- X. Su and W. Chen, "Reliability-guided phase unwrapping algorithm: A review," *Opt. Lasers Eng.* **42**, 245–261 (2004).
- R. R. Garcia and A. Zakhor, "Consistent stereo-assisted absolute phase unwrapping methods for structured light systems," *IEEE J. Sel. Top. Signal Process.* **6**, 411–424 (2012).
- W. Lohry and S. Zhang, "High-speed absolute three-dimensional shape measurement using three binary dithered patterns," *Opt. Express* **22**, 26752–26762 (2014).
- T. Tao, Q. Chen, J. Da, S. Feng, Y. Hu, and C. Zuo, "Real-time 3-D shape measurement with composite phase-shifting fringes and multi-view system," *Opt. Express* **24**, 20253–20269 (2016).



- <sup>19</sup>T. Tao, Q. Chen, S. Feng, Y. Hu, M. Zhang, and C. Zuo, "High-precision real-time 3D shape measurement based on a quad-camera system," *J. Opt.* **20**, 014009 (2017).
- <sup>20</sup>C. Bräuer-Burchardt, C. Munkelt, M. Heinze, P. Kühmstedt, and G. Notni, "Using geometric constraints to solve the point correspondence problem in fringe projection based 3D measuring systems," in *International Conference on Image Analysis and Processing* (Springer, 2011), pp. 265–274.
- <sup>21</sup>Z. Li, K. Zhong, Y. F. Li, X. Zhou, and Y. Shi, "Multiview phase shifting: A full-resolution and high-speed 3D measurement framework for arbitrary shape dynamic objects," *Opt. Lett.* **38**, 1389–1391 (2013).
- <sup>22</sup>X. Liu and J. Kofman, "High-frequency background modulation fringe patterns based on a fringe-wavelength geometry-constraint model for 3D surface-shape measurement," *Opt. Express* **25**, 16618–16628 (2017).
- <sup>23</sup>J. Qian, T. Tao, S. Feng, Q. Chen, and C. Zuo, "Motion-artifact-free dynamic 3D shape measurement with hybrid fourier-transform phase-shifting profilometry," *Opt. Express* **27**, 2713–2731 (2019).
- <sup>24</sup>C. Zuo, S. Feng, L. Huang, T. Tao, W. Yin, and Q. Chen, "Phase shifting algorithms for fringe projection profilometry: A review," *Opt. Lasers Eng.* **109**, 23–59 (2018).
- <sup>25</sup>X. Su and Q. Zhang, "Dynamic 3-D shape measurement method: A review," *Opt. Lasers Eng.* **48**, 191–204 (2010).
- <sup>26</sup>L. Huang, Q. Kema, B. Pan, and A. K. Asundi, "Comparison of fourier transform, windowed fourier transform, and wavelet transform methods for phase extraction from a single fringe pattern in fringe projection profilometry," *Opt. Lasers Eng.* **48**, 141–148 (2010).
- <sup>27</sup>S. Van der Jeught and J. J. J. Dirckx, "Deep neural networks for single shot structured light profilometry," *Opt. Express* **27**, 17091–17101 (2019).
- <sup>28</sup>W. Yin, Q. Chen, S. Feng, T. Tao, L. Huang, M. Trusiak, A. Asundi, and C. Zuo, "Temporal phase unwrapping using deep learning," *Sci. Rep.* **9**, 1–12 (2019).
- <sup>29</sup>Y. Hu, Q. Chen, Y. Liang, S. Feng, T. Tao, and C. Zuo, "Microscopic 3D measurement of shiny surfaces based on a multi-frequency phase-shifting scheme," *Opt. Lasers Eng.* **122**, 1–7 (2019).
- <sup>30</sup>Y. An, J.-S. Hyun, and S. Zhang, "Pixel-wise absolute phase unwrapping using geometric constraints of structured light system," *Opt. Express* **24**, 18445–18459 (2016).
- <sup>31</sup>Y. Yin, X. Peng, A. Li, X. Liu, and B. Z. Gao, "Calibration of fringe projection profilometry with bundle adjustment strategy," *Opt. Lett.* **37**, 542–544 (2012).
- <sup>32</sup>K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (IEEE, 2016), pp. 770–778.
- <sup>33</sup>D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014).
- <sup>34</sup>W. Lohry, V. Chen, and S. Zhang, "Absolute three-dimensional shape measurement using coded fringe patterns without phase unwrapping or projector calibration," *Opt. Express* **22**, 1287–1301 (2014).
- <sup>35</sup>L. Huang, Q. Zhang, and A. Asundi, "Camera calibration with active phase target: Improvement on feature detection and optimization," *Opt. Lett.* **38**, 1446–1448 (2013).



# Optics Letters

## Single-shot absolute 3D shape measurement with deep-learning-based color fringe projection profilometry

JIAMING QIAN,<sup>1,2</sup> SHIJIE FENG,<sup>1,2</sup> YIXUAN LI,<sup>1,2</sup> TIANYANG TAO,<sup>1,2</sup>  JING HAN,<sup>2</sup>  
QIAN CHEN,<sup>2,3</sup>  AND CHAO ZUO<sup>1,2,\*</sup> 

<sup>1</sup>Smart Computational Imaging Laboratory (SCILab), School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China

<sup>2</sup>Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China

<sup>3</sup>e-mail: chenqian@njust.edu.cn

\*Corresponding author: zuochao@njust.edu.cn

Received 23 January 2020; revised 23 February 2020; accepted 27 February 2020; posted 28 February 2020 (Doc. ID 388994); published 20 March 2020

Recovering the high-resolution three-dimensional (3D) surface of an object from a single frame image has been the ultimate goal long pursued in fringe projection profilometry (FPP). The color fringe projection method is one of the technologies with the most potential towards such a goal due to its three-channel multiplexing properties. However, the associated color imbalance, crosstalk problems, and compromised coding strategy remain major obstacles to overcome. Inspired by recent successes of deep learning for FPP, we propose a single-shot absolute 3D shape measurement with deep-learning-based color FPP. Through “learning” on extensive data sets, the properly trained neural network can “predict” the high-resolution, motion-artifact-free, crosstalk-free absolute phase directly from one single color fringe image. Compared with the traditional approach, our method allows for more accurate phase retrieval and more robust phase unwrapping. Experimental results demonstrate that the proposed approach can provide high-accuracy single-frame absolute 3D shape measurement for complicated objects. © 2020 Optical Society of America

<https://doi.org/10.1364/OL.388994>

Fringe projection profilometry (FPP) [1] is one of the most widely used three-dimensional (3D) measurement techniques because of its simple hardware facilities, flexible implementation, and high measurement precision. Recently, with the increasing demands of 3D information acquisition in various applications, such as online quality inspection and rapid reverse engineering, high-speed 3D shape measurement technologies based on FPP are becoming more and more popular and essential [2–5].

To achieve high-speed 3D imaging, it is necessary to improve the measurement efficiency, i.e., to reduce the number of patterns required per reconstruction [6,7]. Ideally, the absolute

3D surface of an object should be recovered from only a single image. The color-coded projection methods [8–13] have great advantages in dynamic scene measurement, since it can encode three independent patterns in its red, blue, and green channels. However, few of them can be used for high-accuracy measurements of complex objects. On one hand, to obtain high-accuracy phase information, the phase-shifting (PS) method [14] with high measurement resolution is preferred. However, PS requires at least three fringe images, which occupy all channels of an RGB image, so that only the spatial phase unwrapping method [15] (which will become vulnerable when encountering isolated or unjoined phase) can remove the phase ambiguity [9]. On the other hand, to achieve robust phase unwrapping, the strategies of combining fringe patterns with the fringe-order-coded information (gray-code patterns or stair intensities) or combining multi-frequency fringe images are adopted [10–13]. The former still cannot unwrap the phase stably due to the difficulty in identifying the edge of gray-code patterns or the sensitivity of intensity to ambient light noise and large surface reflectivity variations of the objects [10,11]. The latter can recover the absolute phase by the three-fringe number selection method [16] but compromises the accuracy due to the use of the Fourier transform (FT) method [15,17] (well known for its single-shot nature but with limited quality around discontinuities and isolated areas in the phase map) [12,13]. In addition, there are some inherent defects in color-coded projection methods, such as chromatic aberration and crosstalk between color channels, which will affect the quality of phase calculations. Some researchers propose some pre-processing methods [9,12,18] to compensate for these factors, but only to some extent reduce their impact.

Recently, the successes of deep learning in FPP [19–23] have provided new opportunities for color-coded projection technologies. In this Letter, we propose a deep-learning-based color fringe projection profilometry. With the help of a properly trained neural network, a high-accuracy absolute

phase can be “learned” from a color-coded image without any pre/post-processing or phase error compensation.

The flowchart of our method is shown in Fig. 1. Notably, instead of adopting an end-to-end network structure directly linking fringe images to the absolute phase, our network predicts a high-quality wrapped phase map along with a “coarse” absolute phase from the three-channel fringe patterns of different frequencies. There are three basic considerations for this strategy: (1) due to the inherent depth ambiguities in FPP, one single fringe pattern is generally insufficient to uniquely determine the fringe order in the presence of isolated surfaces or surface discontinuities [5]. Thus, to guarantee absolute 3D shape measurement independent of any assumptions and prior knowledge, multi-frequency fringe patterns and temporal phase unwrapping (TPU) [5] are necessary. (2) Without additional auxiliary information, a simple input–output network structure (linking fringe images to the absolute phase directly) usually produces an estimate with compromised accuracy; especially, the measured surface contains sharp edges, discontinuities, or large surface reflectivity variations [23]. Based on this consideration, we incorporate the idea of TPU into our network to determine the fringe order of the wrapped phase with the predicted “coarse” absolute phase on a pixel-by-pixel basis. (3) For the unwrapped phase map, our deep neural network is trained to predict the intermediate numerator and denominator of the arctangent function, to bypass the difficulties associated with reproducing abrupt  $2\pi$  phase wraps and thus obtain a high-quality phase estimate [19].

Our pattern design is similar to that of [13], which encodes three fringe patterns of different wavelengths  $\lambda_R$ ,  $\lambda_G$ , and  $\lambda_B$  into the red, green, and blue channels of one color image. The color-coded image is projected by a projector, modulated by

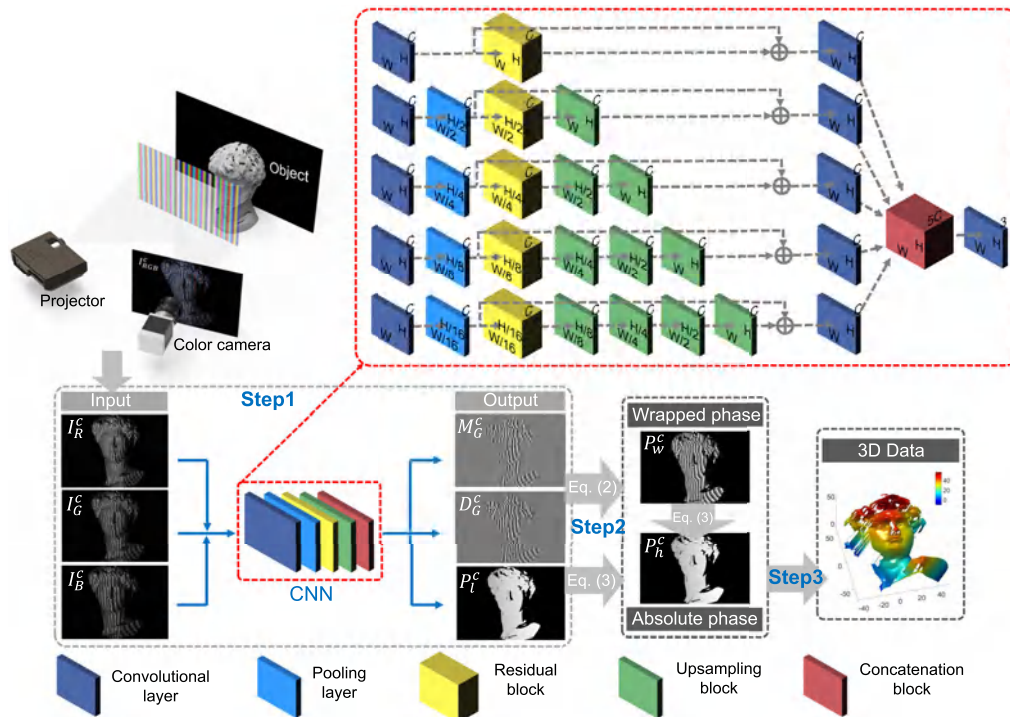
the object, and finally captured by a color camera. The captured image can be represented as  $I_{RGB}^c$ , and the gray images of its three channels can be expressed as  $I_R^c$ ,  $I_G^c$ , and  $I_B^c$ , respectively. Since three fringe images with different wavelengths are used, the phase unwrapping can be achieved by the projection distance minimization (PDM) approach [24], which obtains the pixel-wise qualified fringe order according to wrapped phase distribution in three fringe images. Because the wrapped phase is encoded in the fringe pattern and the deep learning can organically synthesize the spatial and temporal information [23], we can expect that the properly trained neural network can directly obtain the absolute phase from the fringe patterns of three frequencies. However, only the low-accuracy absolute phase can be predicted. To obtain high-quality phase information, the physical model of the traditional PS method is considered. For the  $N$ -step PS algorithm, the  $N$  fringe patterns can be expressed as

$$I_n = A + B \cos(\Phi + 2\pi n/N), \quad (1)$$

where  $I_n$  represents the  $(n + 1)$ th fringe image,  $n \in [0, N - 1]$ ,  $A$  is the average intensity,  $B$  is the amplitude,  $\Phi$  is the absolute phase, and  $2\pi/N$  is the phase shift. The wrapped phase  $\phi$  can be obtained by [14]

$$\phi = \arctan \frac{M}{D} = \arctan \frac{\sum_{n=0}^{N-1} I_n \sin(2\pi n/N)}{\sum_{n=0}^{N-1} I_n \cos(2\pi n/N)}, \quad (2)$$

where  $M$  and  $D$  represent the numerator and denominator of arctangent function, respectively. If the network predicts  $M$  and  $D$ , the phase information can be obtained by Eq. (2). Such an operation provides higher phase accuracy than the network structure of directly linking fringe pattern to phase



**Fig. 1.** Flowchart of our method. Step 1: input three gray fringe images  $I_R^c$ ,  $I_G^c$ , and  $I_B^c$  of the color image channels, and output the numerator and denominator terms  $M_G^c$  and  $D_G^c$  and the low-accuracy absolute phase  $P_f^c$  in the green channel. Step 2: obtain the high-accuracy absolute phase  $P_h^c$  by Eqs. (2) and (3). Step 3: reconstruct the 3D information by the calibration parameters.



[19]. Therefore,  $I_R^c$ ,  $I_G^c$ , and  $I_B^c$  are fed into the constructed neural network, and the outputs are the numerator  $M_G^c$  and denominator  $D_G^c$ , and a rough absolute phase map  $P_l^c$  (whose error is between  $-\pi$  and  $\pi$ ) in the green channel. In addition, to enable the network to resist crosstalk and chromatic aberration problems, the data without these factors are used as labels to train our network. After the network predicts  $M_G^c$ ,  $D_G^c$ , and  $P_l^c$ , the wrapped phase  $P_w^c$  of wavelength  $\lambda_G$  can be obtained by Eq. (2). Then the high-quality absolute phase  $P_b^c$  can be acquired by

$$P_b^c = P_w^c + 2\pi \text{Round}[(P_l^c - P_w^c)/2/\pi], \quad (3)$$

where Round represents the rounding function. Finally, the 3D reconstruction can be performed by utilizing the pre-calibrated parameters [25] of the system.

The upper right part of Fig. 1 reveals the internal structure of our network, which is a five-path convolutional neural network (CNN) and is more powerful than our previous CNNs [19,21–23]. For each convolutional layer, the kernel size is  $3 \times 3$  with convolution stride one, and the output is a 3D tensor of shape  $(H, W, C)$ , where  $(H, W)$  is the size of the input pattern, and  $C$  represents the number of filters used in each convolutional layer ( $C = 64$ ). In the first path of the CNN, the inputs are processed by a convolutional layer, followed by four residual blocks and another convolutional layer. Also, implementing shortcuts between residual blocks contributes to the convolution stability [22]. In the other four paths, the data are down-sampled by the max-pooling layers by two times, four times, eight times, and 16 times, respectively, for better feature extraction, and then up-sampled by the up-sampling blocks to match the original size. The outputs of all paths are concatenated into a tensor with five channels. Finally, three channels are generated in the last convolution layer.

We construct an FPP system, which includes a LightCrafter 4500 ( $912 \times 1140$  resolution) and a Basler acA640-750uc

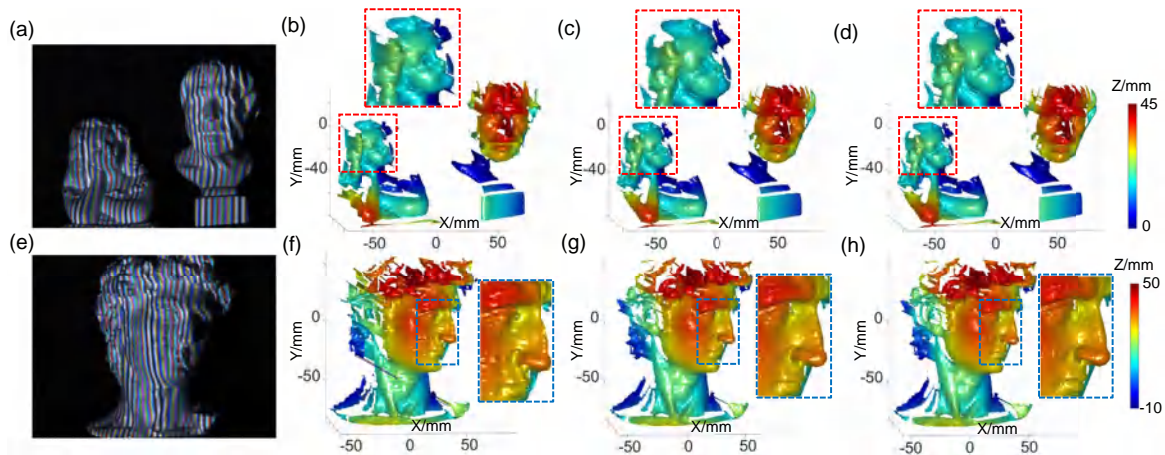
camera ( $640 \times 480$  resolution), to test the effectiveness of our method. The wavelengths of three color channels are selected as  $\lambda_R = 9$ ,  $\lambda_G = 11$ , and  $\lambda_B = 13$ , which provide unambiguous 3D reconstructions for the whole projection range. During the training session, 600 groups of images are captured. Each group contains a color-coded fringe pattern [Fig. 2(a)], the three channels of which are used as inputs of the network, as well as the 12-step PS fringe patterns of three frequencies [Figs. 2(b)–2(d)], which are used for the calculation of the ground-truth data (the three frequencies of the latter are consistent with those of the three channels of the former). To avoid crosstalk and chromatic aberration problems at the source, when collecting the PS images, the green non-composite fringe patterns are projected, and only the green channels of the captured images are utilized for the label. The numerator and denominator terms are calculated by Eq. (2). The absolute phase is obtained based on PDM. When training, 450 groups of data are used for training, and 150 groups are used for verification.

To verify the superiority of our method over traditional methods, we measure two scenarios with our method and the method of [13]. The measured results are shown in Fig. 3. The shadow-noised regions, with  $B$  smaller than a pre-defined threshold, are excluded in the subsequent processing [14]. Figures 3(b) and 3(f) are results of the method of [13], from which we can see that despite the crosstalk compensation in advance, the resolution of the details is limited due to the use of FT method. Also, the phase errors lead to the phase unwrapping errors at the edges. In contrast to the traditional method, our method produces more accurate absolute reconstruction, whose quality is even comparable to that obtained by PS and PDM methods.

Next, we apply our approach to the dynamic 360-deg 3D model reconstruction. A metallic workpiece is continually rotated for one cycle on a mechanical stage and measured with our approach (note that there are no objects of such type in our training and verification set). The measured results from two

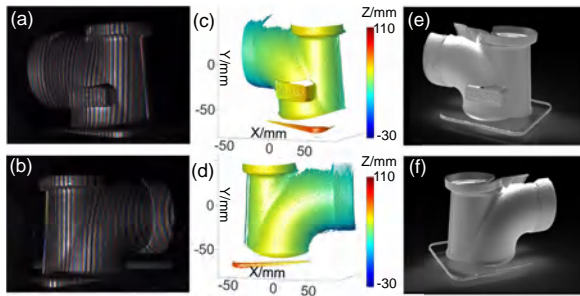


**Fig. 2.** Group of data sets. (a) Color image. (b)–(d) 12-step PS images of  $\lambda_R$ ,  $\lambda_G$ , and  $\lambda_B$ .

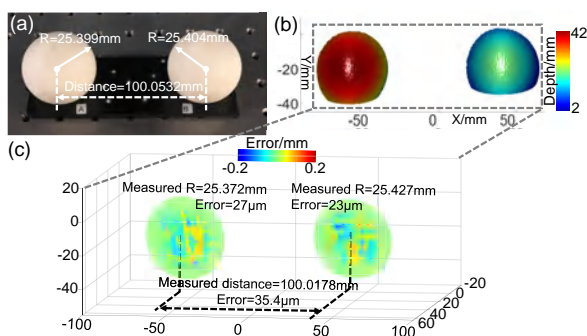


**Fig. 3.** Measurement results of two scenes. (a), (e) Captured color images. (b), (f) Results measured by the method of [13]. (c), (g) Results of our method. (d), (h) Ground truth obtained by 12-step PS and the projection distance minimization.





**Fig. 4.** Measured results of a 360-deg rotated workpiece with our method. (a), (b) Color images of different times. (c), (d) Results corresponding to (a), (b). (e), (f) Registration results (see Visualization 1, Visualization 2, and Visualization 3 for the whole process).



**Fig. 5.** Quantitative analysis of the accuracy of our method. (a) Measured standard spheres. (b) Measured results of our method. (c) Error distribution of the measured result.

different perspectives are shown in Figs. 4(a)–4(d). Due to the single-shot nature of our approach, the measurement can be performed uninterruptedly without being affected by motion artifacts. We register all the independent measurements into an integral 3D model, which is shown in Figs. 4(e) and 4(f). This experiment demonstrates the potential of our method for rapid reverse engineering applications.

Finally, to quantify the accuracy of our method, two standard spheres as shown in Fig. 5(a) are measured. The measured result is shown in Fig. 5(b). We perform sphere fitting to the results, and the error distribution is shown in Fig. 5(c). The radii of reconstructed spheres are 25.372 mm and 25.427 mm, with errors of 27  $\mu\text{m}$  and 23  $\mu\text{m}$ . The measured center distance is 100.0178 mm, with an error of 35.4  $\mu\text{m}$ . This experiment validates the high measurement accuracy of our method.

In this Letter, we have presented a deep-learning-based color FPP. With deep learning, color-coded projection technologies are rejuvenated in single-frame, high-precision 3D imaging. Deep learning breaks the dependence of traditional methods on prior knowledge and can more efficiently utilize the raw information “hidden” in the original fringe pattern. However, it should be mentioned that for some other disadvantages of

color-coded projection methods, such as being unsuited for measuring color objects, the proposed depth learning approach is still inadequate because the information cannot “come out of thin air.” The fringe in a certain channel will be blended onto the object surface, leading to extremely poor fringe contrast and information lost. One possible solution is to create “invisible fringes” with infrared or ultraviolet light sources, which is an interesting direction for future work.

**Funding.** National Natural Science Foundation of China (61705105, 61722506); National Key Research and Development Program of China (2017YFF0106403); Outstanding Youth Foundation of Jiangsu Province of China (BK20170034); Fundamental Research Funds for the Central Universities (30917011204, 30919011222).




**Disclosures.** The authors declare no conflicts of interest.

## REFERENCES

1. S. S. Gorthi and P. Rastogi, *Opt. Lasers Eng.* **48**, 133 (2010).
2. S. Feng, C. Zuo, T. Tao, Y. Hu, M. Zhang, Q. Chen, and G. Gu, *Opt. Lasers Eng.* **103**, 127 (2018).
3. J. Qian, S. Feng, T. Tao, Y. Hu, K. Liu, S. Wu, Q. Chen, and C. Zuo, *Opt. Lett.* **44**, 5751 (2019).
4. T. Tao, Q. Chen, S. Feng, J. Qian, Y. Hu, L. Huang, and C. Zuo, *Opt. Express* **26**, 22440 (2018).
5. C. Zuo, L. Huang, M. Zhang, Q. Chen, and A. Asundi, *Opt. Lasers Eng.* **85**, 84 (2016).
6. C. Zuo, Q. Chen, G. Gu, S. Feng, F. Feng, R. Li, and G. Shen, *Opt. Lasers Eng.* **51**, 953 (2013).
7. J. Qian, T. Tao, S. Feng, Q. Chen, and C. Zuo, *Opt. Express* **27**, 2713 (2019).
8. Z. Zhang, *Opt. Lasers Eng.* **50**, 1097 (2012).
9. P. S. Huang, Q. Hu, F. Jin, and F. P. Chiang, *Opt. Eng.* **38**, 1065 (1999).
10. W.-H. Su, *Opt. Express* **16**, 2590 (2008).
11. N. Karpinsky and S. Zhang, *Opt. Eng.* **49**, 063604 (2010).
12. Z. Zhang, C. E. Towers, and D. P. Towers, *Opt. Express* **14**, 6444 (2006).
13. Z. Zhang, D. P. Towers, and C. E. Towers, *Appl. Opt.* **49**, 5947 (2010).
14. C. Zuo, S. Feng, L. Huang, T. Tao, W. Yin, and Q. Chen, *Opt. Lasers Eng.* **109**, 23 (2018).
15. X. Su and Q. Zhang, *Opt. Lasers Eng.* **48**, 191 (2010).
16. D. Towers, C. Towers, and Z. Zhang, “Optical imaging of physical objects,” U.S. patent App. 12/377,180 (7 July 2010).
17. L. Huang, Q. Kemao, B. Pan, and A. K. Asundi, *Opt. Lasers Eng.* **48**, 141 (2010).
18. Z. Zhang, C. Towers, and D. Towers, *Opt. Lasers Eng.* **48**, 159 (2010).
19. S. Feng, Q. Chen, G. Gu, T. Tao, L. Zhang, Y. Hu, W. Yin, and C. Zuo, *Adv. Photon.* **1**, 025001 (2019).
20. S. Van der Jeught and J. J. Dirckx, *Opt. Express* **27**, 17091 (2019).
21. S. Feng, C. Zuo, W. Yin, G. Gu, and Q. Chen, *Opt. Lasers Eng.* **121**, 416 (2019).
22. W. Yin, Q. Chen, S. Feng, T. Tao, L. Huang, M. Trusiak, A. Asundi, and C. Zuo, arXiv preprint arXiv:1903.09836 (2019).
23. J. Qian, S. Feng, T. Tao, Y. Hu, Y. Li, Q. Chen, and C. Zuo, arXiv preprint arXiv:2001.01439 (2020).
24. C. Zuo, T. Tao, S. Feng, L. Huang, A. Asundi, and Q. Chen, *Opt. Lasers Eng.* **102**, 70 (2018).
25. Y. Yin, X. Peng, A. Li, X. Liu, and B. Z. Gao, *Opt. Lett.* **37**, 542 (2012).



# Single-shot 3D shape measurement using an end-to-end stereo matching network for speckle projection profilometry

WEI YIN,<sup>1,2,3,4</sup>  YAN HU,<sup>1,2,3,4</sup>  SHIJIE FENG,<sup>1,2,3,4,7</sup>  LEI HUANG,<sup>5</sup>  QIAN KEMAO,<sup>6</sup>  QIAN CHEN,<sup>1,2,8</sup>  AND CHAO ZUO<sup>1,2,3,4,9,10</sup> 

<sup>1</sup>*School of Electronic and Optical Engineering, Nanjing University of Science and Technology, No. 200 Xiaolingwei Street, Nanjing, Jiangsu Province 210094, China*

<sup>2</sup>*Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing, Jiangsu Province 210094, China*

<sup>3</sup>*Smart Computational Imaging (SCI) Laboratory, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China*

<sup>4</sup>*Smart Computational Imaging Research Institute (SCRI) of Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210019, China*

<sup>5</sup>*Brookhaven National Laboratory, NSLS II 50 Rutherford Drive, Upton, New York 11973-5000, USA*

<sup>6</sup>*School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore*

<sup>7</sup>*geniussjiejie@163.com*

<sup>8</sup>*chenqian@njust.edu.cn*

<sup>9</sup>*zuochao@njust.edu.cn*

<sup>10</sup>*surpasszuo@163.com*

**Abstract:** Speckle projection profilometry (SPP), which establishes the global correspondences between stereo images by projecting only a single speckle pattern, has the advantage of single-shot 3D reconstruction. Nevertheless, SPP suffers from the low matching accuracy of traditional stereo matching algorithms, which fundamentally limits its 3D measurement accuracy. In this work, we propose a single-shot 3D shape measurement method using an end-to-end stereo matching network for SPP. To build a high-quality SPP dataset for training the network, by combining phase-shifting profilometry (PSP) and temporal phase unwrapping techniques, high-precision absolute phase maps can be obtained to generate accurate and dense disparity maps with high completeness as the ground truth by phase matching. For the architecture of the network, a multi-scale residual subnetwork is first leveraged to synchronously extract compact feature tensors with 1/4 resolution from speckle images for constructing the 4D cost volume. Considering that the cost filtering based on 3D convolution is computationally costly, a lightweight 3D U-net network is proposed to implement efficient 4D cost aggregation. In addition, because the disparity maps in the SPP dataset should have valid values only in the foreground, a simple and fast saliency detection network is integrated to avoid predicting the invalid pixels in the occlusions and background regions, thereby implicitly enhancing the matching accuracy for valid pixels. Experiment results demonstrated that the proposed method improves the matching accuracy by about 50% significantly compared with traditional stereo matching methods. Consequently, our method achieves fast and absolute 3D shape measurement with an accuracy of about 100 $\mu\text{m}$  through a single speckle pattern.

© 2021 Optical Society of America under the terms of the [OSA Open Access Publishing Agreement](#)

## 1. Introduction

Optical 3D measurements based on structured light projection have become a prevalent non-contact 3D shape measurement technique [1]. With the advantages of simple hardware configuration, high measurement accuracy, and high 3D point cloud density, it has been proven one of the most

promising techniques and is extensively applied in industry inspection and scientific research [2–5]. Essentially, the structured light-based 3D measurement methods can be regarded as an improved form of stereo vision, which is achieved by introducing an additional light source generator (such as a projector) in the system configuration [6]. The light source generator projects a series of specifically coded patterns onto the measured scenes [7]. Compared with stereo vision-based methods, the structured light-based 3D measurement methods can easily overcome the problem of low matching accuracy caused by weak texture regions.

Among the 3D shape measurement methods based on structured light projection, two commonly used structured light patterns are fringe patterns and speckle patterns. Correspondingly, there are two mainstream methods: fringe projection profilometry (FPP) [8–11] and speckle projection profilometry (SPP) [12–14]. In FPP, the projector projects a series of fringe patterns onto the measured scenes. The fringe images modulated by the measured objects are captured synchronously by the camera and then processed to obtain the phase information by using various phase retrieval techniques, such as Fourier transform profilometry (FTP) [15–17] and phase-shifting profilometry (PSP) [18]. However, these methods both adopt the arctangent function which can only provide a wrapped phase with  $2\pi$  phase jumps. Therefore, it is necessary to perform phase unwrapping to eliminate the phase ambiguity and convert the wrapped phase to the absolute phase [19–25]. To address this issue, several composite phase-shifting schemes (e.g. dual-frequency PSP [26], bi-frequency PSP [27], and 2+2 PSP [28]) have been proposed, which can solve the phase ambiguity problem without significantly increasing the number of projected patterns. However, these methods still require a certain number of projection patterns. As a result, it is difficult to obtain high-precision and absolute phase information from a single fringe image in FPP, which limits its applications in dynamic 3D measurement [29,30].

Different from FPP, the projector in SPP projects a speckle pattern onto the measured scenes. The speckle images modulated by the measured objects are captured synchronously by the stereo camera and then processed to obtain the disparity map by using various stereo matching techniques. The projected speckle pattern designed using a spatial encoding strategy has inherently global uniqueness, which makes the SPP-based 3D measurement methods have the advantage of single-shot 3D reconstruction. Therefore, the key idea of the design method for the speckle pattern is how to ensure that the local speckles are globally unique with respect to the whole projection pattern [31]. These design methods for projection patterns can be grouped into three main classes based on various spatial encoding strategies [7,32,33]: strategies based on non-formal codification [34,35], strategies based on De Bruijn sequences [36–38], and strategies based on M-arrays [39]. In the last few decades, researchers have proposed numerous design methods for the speckles. However, due to the measured objects with complex reflection characteristics and the perspective differences between the stereo camera, it is still difficult to ensure the global uniqueness of each pixel in the whole measurement space by only projecting one speckle pattern [12,14,40], which leads to the common mismatching in actual measurements. In order to solve this problem in SPP, some robust stereo matching algorithms such as SGM [41–43] and ELAS [44] are proposed to acquire dense disparity maps, thus enabling robust absolute 3D measurement. However, these methods achieve reliable stereo matching by smoothing the disparity map, at the cost of matching accuracy. It is easy to understand that projecting multiple speckle images will improve the accuracy of 3D measurement, because more constraints can be exploited to completely guarantee the global uniqueness of the measured scenes. Following this idea, Zhou *et al.* [14] proposed a high-precision 3D surface profile measurement scheme by only projecting a single-shot color binary speckle pattern (CBSP) and a temporal-spatial correlation matching algorithm, which can be applied to measurements of dynamic and static objects. In order to improve the 3D measurement speed, Schaffer *et al.* [12,13] used laser speckles as projected patterns which are switched using an acousto-optical deflector. Its projection rate is more than 10 times higher than the common projection systems. Capturing images of encoded

objects through two synchronized high-speed cameras, this proposed system achieves high-speed, dense, and accurate 3D measurements of spatially separated objects at 350 frames per second. These proposed SPP methods can achieve high-performance 3D measurement based on speckle projection, but it is impossible to obtain accurate 3D data from a single speckle image. For SPP, it still lacks a stereo matching algorithm using a single speckle pattern that can achieve high-robustness and high-accuracy 3D measurement for the recovery of the fine details of complex surfaces.

Compared with traditional stereo matching methods, recently, many deep learning methods for stereo vision are proposed and have achieved excellent performance of stereo matching [45–52]. There is generally a four-step pipeline for stereo matching, including matching cost calculation, cost aggregation, disparity computation, and disparity refinement, while traditional stereo matching methods perform all four steps using non-learning techniques. Existing learning-based stereo matching methods attempt to exploit deep learning to implement one or multiple of the four steps to obtain better matching results. LeCun *et al.* [45] first adopted the Siamese network to perform block matching for obtaining the initial matching cost and then exploited typical stereo matching procedures, including SGM-based cost aggregation, disparity computation, and disparity refinement to further improve matching results. Luo *et al.* [46] inputted left and right image patches with different sizes into the CNNs for computing the initial matching cost, which will convert the binary classification problem into a multi-classification task, enabling high-efficiency stereo matching. Currently, some end-to-end stereo matching networks have been developed to predict whole disparity maps without post-processing. Kendall *et al.* [49] proposed to generate a 4D cost volume of size  $C \times D \times H \times W$  (*i.e.*,  $Features \times Disparity \times Height \times Width$ ) by combining the features of all pixels from the reference image and all candidates among disparity ranges along the epipolar line of the target image. The 4D cost volume is filtered through a series of 3D convolutional layers. The final disparity maps are regressed from the filtered cost volume using a differentiable soft argmin operation, which allows it to achieve matching results with sub-pixel accuracy without any additional post-processing or regularization. Later, Chang *et al.* [51] proposed a pyramid stereo matching network (PSMNet) to further improve the matching accuracy by using the spatial pyramid pooling and multiple hourglass networks based on the 3D CNN. Zhang *et al.* [52] introduced SGM-based cost aggregation and local guided filter into the existing cost aggregation subnetwork to obtain better matching accuracy and the generalization ability of the network.

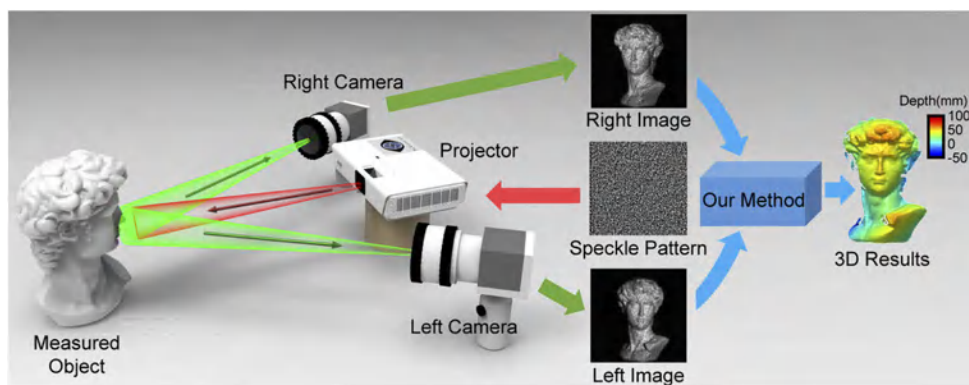
In this work, we propose a single-shot 3D shape measurement method using an end-to-end stereo matching network for speckle projection profilometry. In supervised learning, the use of high-quality datasets, including input data and ground truth, is very important for learning-based methods. KITTI is a prominent stereo dataset, which promoted the development of deep learning in stereo vision [53]. It is worth noting that KITTI is very challenging because its labels obtained by 3D Lidar are extremely sparse and low-precision. In our method, different from KITTI, by combining 12-step PSP [18] and multi-frequency temporal phase unwrapping techniques [22], high-precision absolute phase maps with high completeness can be obtained to generate dense disparity maps with subpixel precision by phase matching, which will be as the high-quality ground truth for our stereo matching network. For the architecture of our proposed network, a multi-scale residual subnetwork is first leveraged to synchronously extract compact feature tensors with 1/4 resolution from speckle images for constructing the 4D cost volume. Considering that the cost filtering operation using 3D convolutional layers is computationally expensive, a lightweight 3D U-net network is proposed to implemented efficient 4D cost aggregation for achieving higher matching performance. In addition, because the disparity maps (as the ground truth) in the SPP dataset has valid values only in the foreground, a simple and fast saliency detection network is integrated into our end-to-end network to avoid predicting the invalid pixels in the disparity maps including occlusions and backgrounds, thereby implicitly enhancing the



matching accuracy for valid pixels. Based on the proposed method, the matching accuracy is improved by about 50% significantly compared with traditional stereo matching methods. The experiment results demonstrated that the proposed method can achieve fast and absolute 3D shape measurement with an accuracy of about  $100\mu\text{m}$  through a single speckle pattern.

## 2. Principle

In this section, a single-shot 3D shape measurement method using an end-to-end stereo matching network for speckle projection profilometry will be presented. In our method, a speckle pattern and a series of fringe patterns need to be projected by the projector onto the measured scenes and captured synchronously by the stereo camera. The acquired speckle image pair is first processed by epipolar rectification, and then fed directly into the proposed end-to-end stereo matching network to obtain the corresponding disparity map without the background. The disparity map is converted into the final 3D results after disparity-to-height mapping as shown in Fig. 1. It is clear that the projected speckle pattern and the end-to-end stereo matching network together determine the actual 3D measurement performance of the proposed method.



**Fig. 1.** The diagram of the proposed single-shot 3D shape measurement method using an end-to-end stereo matching network for speckle projection profilometry.

For the speckle pattern, we follow a simple and effective design and evaluation method proposed in our previous work [31]. By introducing epipolar rectification and depth constraint, the only thing the stereo matching algorithms need to do is to search the corresponding pixel within the pre-defined local 1D range rather than the traditional global 2D range, which means that our optimized design method of the speckle pattern just requires the local speckles in the speckle patterns are unique with respect to the local 1D projection space. Based on this idea, the projected speckle pattern is designed and evaluated to assist in improving the 3D measurement performance.

For the proposed end-to-end stereo matching network, there are two aspects that affect its final stereo matching performance. First, for the deep learning-based network approach, the datasets, including input data and ground truth, are very important to efficiently train the stereo matching network. In our method, a series of acquired fringe images are used to generate dense disparity maps with subpixel precision as the high-quality ground truth for our SPP datasets, which potentially determines the trained network's highest matching accuracy and robustness when measuring objects with complex surfaces. In the next subsection, we will discuss in detail how to construct a high-quality SPP dataset using phase-shifting methods and multi-frequency temporal phase unwrapping techniques in FPP. Secondly, for the architecture of our proposed network, although a large number of high-performance learning-based stereo matching networks

exist, these networks are generally trained and validated on the KITTI stereo dataset and cannot be directly applied to SPP. KITTI is a prominent stereo dataset, which promoted the development of deep learning in stereo vision [53]. It is worth noting that KITTI is very challenging because its labels obtained by 3D Lidar are extremely sparse and low-precision. Specifically, KITTI is a dataset in the field of autonomous driving, in which the data has the properties of large scale and sparse texture, and its 3D reconstruction accuracy is millimeter precision. In contrast, our stereo matching network aims to achieve high-precision and robust 3D measurements with micron-level accuracy by matching the objects with strong speckle texture information. The specific structure of the proposed network will be presented in detail according to Section 2.2.

### 2.1. High-quality SPP dataset constructed by using FPP

To build a high-quality SPP dataset, fringe projection profilometry (FPP) is used to obtain high-precision and dense disparity maps as the ground truth. In a common FPP system, there are three main processing steps in FPP: phase extraction, phase unwrapping, and phase-to-height mapping. During phase recovery, sinusoidal fringe-based FPP methods are more prevalent to retrieval the wrapped phase using Fourier transform methods in frequency domain [15] or phase-shifting methods in time domain [18]. Fourier transform profilometry (FTP) has the advantage of single-shot phase extraction but suffers from the spectrum overlapping problem. These methods generally produce coarse wrapped phases with low quality, making it difficult to achieve high-precision 3D acquisition. Different from FTP, phase-shifting profilometry (PSP) can realize pixel-wise phase measurements with higher accuracy unaffected by ambient light, but it needs to project at least three fringe patterns to obtain a phase map theoretically.

In this work, the standard 12-step phase-shifting fringe patterns with shift offset of  $2\pi/12$  are adopted because it is quite robust to ambient illumination and varying surface properties:

$$I_n^p(x, y) = 0.5 + 0.5 \cos(2\pi fx - 2\pi n/12), \quad (1)$$

where  $I_n^p(x, y)$  ( $n = 0, 1, 2, \dots, 11$ ) represent fringe patterns to be projected,  $f$  is the frequency of fringe patterns. Then the fringe images captured by the camera can be described as

$$I_n^c(x, y) = A^c(x, y) + B^c(x, y) \cos(\phi^c(x, y) - 2\pi n/12), \quad (2)$$

where  $I_n^c(x, y)$  represent the intensity of captured fringe images,  $A^c(x, y)$ ,  $B^c(x, y)$ , and  $\phi^c(x, y)$  are the average intensity, the intensity modulation, and the phase distribution of the measured object. According to the least-squares algorithm, the wrapped phase  $\phi^c(x, y)$ ,  $B^c(x, y)$ , and  $Mask_v^c(x, y)$  can be obtained:

$$\phi^c(x, y) = \tan^{-1} \frac{\sum_{n=0}^{11} I_n^c(x, y) \sin(2\pi n/12)}{\sum_{n=0}^{11} I_n^c(x, y) \cos(2\pi n/12)}, \quad (3)$$

$$B^c(x, y) = \frac{2}{12} \sqrt{\left[ \sum_{n=0}^{11} I_n^c(x, y) \sin(2\pi n/12) \right]^2 + \left[ \sum_{n=0}^{11} I_n^c(x, y) \cos(2\pi n/12) \right]^2}, \quad (4)$$

$$Mask_v^c(x, y) = B^c(x, y)/255 > Thr1, \quad (5)$$

where  $Thr1$  is the preset threshold for the tested object,  $Mask_v^c(x, y)$  can be used to identify the valid points in the whole image. The threshold  $Thr1$  should be changed for object surfaces with different reflectivity, theoretically. In most cases,  $Thr1 = 0.01$  is acceptable for various objects in our measurement. In our method,  $Mask_v^c(x, y)$  is exploited to preprocess the ground truth for enhancing the learning ability of the network to the valid information of the measured scenes.

Due to the truncation effect of the arctangent function in Eq. (3), the obtained phase  $\phi^c(x, y)$  is wrapped within the range of  $(-\pi, \pi]$ , and its relationship with  $\Phi^c(x, y)$  is:

$$\Phi^c(x, y) = \phi^c(x, y) + 2\pi k^c(x, y), \quad (6)$$

where  $k^c(x, y)$  represents the fringe order of  $\Phi^c(x, y)$ , and its value range is from 0 to  $f - 1$ .

In our method, multi-frequency temporal phase unwrapping method (MF-TPU) is exploited to obtain  $k^c(x, y)$  for each pixel in the phase map accurately. In MF-TPU, the wrapped phase  $\phi^c(x, y)$  is unwrapped with the aid of one (or more) additional wrapped phase map with different frequency. For instance, two wrapped phases  $\phi_h^c(x, y)$  and  $\phi_l^c(x, y)$  are both retrieved from phase-shifting algorithms by using Eq. (3), ranging from  $-\pi$  to  $\pi$ . It is easy to find that the two absolute phases  $\Phi_h^c(x, y)$  and  $\Phi_l^c(x, y)$  corresponding to  $\phi_h^c(x, y)$  and  $\phi_l^c(x, y)$  have the following relationship:

$$\begin{cases} \Phi_h^c(x, y) = \phi_h^c(x, y) + 2\pi k_h^c(x, y), \\ \Phi_l^c(x, y) = \phi_l^c(x, y) + 2\pi k_l^c(x, y), \\ \Phi_h^c(x, y) = (f_h/f_l)\Phi_l^c(x, y), \end{cases} \quad (7)$$

where  $f_h$  and  $f_l$  are the frequency of high-frequency fringes and low-frequency fringes. Based on Eq. (7),  $k_h^c(x, y)$  can be calculated by the following formula:

$$k_h^c(x, y) = \frac{(f_h/f_l)\Phi_l^c(x, y) - \phi_h^c(x, y)}{2\pi}. \quad (8)$$

Since the fringe order  $k_h^c(x, y)$  is integer, ranging from 0 to  $f_h - 1$ , Eq. (8) can be adapted as

$$k_h^c(x, y) = \text{Round} \left[ \frac{(f_h/f_l)\Phi_l^c(x, y) - \phi_h^c(x, y)}{2\pi} \right], \quad (9)$$

where  $\text{Round}()$  is the rounding operation. When  $f_l$  is 1, there will be no phase ambiguity so that  $\phi_l^c(x, y)$  is inherently an unwrapped phase. Theoretically, for MF-TPU, this single-period phase can be used to directly assist phase unwrapping of  $\phi_h^c(x, y)$  with relatively higher frequency. However, the phase unwrapping capability of MF-TPU is greatly constrained due to the influence of noise in practice. For a normal FPP system, MF-TPU can only reliably unwrap the phase with about 16 periods due to the non-negligible noises and other error sources in actual measurement. Thus, it generally exploits multiple ( $>2$ ) sets of phases with different frequencies to hierarchically unwrap the wrapped phase step by step, and finally arrives at the absolute phase with high frequency instead of only using the phase with a single period. In our method, three wrapped phases with different frequencies (including 1, 8 and 57) are used to obtain high-precision and dense (57-period) absolute phase.

Finally, phase matching based on the phase information is implemented to obtain the disparity map with integer-pixel precision by minimizing the difference between absolute phases from two perspectives:

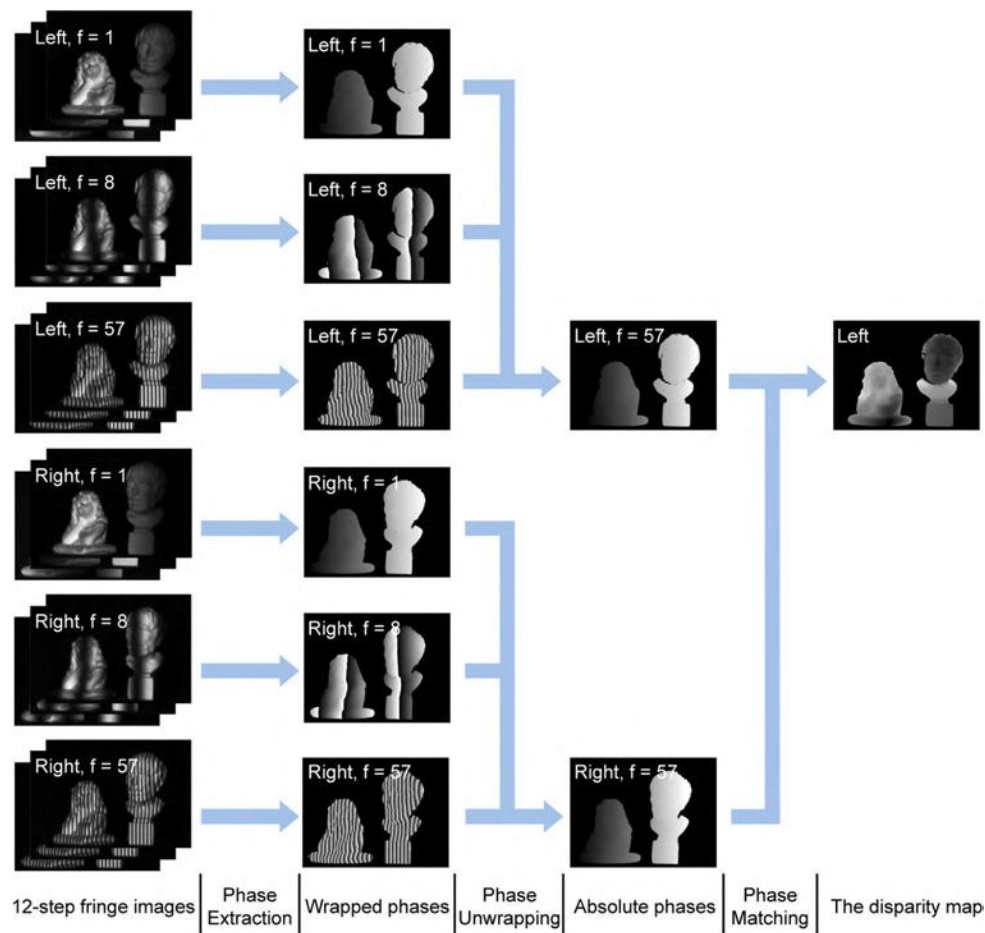
$$\Delta\Phi(i) = \text{abs}(\Phi_L(x, y) - \Phi_R(x + i, y)), \quad (10)$$

$$\Delta\Phi_{\min}(D_{int}) = \min_i \Delta\Phi(i), \quad (11)$$

where  $i$  is the candidate disparity value locally in our SPP system based on epipolar rectification and depth constraint, the disparity  $D_{int}$  represents the pixel-to-pixel correspondence between two camera views. Then, the disparity refinement is realized to obtain the disparity map with subpixel precision by a simple linear interpolation:

$$D_{sub} = D_{int} + \begin{cases} \frac{\Phi_L(x, y) - \Phi_R(x + D_{int}, y)}{\Phi_R(x + D_{int} + 1, y) - \Phi_R(x + D_{int}, y)}, & \Phi_L(x, y) - \Phi_R(x + D_{int}, y) > 0, \\ \frac{\Phi_L(x, y) - \Phi_R(x + D_{int}, y)}{\Phi_R(x + D_{int}, y) - \Phi_R(x + D_{int} - 1, y)}, & \Phi_L(x, y) - \Phi_R(x + D_{int}, y) < 0. \end{cases} \quad (12)$$

By phase matching, the high-precision and dense disparity map  $D_{sub}$  can be obtained as the ground truth of our high-quality SPP dataset in Fig. 2.

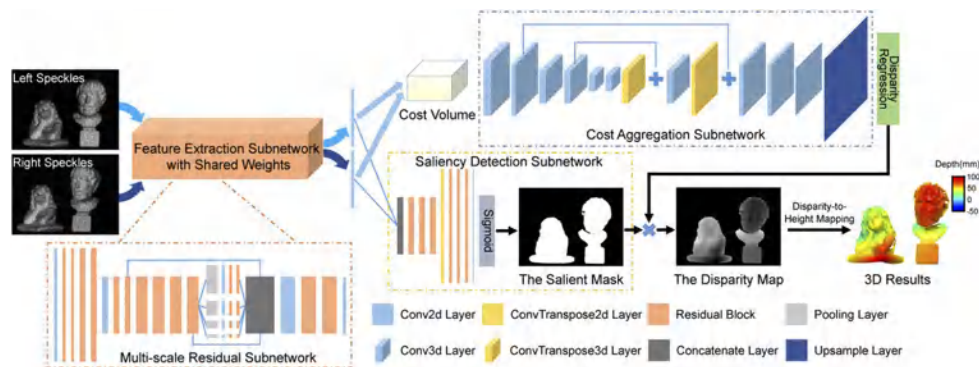


**Fig. 2.** The diagram of constructing high-quality SPP dataset by using FPP.

## 2.2. End-to-end stereo matching network

In this subsection, an end-to-end stereo matching network, which is used to solve the stereo matching problem in SPP, is proposed to substantially promote the matching accuracy compared with the state-of-the-art stereo matching methods. Existing high-performance learning-based stereo matching networks are generally trained and validated on the KITTI stereo dataset. In the KITTI stereo dataset, the data has the properties of large scale and sparse texture, and the corresponding 3D reconstruction results have only millimeter precision. In contrast, based on our high-quality SPP dataset, our stereo matching network aims to achieve robust 3D measurements with micron-level accuracy using a speckle image pair. In addition, for the ground truth of our SPP dataset, the disparity map of the sample data has valid values only in the foreground as shown in Fig. 2. Thus, it is difficult to naively exploit these existing end-to-end networks [50–52] to directly obtain the final disparity map, but a simple and fast saliency detection network is integrated into our network to avoid predicting the invalid pixels in the disparity maps including occlusions and backgrounds. Specifically, the schematic diagram of the proposed stereo matching network is shown in Fig. 3.





**Fig. 3.** The schematic diagram of the proposed end-to-end stereo matching network. The whole stereo matching network is composed of a multi-scale residual subnetwork (as the shared feature extraction subnetwork), construction of the 4D cost volume, cost aggregation using 3D convolutional layers, disparity regression, and a saliency detection subnetwork.

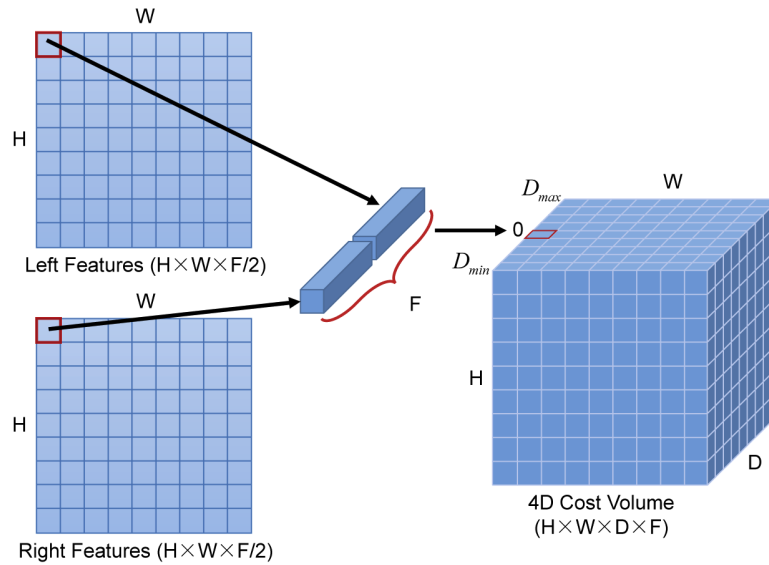
In Fig. 3, the whole stereo matching network is composed of a multi-scale residual subnetwork (as the shared feature extraction subnetwork), construction of the 4D cost volume, cost aggregation using 3D convolutional layers, disparity regression, and a saliency detection subnetwork. It is worth noting that before stereo matching epipolar rectification is first executed to simplify the two-dimensional search problem to a one-dimensional matching problem [54]. Then, in feature extraction for matching cost calculation, different from the traditional methods that directly exploit the gray information or color value of the pixel for correspondence matching, our purpose is to calculate the feature representation of each pixel to be matched for the subsequent matching process. Specifically, learning-based methods usually implement feature extraction on the input stereo images simultaneously to obtain rich feature information, which is used to construct a 4D cost volume as the initial matching cost. Therefore, the initial matching accuracy corresponding to the initial matching cost strongly depends on the quality of the extracted feature information.

For the feature extraction subnetwork in our work, a multi-scale residual network is proposed to process the input stereo image pair to obtain rich multi-scale feature information. In this subnetwork, speckle images are first processed by a 2D convolution layer and four residual blocks to obtain 64-channel feature tensors. Considering that the high-resolution matching costs in the subsequent cost aggregation will consume a lot of computational overhead and take up expensive GPU memories, it is necessary to perform a 1/4 downsample operation on the feature tensors. It is worth noting that the extraction of low-resolution feature tensors is not so much a compromise to the expensive computational cost but to keep the feature tensors more compact and achieve high-efficiency feature extraction. Then, the low-resolution feature tensors successively go through six residual blocks for further expanding the receptive field of each pixel of output tensors. It is crucially important that each pixel of feature tensors yielded by the network must have a larger receptive field so that the network will not ignore any important feature information during the prediction period [55]. And then, the multi-scale pooling layers are introduced to downsample the input tensors by 1/4, 1/16, 1/64, and 1/256, which can further compress and extract the main features of the tensors to reduce computation complexity and prevent over-fitting. For these four downsample paths, the feature tensors are all processed sequentially by a convolutional layer, a group of residual blocks, and an upsample layer implemented by bilinear interpolation. After the feature tensors from these six paths are gathered, the concatenate layer is applied for the feature combination along the channel axis. Finally, the feature tensors are processed by a 2D convolution layer, two residual blocks, and a 2D convolution layer without ReLU to obtain 32-channel feature tensors with 1/4 resolution.

At the next stage, for constructing the 4D cost volume, feature tensors of each pixel in the left image and all corresponding candidates in the local disparity range on the epipolar line of the right image are concatenated. The initial 4D cost volume of dimensionality  $H \times W \times D \times F$  (i.e., *Height*  $\times$  *Width*  $\times$  *Disparity*  $\times$  *Features*) is built as shown in Fig. 4:

$$\begin{aligned} \text{Cost}(:, 1 : (W - D_i), D_i - D_{\min} + 1, 1 : \frac{F}{2}) &= \text{Feature}_{\text{left}}(:, 1 : (W - D_i), :), \\ \text{Cost}(:, 1 : (W - D_i), D_i - D_{\min} + 1, (\frac{F}{2} + 1) : F) &= \text{Feature}_{\text{right}}(:, (D_i + 1) : W, :), \end{aligned} \quad (13)$$

where  $\text{Feature}_{\text{left}}$  and  $\text{Feature}_{\text{right}}$  represent the feature tensors with 1/4 resolution from two perspectives output by the feature extraction subnetwork, their size ( $H \times W \times F/2$ ) is  $240 \times 320 \times 32$  for the  $480 \times 640$  resolution of the cameras.  $[2D_{\min}, 2D_{\max}]$  is the disparity range of our SPP system. For feature tensors with 1/4 resolution, the initial 4D cost volume is built based on the range  $[D_{\min}, D_{\max}]$ .  $D_i$  is a candidate disparity in the range  $[D_{\min}, D_{\max}]$ .  $D$  is the absolute disparity range ( $D_{\max} - D_{\min} + 1$ ).



**Fig. 4.** The schematic diagram of the construction of the 4D cost volume. Based on the disparity range of our SPP system, the initial 4D cost volume is built by combining feature tensors of each pixel in the left image and all corresponding candidates along the epipolar line of the right image.

In cost aggregation, the initial 4D cost volume will be further optimized using 3D convolutional layers. Although some downsample operations have been done during feature extraction, in fact, the 4D cost volume with 1/4 resolution still occupies a lot of GPU memories. Therefore, a lightweight 3D U-net network is proposed to achieve efficient 4D cost aggregation. First of all, three sets of 3D convolutional layers are adopted to realize cost filtering and downsample the 4D cost volume by 1/4. Then, the ConvTranspose3d layer is used to upsample the cost volume, and combined with shortcut operations to achieve residual aggregation. According to the output of the residual operations, three 3D convolutional layers are used to acquire a 4D cost volume with a single-channel feature, and subsequently obtain the final full-resolution 4D cost volume through an upsample layer.

Disparity regression in [49] is introduced to estimate the disparity map based on the final 4D cost volume with a single-channel feature. The probability of each candidate disparity  $D_i$  is first

calculated using the softmax operation for the predicted cost volume. The predicted disparity map  $Disparity(x, y)$  is procured by the weighted sum of the normalized probability for each candidate disparity  $D_i$ :

$$Disparity(x, y) = \sum_{D_i=2D_{min}}^{2D_{max}} D_i \times softmax(Cost(x, y, D_i)). \quad (14)$$

The traditional stereo matching network directly calculates the loss between the predicted disparity map and the ground-truth for training. But for the dataset built in our SPP system, the disparity map of the sample data has valid values only in the foreground. Therefore, it is necessary to integrate an additional saliency detection network into our existing network. Currently, the learning-based saliency detection method has been widely investigated with its advantages of high accuracy, high efficiency, and low cost. Among them, fully convolutional network (FCN) is one of the most promising network architectures and has achieved significant results on various well-known datasets [56]. However, given the dataset of SPP that the spatial structure of the tested scenes is relatively simple and the saliency objects have strong speckle texture information, a saliency detection network based on a simple network structure can also achieve good detection results. In order to avoid extracting redundant features, the feature tensors from two perspectives output by the feature extraction subnetwork are directly stacked through a concatenate layer. And then, through a group of residual blocks, a ConvTranspose2d layer, another group of residual blocks, and a convolutional layer, the feature tensors are sequentially filtered and upsampled to obtain a single-channel feature tensor with full resolution. Finally, the sigmoid function is used to achieve the regression of the saliency detection mask  $Mask(x, y)$ , enabling the prediction of the disparity map without the background:

$$Disparity_{train}(x, y) = Disparity(x, y) \times Mask(x, y). \quad (15)$$

During training, we used *Adam* to minimize the joint loss, thereby updating the weights that parameterize the network. The joint loss consists of a smooth L1 loss for the disparity map and a binary cross-entropy loss for the saliency mask:

$$Loss = Loss_{Mask} + Loss_{Disparity}, \quad (16)$$

$$Loss_{Mask} = -\frac{1}{N} \sum_{n=1}^N [Mask_v^c(n) \ln Mask(n) + (1 - Mask_v^c(n)) \ln(1 - Mask(n))], \quad (17)$$

$$Loss_{Disparity} = \frac{1}{N} \sum_{n=1}^N smooth_{L_1}(D_{sub}(n) - Disparity_{train}(n)), \quad (18)$$

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1, \\ |x| - 0.5, & \text{otherwise.} \end{cases} \quad (19)$$

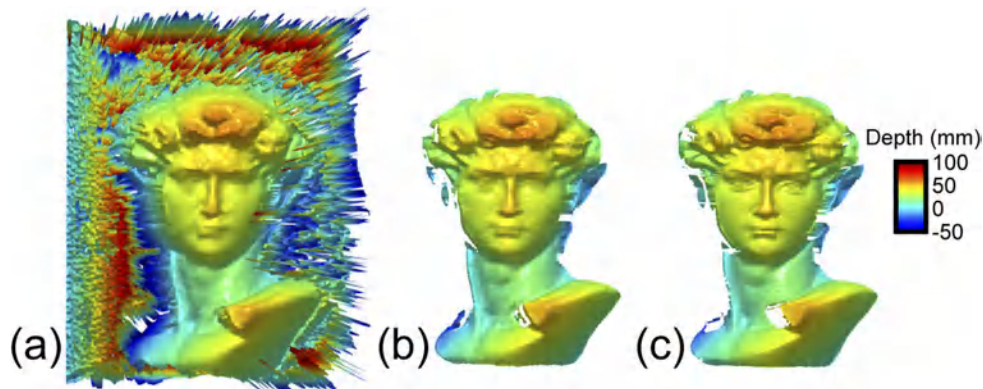
where  $Mask_v^c$  and  $D_{sub}$  are the corresponding ground truth of the saliency mask and the disparity map according to Section 2.1.

During testing, the saliency detection mask  $Mask(x, y)$  needs to be binarized to distinguish the foreground from the background, and the final disparity map is obtained:

$$Disparity_{final}(x, y) = Disparity(x, y), \quad \text{if } Mask(x, y) >= 0.75. \quad (20)$$

To verify the actual impact of the saliency detection network, the comparison of the 3D reconstruction results without/with the saliency detection network is presented as shown in Fig. 5.

It can be found in Fig. 5 that our measurement results without the saliency detection network have serious mismatches in the background, which will affect the convergence of the network during training and reduce the actual performance of the network. Therefore, the saliency detection network is an additional but necessary module in our approach, implicitly enhancing the matching accuracy for valid pixels.



**Fig. 5.** Comparison of the 3D reconstruction results without/with the saliency detection network. (a) the 3D reconstruction results without the saliency detection network. (b) the 3D reconstruction results with the saliency detection network. (c) the ground truth.

### 3. Experiments

To verify the actual 3D measurement performance of the proposed method, a common stereo vision-based SPP system with a wide baseline is built as shown in Fig. 1, which consists of two monochrome cameras (Basler acA640-750um with the resolution of  $640 \times 480$ ) and a DLP projector (LightCrafter 4500Pro with the resolution of  $912 \times 1140$ ). Since the baseline between the stereo cameras is about  $270\text{mm}$ , the disparity constraint of our system should be suitably set to  $-100$  to  $59$  pixels to measure objects with a depth range of  $-100\text{mm}$  to  $100\text{mm}$ . The distance between the measurement system and the objects to be tested is about  $900\text{mm}$ . In addition, the projected speckle pattern has been designed and evaluated based on our previous work [31] to obtain the best 3D measurement performance.

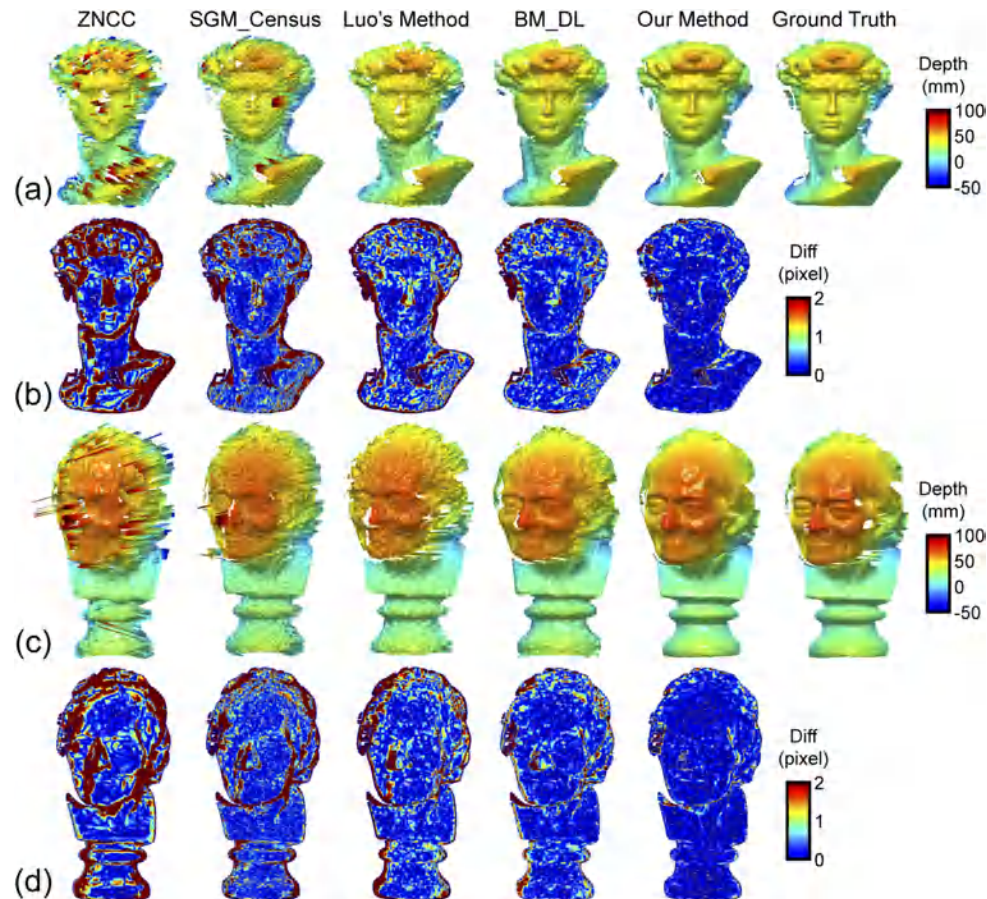
In our experiment, we collected the dataset including 1200 different scenes, which are randomly composed of 30 simple and complex objects. The whole dataset has 1200 image pairs, which are divided into 800 image pairs for training, 200 image pairs for validation, and 200 image pairs for testing. During training, to monitor the accuracy of the neural networks for samples that they have never seen, the scenes in these training, verification, and testing datasets are separate from each other. In addition, to achieve high-robustness and high-accuracy stereo matching, the proposed stereo matching network can only process a pair of stereo images at a time during training, which occupies about 23GB graphic memories. The training epoch is set as 200 which takes about 5 days. The proposed network takes 0.95 seconds for disparity prediction.

#### 3.1. Experimental comparison of different methods

A comparative experiment is first carried out to reveal the high performance of the proposed method compared with two traditional methods (ZNCC [57] and SGM\_Census [41,42]) and two learning-based methods (Luo's method [46] and BM\_DL proposed in our previous work [55]). Measuring the objects with ridged, complex, or discontinuous surfaces is a challenging task for a single-shot SPP. To verify the reliability of these methods for scanning these challenging



surfaces, two different objects are measured including the David model and the statue of Voltaire. The corresponding 3D reconstruction results obtained by ZNCC, SGM\_Census, Luo's method, BM\_DL, and our method are shown in Figs. 6(a) and (c).



**Fig. 6.** Comparison of the 3D reconstruction results using different methods. (a) the 3D reconstruction results of the David model, (b) the matching errors of the David model, (c) the 3D reconstruction results of the statue of Voltaire, (d) the matching errors of the statue of Voltaire.

The ZNCC criterion is highly common for practical use, as it is insensitive to the offset and scale changes in the intensity of the local matched block and provides the most accurate and reliable displacement estimations compared with other criteria [57]. In ZNCC, block matching is performed to calculate the matching costs and acquire the integer-pixel disparity maps, which then can be refined to obtain the sub-pixel disparity maps by a five-point quadratic curve fitting model [14]. In order to enhance the matching performance of ZNCC, the block size in block matching is determined as  $19 \times 19$  after an exhaustive empirical search. However, the fundamental assumption made by block matching is that all the pixels in the matching window have similar disparities. As a consequence, this assumption does not hold at disparity discontinuities, causing the corresponding 3D results with the edge-fattening issue [58,59] in object boundaries and thin structures as shown in Fig. 6.

Compared with ZNCC, SGM\_Census can provide dense 3D measurement results. In SGM\_Census, the census transform with the same block size of  $19 \times 19$  is applied to calculate the initial matching costs, which are then processed to obtain the 3D results using a series of post-processing operations including 1D cost aggregation from 8 paths, Winner-Take-All (WTA), and a quadratic curve fitting [41]. However, SGM\_Census avoids mismatching by smoothing the disparity map for achieving reliable stereo matching, at the cost of 3D measurement accuracy as shown in Fig. 6. It can be found that there are some obvious mismatch areas and low-precision 3D measurement results using ZNCC and SGM\_Census, which proves that these non-parametric matching methods are so difficult to provide reliable and high-precision matching results on the SPP system with a wide baseline.

Different from these traditional methods, two learning-based methods (Luo's method and BM\_DL) are also implemented for comparison. In the two methods, matching cost calculation is implemented using the network. In Luo's method, a pair of block data (centered on the point to be matched in the left image and its all corresponding candidate points in the right image) is inputted into the network at the same time to search the correct candidate point within the pre-defined local disparity range. To realize the high performance of stereo matching, a block matching network based on the Siamese structure is adopted to generate better initial matching costs. Similar to SGM\_Census, a series of same post-processing operations are used to obtain the 3D results as shown in Fig. 6. Furthermore, BM\_DL proposed in our previous work is an enhanced version of Luo's method. In the block matching network of BM\_DL, some additional but necessary convolutional layers and residual blocks are stacked at the head of the network to further enhance the ability of feature extraction. Besides, the fully connected layers with shared weights are used instead of the original inner product to improve the accuracy of the network's similarity measurement. It is easy to find in Fig. 6 that BM\_DL can output more accurate and dense disparity results compared with SGM\_Census and Luo's method. However, the measurement accuracy achieved by BM\_DL cannot meet the requirements of high-precision 3D measurement applications. It is important that how to leverage the end-to-end network to achieve more efficient three-dimensional matching is worth investigating.

Obviously, in Fig. 6, the proposed end-to-end stereo matching network yields the highest-quality 3D reconstruction by the single-shot measurement. Compare with the ground truth using the 12-step phase-shifting fringe patterns as shown in Fig. 6, due to the inherent characteristics of local smoothness for stereo matching, there are some local details with slight distortion and blurred surfaces in our 3D reconstruction results. However, it can be found that our method can obtain high-precision 3D results that are closer to the ground truth. It is easy to conclude based on these experimental results that our matching network can achieve 3D measurements with the best performance among several SPP methods.

Besides, compared with the ground truth, the matching errors for different methods are shown in Figs. 6(b) and 6(d) and the corresponding quantitative analysis results can be found in Table 1. To ensure the objectivity of the analysis results, the differences between the disparity results obtained using these methods and the ground truth are used to make an accurate judgment. The number of points is the sum of valid points in the ground truth. The missing ratio means the proportion of points that are valid points in the ground truth but invalid points in these disparity results. For ZNCC, SGM\_Census, Luo's method, and BM\_DL, the 4-connected image segmentation method is used to process the disparity maps to identify and remove segments with fewer pixels [41]. For our method, the mask generated by the saliency detection subnetwork is exploited to directly remove the invalid pixels in the disparity maps including occlusions and backgrounds. Then the error ratio is easily obtained by counting the number of valid points where their absolute disparity difference between the ground truth and these disparity results are more than 1 pixel. All remaining valid points are regarded as correct points and then further subdivided according to different disparity accuracies including 1 pixel, 0.5 pixels, and 0.2 pixels. It can

be seen from Table 1 that the missing ratio and the error ratio using our method are lower than 2% and 6%. The correctness ratio achieved by our method is higher than 93%, and most of the pixels have a disparity accuracy of lower than 0.5 pixels. The results illustrated that the matching accuracy using the proposed method is improved by about 50% significantly compared with traditional stereo matching methods. Our method can achieve robust 3D shape measurement with a high correctness ratio and high completeness for objects with complex surfaces and geometric discontinuities.

**Table 1. Quantitative analysis results for different methods**

Object	Nop <sup>a</sup>	Method	Mmr <sup>b</sup> (%)	Emr <sup>c</sup> (%)	Cmr <sup>d</sup> (%)		
					$\leq 1^e$	$\leq 0.5^f$	$\leq 0.2^g$
David	62337	ZNCC	16.92	34.77	48.31	33.98	16.93
		SGM_Census	10.09	22.18	67.73	49.21	23.73
		Luo's method	7.91	18.26	73.83	52.36	25.05
		BM_DL	3.44	13.24	83.32	64.26	33.18
		Our method	<b>1.57</b>	<b>5.34</b>	<b>93.09</b>	<b>83.67</b>	<b>56.79</b>
Voltaire	75403	ZNCC	10.91	28.07	61.02	45.92	22.63
		SGM_Census	6.11	18.51	75.38	56.96	28.13
		Luo's method	6.97	14.71	78.32	59.06	39.31
		BM_DL	2.29	9.69	88.02	72.11	39.48
		Our method	<b>0.68</b>	<b>2.84</b>	<b>96.48</b>	<b>89.48</b>	<b>62.64</b>

<sup>a</sup>Nop = Number of points,

<sup>b</sup>Mmr = Missing matching rate,

<sup>c</sup>Emr = Error matching rate,

<sup>d</sup>Cmr = Correct matching rate,

<sup>e</sup> $\leq 1$  = Less than 1 pixel,

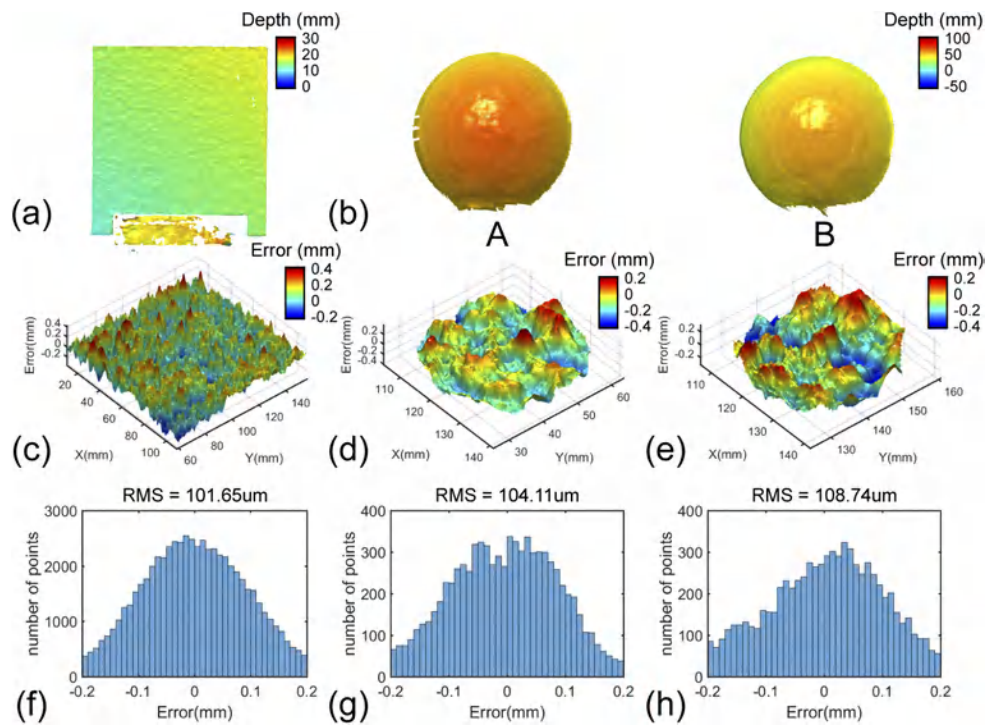
<sup>f</sup> $\leq 0.5$  = Less than 0.5 pixels,

<sup>g</sup> $\leq 0.2$  = Less than 0.2 pixels.

### 3.2. Precision analysis

Further, to quantitatively evaluate the accuracy of our system using the proposed end-to-end stereo matching network, a ceramic plane and a pair of standard ceramic spheres with a diameter of 50.8mm are measured. Figures 7(a) and 7(b) show the corresponding 3D reconstruction results. And then, based on the obtained 3D reconstruction data, the plane fitting is performed to acquire the ideal plane as the ground truth. The difference between the measured plane and the ideal plane is calculated to obtain the 3D measured errors as shown in Fig. 7(c). The quantitative histograms of the differences are displayed as shown in Fig. 7(f). It can be easily found that the major measured errors are less than 200 $\mu\text{m}$  with the RMS of 101.65 $\mu\text{m}$ , respectively. Likewise, for the 3D measurement of a pair of standard ceramic spheres as shown in Fig. 7(b), the sphere fitting is used to obtain the actual measurement error as shown in Figs. 7(d) and 7(e). Then, the RMS of the 3D measurement accuracy is about 100 $\mu\text{m}$  as shown in Figs. 7(g) and 7(h).

In addition, the precision analysis results for different methods are presented in Table 2. For the ceramic plane, the measurement errors achieved using ZNCC are less than 200 $\mu\text{m}$  with the RMS of 103.04 $\mu\text{m}$ . The reason for this result is that based on the basic assumption of block matching all pixels in the matching window have similar disparities. However, this assumption does not hold for measuring objects with ridged, complex, or discontinuous faces. For the standard ceramic spheres, ZNCC can only generate coarse 3D measurement results with many matching errors as shown in Fig. 8. It is noted that by the sphere fitting the actual measurement errors are greater than 1mm. After outlier removal, the measurement accuracy has been improved



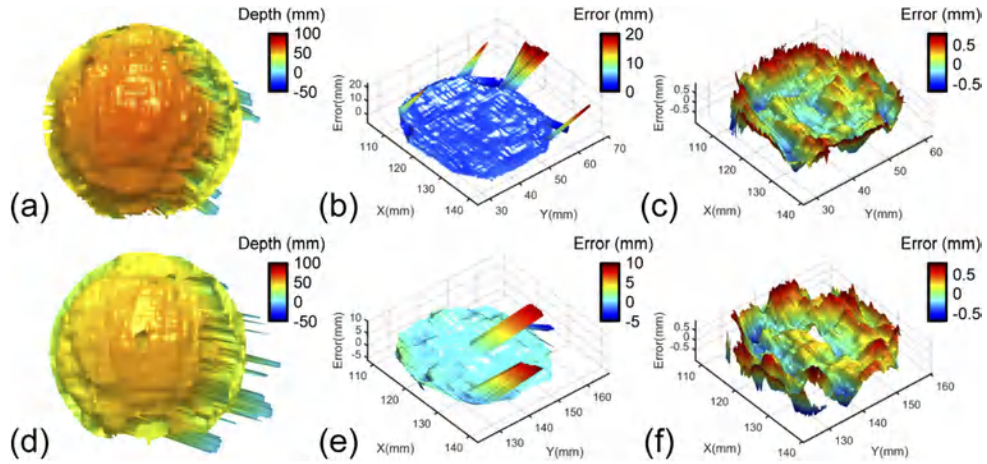
**Fig. 7.** Precision analysis for measuring a ceramic plane and a pair of standard ceramic spheres using our method. (a) The 3D reconstruction results of a ceramic plane, (b) the 3D reconstruction results of a pair of standard ceramic spheres, (c)-(e) the corresponding distributions of the measured errors of (a)-(b), and (f)-(h) the corresponding quantitative histograms of the measured errors of (a)-(b).

significantly but is still greater than  $300\mu m$ . And the radius error of the tested ceramic spheres using ZNCC is greater than  $1mm$  in Table 2. In contrast, SGM\_Census provides measurement results with similar accuracy for measuring planes and spheres. Similarly, Luo's method and BM\_DL can also realize robust and more accurate measurements for measuring planes and spheres. However, these methods all use the same post-processing operations to achieve reliable stereo matching by smoothing the disparity map, at the cost of matching accuracy. Unlike these methods, whether the planes or spheres are measured, and whether RMS or radius errors of the spheres are calculated, our method can achieve robust 3D shape measurement with the best accuracy. This result verifies that the proposed method can significantly increase the matching accuracy of SPP and achieve high-precision 3D reconstruction results.

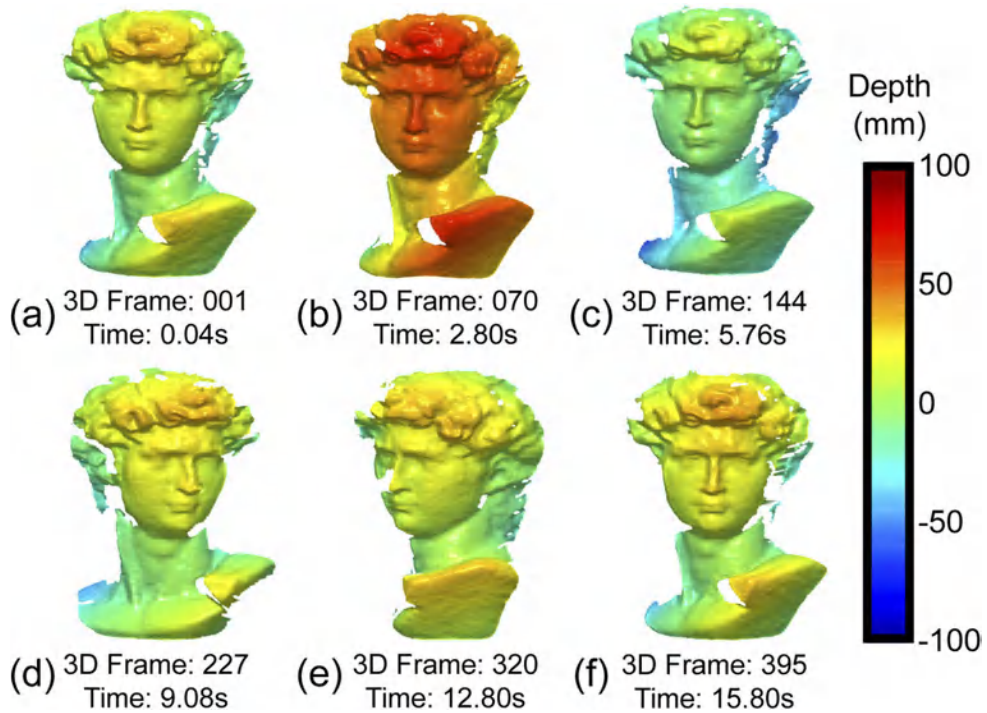
**Table 2. Precision analysis results for different methods**

Method	Ceramic plane		Ceramic sphere A		Ceramic sphere B	
	RMS	RMS	Radius error	RMS	Radius error	
ZNCC	$103.04\mu m$	$340.62\mu m$	$1.31mm$	$367.25\mu m$	$1.02mm$	
SGM_Census	$279.39\mu m$	$332.11\mu m$	$244.35\mu m$	$345.76\mu m$	$260.72\mu m$	
Luo's method	$240.12\mu m$	$292.65\mu m$	$213.05\mu m$	$274.82\mu m$	$229.35\mu m$	
BM_DL	$189.85\mu m$	$213.66\mu m$	$172.85\mu m$	$207.79\mu m$	$166.57\mu m$	
Our method	<b><math>101.65\mu m</math></b>	<b><math>104.11\mu m</math></b>	<b><math>117.85\mu m</math></b>	<b><math>108.74\mu m</math></b>	<b><math>114.13\mu m</math></b>	





**Fig. 8.** Precision analysis for measuring a pair of standard ceramic spheres using ZNCC. (a) The 3D reconstruction results of ceramic spheres A, (b) the corresponding distributions of the measured errors of (a), (c) the corresponding distributions of the measured errors of (a) after outlier removal, (d) the 3D reconstruction results of ceramic spheres B, (e) the corresponding distributions of the measured errors of (d), and (f) the corresponding distributions of the measured errors of (d) after outlier removal.



**Fig. 9.** The 3D reconstruction results for a dynamic scene: a moving David model (Visualization 1). (a)-(c) The David model moves along the Z axis and (d)-(f) the David model rotates around the Y axis.

### 3.3. Fast 3D surface imaging

Last of all, our system is applied to record a dynamic scene for fast 3D shape measurement: a moving David model as shown in Fig. 9. In this experiment, the exposure time of cameras is set 39.2ms to capture the speckle images at the speed of 25Hz for achieving 3D reconstruction at 25fps. Figure 9 shows the color-coded 3D reconstruction results at different time points. During the whole dynamic measurement, the David model first moves forward along the Z axis, and arrives at the boundary of the predefined measurement space at 2.8 seconds. Then, the David model moves in reverse along the Z axis to another boundary of the predefined measurement space at 5.76 seconds. Furthermore, the David model returns to the initial position and starts to rotate around the Y axis. Finally, it is back to the origin position again in 15.8 seconds. The whole 3D measurement results can refer to Visualization 1. In the whole measuring procedures, the 3D surfaces of the David model are correctly and high-quality reconstructed, verifying the reliability of the proposed method to perform the absolute 3D shape measurement with high completeness at high speed.

## 4. Conclusion

In summary, we proposed a single-shot 3D shape measurement method using an end-to-end stereo matching network based on a common stereo vision-based SPP system. To efficiently train the stereo matching network, a high-quality SPP dataset is first built by combining phase-shifting profilometry (PSP) and temporal phase unwrapping techniques in FPP. High-precision absolute phase maps obtained using FPP are used to generate accurate and dense disparity maps with high completeness as the ground truth of the dataset by phase matching. For the architecture of the network, the proposed network first leverages a multi-scale residual subnetwork to synchronously extract compact feature tensors with 1/4 resolution from speckle images for constructing the 4D cost volume. Although some downsample operations have been done during feature extraction, in fact, the 4D cost volume with 1/4 resolution still occupies a lot of GPU memories. Therefore, a lightweight 3D U-net network is proposed to implemented efficient 4D cost aggregation for achieving higher matching performance. Considering that the disparity maps (as the ground truth) in the SPP dataset has valid values only in the foreground, a simple and fast saliency detection network is proposed and integrated into our network to avoid enhancing the invalid pixels in the disparity maps including occlusions and backgrounds, thereby implicitly enhancing the matching accuracy for valid pixels. The experimental comparison of different methods illustrated that compared with traditional methods our method can achieve robust 3D shape measurement with a high correctness ratio and high completeness for objects with complex surfaces. Besides, the quantitative analysis results proved again that the matching accuracy using the proposed method is improved by about 50% significantly compared with traditional stereo matching methods. The experiment results of the precision analysis demonstrated that the proposed method can achieve absolute 3D shape measurement with an accuracy of about  $100\mu\text{m}$  through only a single speckle pattern. The dynamic measurement experiment has verified the success of the proposed method in its ability to effectively achieve fast and accurate 3D shape measurements with high completeness for complex scenes at 25fps.

Finally, there are several aspects that need to be further improved in the proposed method. First, since there are many costly 3D convolutions for cost aggregation in the proposed network, the initial cost volume is 1/4 downsampled in advance, which undoubtedly reduces the accuracy of stereo matching significantly. Therefore, how to achieve more efficient cost aggregation is still a problem to be solved. Second, it is easy to understand that projecting multiple speckle images will improve the accuracy of 3D measurement, because more constraints can be exploited to completely guarantee the global uniqueness of the measured scenes. How to improve the measurement accuracy of the stereo matching network by inputting multiple speckle images at the same time is another interesting direction for further investigation. Third, the proposed

network takes 0.95 seconds for disparity prediction that is slower compared with most of the existing algorithms running on GPU. How to achieve fast stereo matching should be considered. It can be found that cost aggregation in the proposed network take accounts for most of the total run time. Similarly, the cost aggregation sub-network should be further optimized to improve the accuracy of stereo matching and reduce the run time. At last, different from traditional non-learning methods, it is noted that the generalization ability of learning methods needs to be further researched and discussed for measuring different objects with complex reflection characteristics or high reflectivity, enabling more reliable 3D shape measurement. Based on the above analysis, we will explore more other methods to design a single-shot SPP system with higher performance.

**Funding.** National Defense Science and Technology Foundation of China (2019-JCJQ-JJ-381); National Key Research and Development Program of China (2017YFF0106403); Leading Technology of Jiangsu Basic Research Plan (BK20192003); “333 Engineering” Research Project of Jiangsu Province (BRA2016407); Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging and Intelligence Sense (3091801410411); Fundamental Research Funds for the Central Universities (30919011222, 30920032101); National Natural Science Foundation of China (61705105, 61722506, 62005121, 62075096).

**Disclosures.** The authors declare no conflicts of interest.

## References

1. S. S. Gorthi and P. Rastogi, “Fringe projection techniques: whither we are?” *Opt. Laser Eng.* **48**(2), 133–140 (2010).
2. S. Feng, L. Zhang, C. Zuo, T. Tao, Q. Chen, and G. Gu, “High dynamic range 3d measurements with fringe projection profilometry: a review,” *Meas. Sci. Technol.* **29**(12), 122001 (2018).
3. Z. Zhang, “Review of single-shot 3d shape measurement by phase calculation-based fringe projection techniques,” *Opt. Laser Eng.* **50**(8), 1097–1106 (2012).
4. W. Yin, S. Feng, T. Tao, L. Huang, S. Zhang, Q. Chen, and C. Zuo, “Calibration method for panoramic 3d shape measurement with plane mirrors,” *Opt. Express* **27**(25), 36538–36550 (2019).
5. Q. Zhang and X. Su, “High-speed optical measurement for the drumhead vibration,” *Opt. Express* **13**(8), 3110–3116 (2005).
6. Z. Zhang, S. Huang, S. Meng, F. Gao, and X. Jiang, “A simple, flexible and automatic 3d calibration method for a phase calculation-based fringe projection imaging system,” *Opt. Express* **21**(10), 12218–12227 (2013).
7. J. Salvi, J. Pages, and J. Batlle, “Pattern codification strategies in structured light systems,” *Pattern Recognition* **37**(4), 827–849 (2004).
8. S. Zhang, “High-speed 3d shape measurement with structured light methods: A review,” *Opt. Laser Eng.* **106**, 119–131 (2018).
9. C. Zuo, T. Tao, S. Feng, L. Huang, A. Asundi, and Q. Chen, “Micro fourier transform profilometry ( $\mu$ ftp): 3d shape measurement at 10,000 frames per second,” *Opt. Laser Eng.* **102**, 70–91 (2018).
10. S. Zhang, “Absolute phase retrieval methods for digital fringe projection profilometry: A review,” *Opt. Laser Eng.* **107**, 28–37 (2018).
11. W. Yin, C. Zuo, S. Feng, T. Tao, Y. Hu, L. Huang, J. Ma, and Q. Chen, “High-speed three-dimensional shape measurement using geometry-constraint-based number-theoretical phase unwrapping,” *Opt. Laser Eng.* **115**, 21–31 (2019).
12. M. Schaffer, M. Grosse, B. Harendt, and R. Kowarschik, “High-speed three-dimensional shape measurements of objects with laser speckles and acousto-optical deflection,” *Opt. Lett.* **36**(16), 3097–3099 (2011).
13. M. Schaffer, M. Grosse, and R. Kowarschik, “High-speed pattern projection for three-dimensional shape measurement using laser speckles,” *Appl. Opt.* **49**(18), 3622–3629 (2010).
14. P. Zhou, J. Zhu, and H. Jing, “Optical 3-d surface reconstruction with color binary speckle pattern encoding,” *Opt. Express* **26**(3), 3452–3465 (2018).
15. X. Su and W. Chen, “Fourier transform profilometry: a review,” *Opt. Laser Eng.* **35**(5), 263–284 (2001).
16. Q. Kemao, “Two-dimensional windowed fourier transform for fringe pattern analysis: principles, applications and implementations,” *Opt. Laser Eng.* **45**(2), 304–317 (2007).
17. S. Feng, Q. Chen, G. Gu, T. Tao, L. Zhang, Y. Hu, W. Yin, and C. Zuo, “Fringe pattern analysis using deep learning,” *Adv. Photonics* **1**(2), 025001 (2019).
18. C. Zuo, S. Feng, L. Huang, T. Tao, W. Yin, and Q. Chen, “Phase shifting algorithms for fringe projection profilometry: A review,” *Opt. Laser Eng.* **109**, 23–59 (2018).
19. X. Su and W. Chen, “Reliability-guided phase unwrapping algorithm: a review,” *Opt. Laser Eng.* **42**(3), 245–261 (2004).
20. M. Zhao, L. Huang, Q. Zhang, X. Su, A. Asundi, and Q. Kemao, “Quality-guided phase unwrapping technique: comparison of quality maps and guiding strategies,” *Appl. Opt.* **50**(33), 6214–6224 (2011).
21. Y. Wang and S. Zhang, “Novel phase-coding method for absolute phase retrieval,” *Opt. Lett.* **37**(11), 2067–2069 (2012).

22. C. Zuo, L. Huang, M. Zhang, Q. Chen, and A. Asundi, "Temporal phase unwrapping algorithms for fringe projection profilometry: A comparative review," *Opt. Laser Eng.* **85**, 84–103 (2016).
23. K. Zhong, Z. Li, Y. Shi, C. Wang, and Y. Lei, "Fast phase measurement profilometry for arbitrary shape objects without phase unwrapping," *Opt. Laser Eng.* **51**(11), 1213–1222 (2013).
24. X. Liu, Y. Yang, Q. Tang, Z. Cai, X. Peng, M. Liu, and Q. Li, "A method for fast 3d fringe projection measurement without phase unwrapping," in *Sixth International Conference on Optical and Photonic Engineering (icOPEN 2018)*, vol. 10827 (International Society for Optics and Photonics, 2018), p. 1082713.
25. W. Yin, Q. Chen, S. Feng, T. Tao, L. Huang, M. Trusiak, A. Asundi, and C. Zuo, "Temporal phase unwrapping using deep learning," *Sci. Rep.* **9**(1), 20175 (2019).
26. K. Liu, Y. Wang, D. L. Lau, Q. Hao, and L. G. Hassebrook, "Dual-frequency pattern scheme for high-speed 3-d shape measurement," *Opt. Express* **18**(5), 5229–5244 (2010).
27. C. Zuo, Q. Chen, G. Gu, S. Feng, and F. Feng, "High-speed three-dimensional profilometry for multiple objects with complex shapes," *Opt. Express* **20**(17), 19493–19510 (2012).
28. C. Zuo, Q. Chen, G. Gu, S. Feng, F. Feng, R. Li, and G. Shen, "High-speed three-dimensional shape measurement for dynamic scenes using bi-frequency tripolar pulse-width-modulation fringe projection," *Opt. Laser Eng.* **51**(8), 953–960 (2013).
29. X. Su and Q. Zhang, "Dynamic 3-d shape measurement method: a review," *Opt. Laser Eng.* **48**(2), 191–204 (2010).
30. S. Feng, C. Zuo, T. Tao, Y. Hu, M. Zhang, Q. Chen, and G. Gu, "Robust dynamic 3-d measurements with motion-compensated phase-shifting profilometry," *Opt. Laser Eng.* **103**, 127–138 (2018).
31. W. Yin, S. Feng, T. Tao, L. Huang, M. Trusiak, Q. Chen, and C. Zuo, "High-speed 3d shape measurement using the optimized composite fringe patterns and stereo-assisted structured light system," *Opt. Express* **27**(3), 2411–2431 (2019).
32. B. Pan, Z. Lu, and H. Xie, "Mean intensity gradient: an effective global parameter for quality assessment of the speckle patterns used in digital image correlation," *Opt. Laser Eng.* **48**(4), 469–477 (2010).
33. Z. Chen, X. Shao, X. Xu, and X. He, "Optimized digital speckle patterns for digital image correlation by consideration of both accuracy and efficiency," *Appl. Opt.* **57**(4), 884–893 (2018).
34. M. Ito and A. Ishii, "A three-level checkerboard pattern (tcp) projection method for curved surface measurement," *Pattern Recognit.* **28**(1), 27–40 (1995).
35. M. Maruyama and S. Abe, "Range sensing by projecting multiple slits with random cuts," *IEEE Trans. Pattern Anal. Machine Intell.* **15**(6), 647–651 (1993).
36. K. L. Boyer and A. C. Kak, "Color-encoded structured light for rapid active ranging," *IEEE Transactions on Pattern Analysis Mach. Intell.* pp. 14–28 (1987).
37. L. Zhang, B. Curless, and S. M. Seitz, "Rapid shape acquisition using color structured light and multi-pass dynamic programming," in *First International Symposium on 3D Data Processing Visualization and Transmission*, (IEEE, 2002), pp. 24–36.
38. J. Pagès, J. Salvi, C. Collewet, and J. Forest, "Optimised de bruijn patterns for one-shot shape acquisition," *Image Vis. Comput.* **23**(8), 707–720 (2005).
39. H. Morita, K. Yajima, and S. Sakata, "Reconstruction of surfaces of 3-d objects by m-array pattern projection method," in *1988 IEEE Conference on International Conference on Computer Vision*, (IEEE, 1988), pp. 468–473.
40. S. Heist, P. Dietrich, M. Landmann, P. Kühmstedt, G. Notni, and A. Tünnermann, "Gobo projection for 3d measurements at highest frame rates: a performance analysis," *Light: Sci. Appl.* **7**(1), 71 (2018).
41. H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(2), 328–341 (2008).
42. H. Hirschmuller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(9), 1582–1599 (2009).
43. F. Gu, Z. Song, and Z. Zhao, "Single-shot structured light sensor for 3d dense and dynamic reconstruction," *Sensors* **20**(4), 1094 (2020).
44. A. Geiger, M. Roser, and R. Urtasun, "Efficient large-scale stereo matching," in *Asian conference on computer vision*, (Springer, 2010), pp. 25–38.
45. J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *2015 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE, 2015), pp. 1592–1599.
46. W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE, 2016), pp. 5695–5703.
47. J. Pang, W. Sun, J. S. Ren, C. Yang, and Q. Yan, "Cascade residual learning: A two-stage convolutional neural network for stereo matching," in *2017 IEEE Conference on International Conference on Computer Vision Workshops*, (IEEE, 2017), pp. 887–895.
48. N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE, 2016), pp. 4040–4048.
49. A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *2017 IEEE Conference on International Conference on Computer Vision*, (IEEE, 2017), pp. 66–75.



50. S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, and S. Izadi, "Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction," in *2018 IEEE Conference on European Conference on Computer Vision (ECCV)*, (IEEE, 2018), pp. 573–590.
51. J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *2018 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE, 2018), pp. 5410–5418.
52. F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "Ga-net: Guided aggregation net for end-to-end stereo matching," in *2019 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE, 2019), pp. 185–194.
53. A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, (IEEE, 2012), pp. 3354–3361.
54. R. Hartley and A. Zisserman, *Multiple view geometry in computer vision* (Cambridge University, 2003).
55. W. Yin, J. Zhong, S. Feng, T. Tao, J. Han, L. Huang, Q. Chen, and C. Zuo, "Composite deep learning framework for absolute 3d shape measurement based on single fringe phase retrieval and speckle correlation," *JPhysPhotonics* **2**, 045009 (2020).
56. A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Comp. Visual Media* **5**(2), 117–150 (2019).
57. B. Pan, H. Xie, and Z. Wang, "Equivalence of digital image correlation criteria for pattern matching," *Appl. Opt.* **49**(28), 5501–5509 (2010).
58. D. Min, J. Lu, and M. N. Do, "A revisit to cost aggregation in stereo matching: How far can we reduce its computational redundancy?" in *2011 International Conference on Computer Vision*, (IEEE, 2011), pp. 1567–1574.
59. D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.* **47**(1/3), 7–42 (2002).



# Composite fringe projection deep learning profilometry for single-shot absolute 3D shape measurement

YIXUAN LI,<sup>1,2</sup>  JIAMING QIAN,<sup>1,2</sup>  SHIJIE FENG,<sup>1,2,3</sup>  QIAN CHEN,<sup>2,4</sup>  AND CHAO ZUO<sup>1,2,\*</sup> 

<sup>1</sup>Smart Computational Imaging (SCI) Laboratory, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China

<sup>2</sup>Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China

<sup>3</sup>shijiefeng@njust.edu.cn

<sup>4</sup>chenqian@njust.edu.cn

\*zuochao@njust.edu.cn

**Abstract:** Single-shot fringe projection profilometry (FPP) is essential for retrieving the absolute depth information of the objects in high-speed dynamic scenes. High-precision 3D reconstruction using only one single pattern has become the ultimate goal in FPP. The frequency-multiplexing (FM) method is a promising strategy for realizing single-shot absolute 3D measurement by compounding multi-frequency fringe information for phase unwrapping. In order to solve the problem of serious spectrum aliasing caused by multiplexing schemes that cannot be removed by traditional spectrum analysis algorithms, we apply deep learning to frequency multiplexing composite fringe projection and propose a composite fringe projection deep learning profilometry (CDLP). By combining physical model and data-driven approaches, we demonstrate that the model generated by training an improved deep convolutional neural network can directly perform high-precision and unambiguous phase retrieval on a single-shot spatial frequency multiplexing composite fringe image. Experiments on both static and dynamic scenes demonstrate that our method can retrieve robust and unambiguous phases information while avoiding spectrum aliasing and reconstruct high-quality absolute 3D surfaces of objects only by projecting a single composite fringe image.

© 2022 Optica Publishing Group under the terms of the [Optica Open Access Publishing Agreement](#)

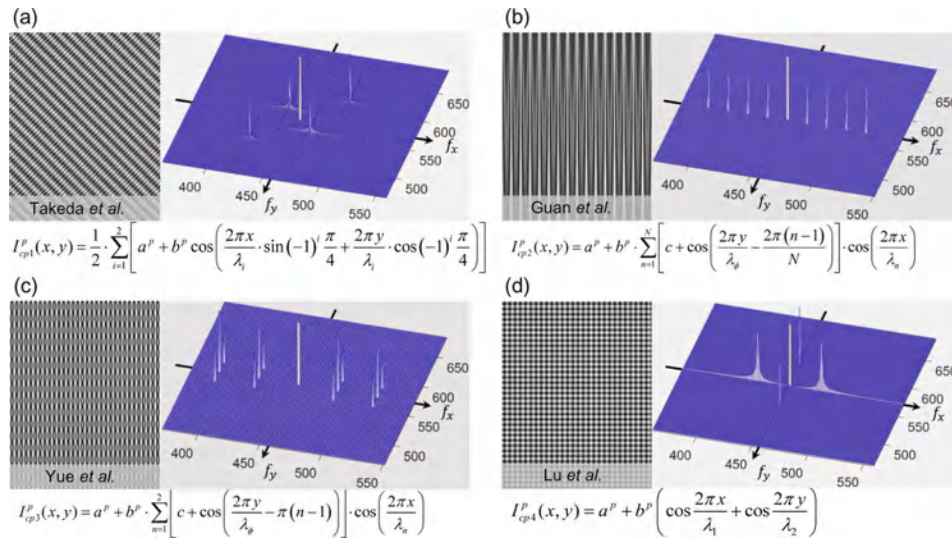
## 1. Introduction

Structured light (SL) projection is one of the most representative 3D optical imaging technologies for macroscopic objects due to its non-contact, high-resolution, and easy-to-implement measurement capabilities [1–4]. Among them, fringe projection profilometry (FPP), has become one of the most prevalent SL methods with the advantages of full-field scanning and high-precision measurement [5–7], which has been widely applied in multiple fields, such as intelligent manufacturing [8] and reverse engineering [9]. For FPP, the projector projects a series of fringe patterns onto the target object, and then the camera captures these images modulated and deformed by the objects. The measured objects' absolute phases and related depth information can be retrieved with the captured fringe patterns by processing the following three steps: fringe analysis, phase unwrapping, and phase-to-height mapping. With the rapid development of optoelectronic information technology [10–12], people subsequently set higher expectations on FPP, requiring both higher precision and higher speed. However, these two aspects seem contradictory in nature. Due to the increasing demand for dynamic scene measurement (such as online industrial inspection, stress deformation analysis, fast reverse modeling, etc. [13]), “speed” has gradually become a fundamental factor that must be taken into account when using FPP. There are two factors

determining the 3D measurement speed of FPP: (1) the speed of hardware devices: optoelectronic devices (e.g., digital light projectors, spatial light modulators, and high-speed image sensors) and digital signal processing units (e.g., high-performance computers and embedded processors); and (2) the number of patterns required per 3D reconstruction of the software algorithms. From the perspective of hardware, Lei and Zhang *et al.* [14,15] has achieved speed breakthroughs by developing the binary defocusing techniques, where the binary defocusing methods coincide with the inherent operation mechanism of the digital-light-processing (DLP) technology, and permit tens of kHz fringe projection speed by using a digital micromirror array device (DMD). Heist *et al.* [16] introduced a GOBO projector that projects aperiodic sinusoidal fringe patterns with high frame rates and high radiant flux, which can generate more than 1,000 independent point clouds per second. In addition, Zuo *et al.* [17] proposed micro Fourier profilometry ( $\mu$ FTP), which used high-speed fringe projection hardware as well as the number of patterns reduction strategy to achieve 3D shapes reconstruction at 10,000 fps. From the perspective of algorithms, several composite phase-shifting methods have been proposed to reduce the number of projected patterns required per unambiguous 3D reconstruction [8,17–24]. Liu *et al.* [18] proposed a dual-frequency pattern strategy that embedded low and high frequency components into a single pattern, at least five patterns were required to reconstruct the 3D point cloud. Zuo *et al.* [20] employed two  $\pi/2$  phase-shifting sinusoidal patterns and two linear increasing/decreasing ramp patterns to reduce the number of patterns required per 3D reconstruction from five to four. Zhang *et al.* [21] embedded the speckle-like signal in three sinusoidal phase-shifted fringe patterns for absolute depth recovery, which can eliminate the phase ambiguity without reducing the fringe amplitude or frequency. Feng *et al.* [22] presented a two-frame fringe projection technique for real-time 3D measurement, using a speckle image and a speckle-embedded fringe image. Tao *et al.* [23] used three composite fringe patterns embedded with the triangular wave into a multi-view system to strengthen the robustness of phase unwrapping. Qian *et al.* [8] further established a complete multi-view fringe projection system, which can achieve real-time high-precision 360-degree 3D model measurement with only three high-frequency fringe patterns. Nevertheless, high-precision 3D reconstruction using only one single pattern is a considerable challenge and has been the ultimate goal of structured light 3D imaging in perpetual pursuit. In 1983, Takeda *et al.* [25,26] proposed Fourier transform profilometry (FTP), which decoded the wrapped phase by Fourier filtering in the spatial frequency domain and achieved the phase demodulation from a single fringe pattern. Afterwards, a series of influential and improved single-shot fringe analysis methods were proposed, such as windowed Fourier transform (WFT) [27–30] and wavelet transform (WT) methods [31]. Particularly, Su *et al.* [32] applied a single high-frequency fringe projection image for drumhead vibration at a speed of kHz level. However, the key to the success of FTP is that the high-frequency fringe information modulated by the object surface can be well separated from the background intensity in the frequency domain. As a result, the FTP technique [33] is limited to measuring smooth surfaces with limited height variations. Besides, the phase distribution retrieved by FTP, ranging between  $-\pi$  to  $\pi$ , suffers from  $2\pi$  periodic ambiguity. Consequently, the wrapped phases require phase unwrapping algorithms to further obtain the absolute phase distribution [34].

To achieve single-shot phase unwrapping, Takeda *et al.* [35] further introduced frequency multiplexing (FM) to FPP to encode two fringe patterns with different spatial carriers into a single snapshot measurement. The projected composite fringe pattern and its spectrum and intensity calculation function are illustrated in Fig. 1(a). After performing the Fourier transform on the FM composite fringe pattern, the spatial frequencies in two orthogonal directions can be extracted from the spectrum simultaneously, with which the periodic phase ambiguity can be removed. Although this method solved the spectrum aliasing problem to some extent and can measure 3D objects with discontinuous and isolated surfaces, the residual phase errors still lead to phase unwrapping errors. Guan *et al.* [36] used four high-frequency carrier information to convolve

the single-frequency phase-shifting fringe patterns to different positions of the Fourier spectrum (Fig. 1(b)), by which the absolute phase maps can be directly recovered from these modulated single-frequency signals through the temporal phase-shifting algorithm [37]. However, due to the weak anti-noise ability of low-frequency image and residual spectrum aliasing, this method cannot be applied to highly precision measurement fields. Yue *et al.* [38] designed another composite structured light pattern formed by modulating two fringe patterns with  $\pi$  phase difference along the orthogonal direction of two distinct carrier frequencies (Fig. 1(c)). Lu *et al.* [39] proposed a fast modulation measuring profilometry based on a single-shot cross grating projection to reconstruct the 3D shape of the objects (Fig. 1(d)). In addition some other single-shot composite coding strategies, such as spatial neighborhood coding schemes [40] (e.g., De Bruijn sequences), color channels multiplexing methods [41], and directly coding methods [42] can also solve the problem of motion. Although the above methods achieve high measurement efficiency, they suffer from compromised measurement accuracy due to the spectrum aliasing problem of FTP. In recent years, deep learning technology has been applied to FPP as a new tool to solve the measurement efficiency and phase/or depth retrieval accuracy [43,44], such as fringe analysis [44–46], fringe enhancement [47], phase demodulation [48,49], phase unwrapping [50–52], and 3D data acquisition [53–55]. These studies have given us an inspiration-whether it is possible to combine fringe projection profilometry with deep learning techniques to achieve higher precision and more robust phase retrieval and 3D reconstruction from only a single composite fringe image.



**Fig. 1.** Spatial frequency multiplexing composite fringe patterns and their corresponding spectrum.

In this work, we present a novel composite fringe projection deep learning profilometry (CDLP), which construct a one-to-three convolutional neural network to analyze the single-shot spatial frequency multiplexing fringe pattern, to reconstruct high-quality 3D shape information in transient scenes. The main contributions of this work are as follows. (1) Under supervised learning, the use of high-quality data sets (including input data and ground truth labels) significantly affects the quality of network model training. In this regard, we first propose a fringe encoding scheme based on spatial frequency multiplexing, which takes into account both unambiguity and multi-spatial information fusion to enhance the training abilities of the deep learning network. Meanwhile, through the N-step phase-shifting (PS) method and multi-frequency temporal phase unwrapping (TPU) combined with the projection distance minimization (PDM) algorithm [17],



we have successively obtained the high-quality wrapped phase numerator, denominator term, and unambiguous unwrapped phase. They serve as three sets of high-quality ground truth labels for our one-to-three single-shot phase retrieval network. (2) For our proposed network framework, we use an improved one-to-three deep convolution neural network to simultaneously achieve high-quality phase analysis and robust phase retrieval. The specific network architecture will be elaborated in subsequent sections. Experimental results validated that the proposed single-shot composite fringe projection deep learning profilometry can directly perform the single-shot robust and unambiguous phase retrieval process and high-quality absolute 3D surface information of the objects under fast, dynamic, and even transient scenes, with a reconstruction accuracy of 60  $\mu\text{m}$ . The remainder of this paper is organized as follows. In Section 2., we start with the basic principles of composite fringe projection deep learning profilometry (CDLP). In Section 3., experimental verifications and comparison results are presented in detail. In the final Section 4., we draw conclusions.

## 2. Principle of composite fringe projection deep learning profilometry (CDLP)

### 2.1. Single composite fringe pattern (CFP) coding scheme

To incorporate multiple patterns in an image as the unique input of a deep learning network to achieve a single-shot fast and robust phase retrieval, we consider the following considerations for the composite fringe coding strategy. (1) It is hard to directly use a absolute single-frequency fringe pattern as a deep learning input to predict the high-precision phase. The reason is that when the single-frequency fringe pattern is projected and captured, the camera resolution is inconsistent with the projector's resolution. Thus only one single-frequency fringe cannot obtain complete sinusoidal intensity information and accurate phase information within a sinusoidal period. (2) Although the absolute phase information of the objects can be obtained directly by using the single-frequency N-step phase-shifting fringe images without phase unwrapping [56], the phase accuracy demodulated by the single-frequency fringe is poor; besides, this strategy always needs to project N fringe images, so the speed cannot exceed the single-frame projection. (3) If we consider combining N single-frequency fringe patterns into an image with a complete sine period, we cannot directly superimpose these N single-frequency phase-shifting images into one image, because the final composite result is a white image. (4) Guan's coding strategy [36], which separated the four single-frequency phase-shifting patterns in the frequency spectrum through four high-frequency carrier frequencies, can directly retrieve the unambiguous phase distributions from a single composite image and avoid the complicated unwrapping process; however, due to the low frequencies lack of high-frequency detail information of the objects, this method cannot fulfill the requirements of high-precision 3D shape measurement. To sum up, we propose a novel three short-wavelength superimposed three carrier-frequency composite fringe pattern coding strategy.

The composite fringe pattern (CFP) and its generation process are shown in Fig. 2(a). Firstly, we generate three sets of sinusoidal fringe patterns with different short wavelengths (or high frequencies), which are recorded as fringe patterns to be modulated:

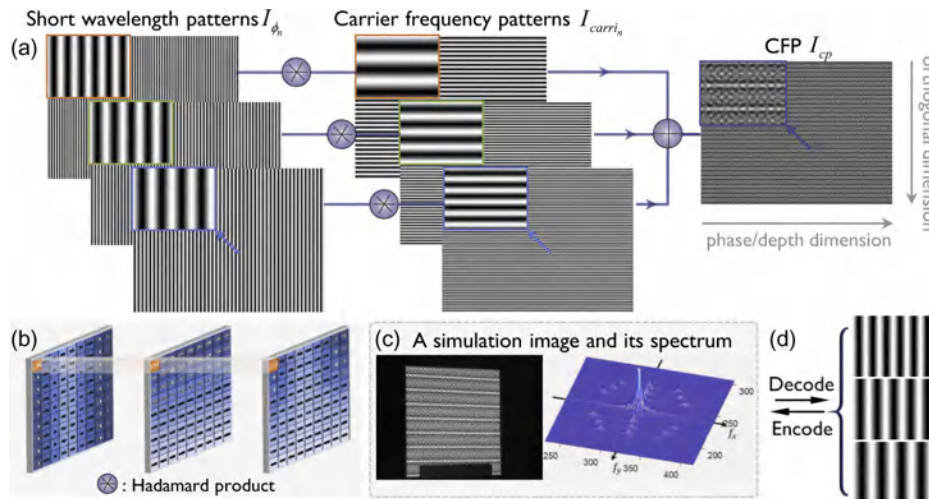
$$I_{\phi_n}^p(x^p, y^p) = a + b \cos\left(\frac{2\pi x^p}{\lambda_{\phi_n}}\right), \quad (1)$$

where  $(x^p, y^p)$  is the projector pixel coordinate. The constants  $a$  and  $b$  are the background intensity and the modulation of the short-wavelength fringe patterns, and their values should strictly meet the following constraints: on the one hand, the  $\cos$  term of  $I_{\phi_n}^p$  after compensation by  $a$  and  $b$  is non-negative, on the other hand, they must ensure that the patterns can be made to reach the maximum contrast ratio. The wavelengths  $\lambda_{\phi_n}$  are changed in the phase direction (the  $x^p$  dimension). Constant  $n$  represents the  $n$ th short-wavelength fringe pattern index,  $n = 1, 2, 3$ .

Then, the short-wavelength fringe patterns  $I_{\phi_n}^p$  are respectively multiplied by three standard cosine fringe patterns  $I_{carrier_n}$  with different carrier frequencies along the orthogonal direction to produce three composite sub-images. By superimposing these three composite sub-images of each channel, a frequency-multiplexed CFP is generated:

$$\begin{aligned}
 I_{cp}^p(x^p, y^p) &= A + B \cdot \sum_{n=1}^3 I_{\phi_n}^p(x^p, y^p) \circ I_{carrier_n}(x^p, y^p) \\
 &= A + B \cdot \sum_{n=1}^3 \left[ a + b \cos\left(\frac{2\pi x^p}{\lambda_{\phi_n}}\right) \right] \cdot \cos(2\pi f_{carrier_n} y^p),
 \end{aligned}
 \tag{2}$$

where  $I_{\phi_n}^p$  is the intensity of the projected CFP,  $A$  and  $B$  are the mean intensity and the amplitude constants to make value of the 8-bit CFP  $I_{\phi_n}^p$  between 0 and 255. The operator  $\circ$  represents the Hadamard product operation, which is a pixel-level calculation process, shown in Fig. 2(b). The frequencies  $f_{carrier_n}$  that change in the orthogonal direction (the  $y^p$  dimension) are recorded as the carrier frequency. The designed CFP contains three short wavelengths (modulation frequencies)  $\lambda_{\phi_n}$  and three carrier frequencies  $f_{carrier_n}$ . The directions between the short-wavelength fringes  $I_{\phi_n}^p$  and the carrier-frequency fringes  $I_{carrier_n}$  are orthogonal, so that the modulation frequencies corresponding to the short wavelengths can be modulated at different positions of the Fourier spectrum through different carrier frequencies. Appropriate short wavelengths and carrier frequencies have to be carefully assigned. The selection conditions of the three short-wavelength  $\lambda_{\phi_n}$  we will discuss in the next section. For the selection of carrier frequency  $f_{carrier_n}$  combinations, in order to expand the bandwidth of each modulation channel and minimize channel leakage, the selected carrier frequencies should be separated as much as possible and far away from zero frequency. However, limited by the spatial resolution of the projector and camera, they have to be restricted within a certain range to ensure reliable phase retrieval.



**Fig. 2.** The composite fringe pattern (CFP) generation process and details. (a) A CFP is formed by modulating and superimposing three short-wavelength fringe patterns with three carrier-frequency fringe patterns along with orthogonal directions. (b) Hadamard product operation between the images of the first channel. (c) A simulation composite fringe image and its spectrum. (d) Conversion between CFP and three fringe patterns.

It should be noted that the intensity range of the projected CFP should be controlled at  $[0,255]$ , so the intensity of the originally generated CFP needs to be normalized

$$I_{cp}^p(x^p, y^p)' = \frac{I_{cp}^p(x^p, y^p) - I_{\min}}{I_{\max} - I_{\min}} \cdot 255, \quad (3)$$

where  $[I_{\min}, I_{\max}]$  is the intensity range of the original CFP. Ideally, after illuminating the object with the composite fringe pattern  $I_{cp}^p$  through a digital projector, the intensities of the captured image can be expressed as:

$$I_{cp}^c(x^c, y^c) = \alpha(x^c, y^c) \cdot \left[ A + B \cdot \sum_{n=1}^3 I_{\phi_n}^c(x^c, y^c) \cdot I_{carrin}(x^c, y^c) \right], \quad (4)$$

where the fringe maps to be demodulated are

$$I_{\phi_n}^c(x^c, y^c) = a + b \cos \Phi_n(x^c, y^c), \quad (5)$$

and  $(x^c, y^c)$  is the pixel coordinate in the camera space,  $\alpha(x^c, y^c)$  is the surface reflectivity of the measured object, and  $\Phi_n(x^c, y^c)$  is the absolute phase. Due to perspective distortion between the projector and the camera, the actual carrier frequencies  $f_{carrin}^c$  of the captured image in the camera view may be different from  $f_{carrin}$ . Thus, the relative position of the projector and the camera should be aligned to share about the same world coordinates both in orthogonal direction and the depth direction.

From Eqs. (4) and (5), we can see that the composite fringe image contains three short-wavelength fringe maps (Fig. 2(d)): the three different wavelength fringe patterns can be encoded as one pattern, and the composite fringe pattern can also be decoded to recover these three different wavelength fringe patterns. The phase information of the three fringe maps can be demodulated separately, and then the absolute phase of the object can be retrieved by the phase retrieval algorithm. Therefore, how to accurately demodulate the phase information of one of the short-wavelength fringe images from the obtained distorted composite fringe images is one of the focuses of this work.

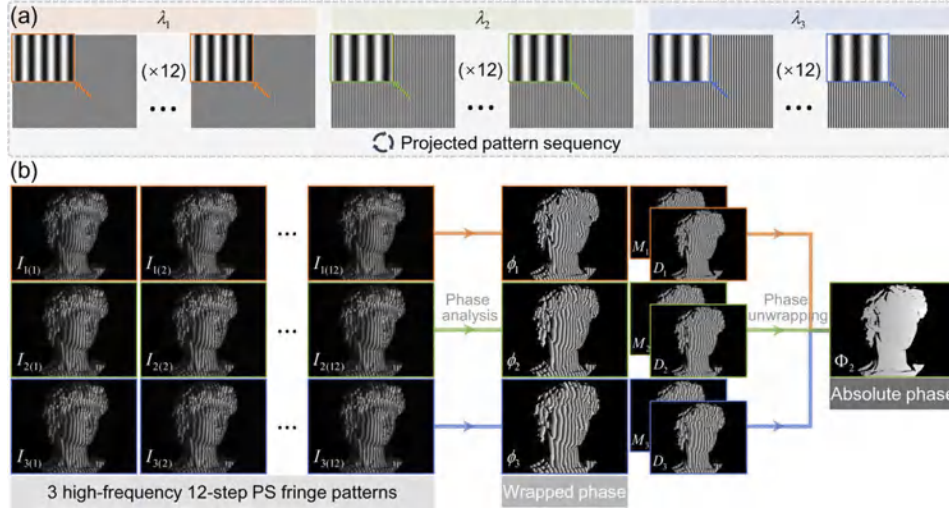
## 2.2. Construction of high-quality network datasets

In order to keep the extracted phase information free from spectrum aliasing, we use deep learning-based methods instead of traditional Fourier transform methods to perform phase demodulation and absolute phase recovery. To begin with, we need to build a high-quality network dataset. The network model learned based on the simulation data may not realistically and comprehensively reflect the actual physical imaging process, and it may not obtain the ideal imaging results. Therefore, we collect and label actual experimental training data rather than simulation data in the deep learning task. In this work, we use a non-composite standard N-step phase-shifting algorithm (PS) for high-quality phase analysis, and use temporal phase unwrapping (TPU) combined with the projection distance minimization (PDM) method to obtain high-precision absolute phase information. The complete process of constructing a high-quality network dataset is shown in Fig. 3.

For the standard N-step phase-shifting algorithm, the fringe images captured by the camera can be expressed as:

$$I_{n(i)}^c(x^c, y^c) = A(x^c, y^c) + B(x^c, y^c) \cos(\Phi_n(x^c, y^c) + 2\pi i/N), \quad (6)$$

where  $I_{n(i)}^c$  represents the  $(i+1)$ th  $n$ -frequency captured image,  $i = 0, 1, \dots, N-1$ ,  $\Phi_n$  is the  $n$ -frequency absolute phase map, and  $2\pi i/N$  is the phase shift. Then, the phase can be calculated



**Fig. 3.** The process of generating training data. (a) The projected sequence consists of three sets of 12-step phase-shifting fringe patterns with different frequencies/wavelengths. (b) The generation process includes projecting and capturing three sets of fringe images, phase analysis to obtain the wrapped phase, and phase unwrapping to retrieve the absolute phase distribution.

through the least-squares algorithm:

$$\phi = \arctan \frac{\sum_{i=0}^{N-1} I_i^c \sin(2\pi i/N)}{\sum_{i=0}^{N-1} I_i^c \cos(2\pi i/N)} = \arctan \frac{M}{D}, \quad (7)$$

where the subscripts  $(x^c, y^c)$  and  $n$  are omitted for convenience,  $M$  and  $D$  represent the numerator and denominator of the arctangent function, respectively. In addition, to improve the image quality and enhance the learning ability of the deep learning network, the image *Mask* function constructed by the modulation function  $B$  (Eq. (8)) is used to remove the invalid points of the entire captured image (Eq. (9)).

$$B(x^c, y^c) = \frac{2}{N} \sqrt{M^2 + D^2}, \quad (8)$$

$$\text{Mask}(x^c, y^c) = \begin{cases} B(x^c, y^c), & B(x^c, y^c) \geq \text{Thr} \\ 0, & B(x^c, y^c) < \text{Thr} \end{cases}. \quad (9)$$

The value of threshold  $\text{Thr}$  is set to 8, which is suitable for most of our measurement scenarios in this work. The initial phase  $\varphi$  we obtained is the relative/wrapped phase within  $(-\pi, \pi)$  due to the truncation of the arctangent function. Thus, we need to perform phase unwrapping to remove the ambiguities and correctly extract the absolute phase contribution. In this work, through the obtained multi-frequency fringe wrapped phase maps, we use the temporal phase unwrapping method to eliminate the phase ambiguity in the time domain pixel by pixel. Projection distance minimization (PDM) is an optimal method for solving multi-frequency temporal phase unwrapping. Assuming that three groups of fringe patterns with fringe wavelength  $\lambda = [\lambda_1, \lambda_2, \lambda_3]^T$  are obtained by the phase-shifting method, the corresponding relative phase is  $\varphi = [\phi_1, \phi_2, \phi_3]^T$ , and the absolute phase  $\Phi = [\Phi_1, \Phi_2, \Phi_3]^T$  and the wrapped phase satisfy the



following relationship:

$$\Phi = \varphi + 2\mathbf{k}\pi, \quad (10)$$

where  $\mathbf{k} = [k_1, k_2, k_3]^T$  is the integer fringe order vector,  $k_{1,2,3} \in [0, K - 1]$ , and  $K$  denotes the number of used fringes. The task of phase unwrapping is to determine the unique fringe orders  $\mathbf{k}$  of the wrapped phase, and then obtain the absolute phase maps  $\Phi$  from Eq. (10). To achieve the goal that the relative phase  $\varphi$  can be successfully unwrapped without ambiguities within the desired measurement range, the fringe wavelength combination should be selected appropriately. On the one hand, given that the projection pattern has  $W$  pixels along the horizontal axis wherein the sinusoidal fringe intensity change, on the other hand, considering the fact that the least common multiple of the wavelength combination determines the maximum range of unambiguous phase wrapping along the absolute phase axial direction [17,57], the selected three different wavelengths  $\lambda_1, \lambda_2, \lambda_3$  should satisfy the following inequality to exclude phase ambiguous:

$$LCM(\lambda_1, \lambda_2, \lambda_3) \geq W, \quad (11)$$

where  $LCM()$  represents the least common multiple function. Refer to the optimal wavelength selection strategy [17,58], the wavelengths  $\lambda_n$  should be sufficiently small to allow more higher accuracy measurement. In particular, we also select the same wavelength combination as the three high-frequency modulation wavelength of the generated CFP (refer to Section 2.1).

$$\lambda_n = \lambda_{\phi_n}, \quad (12)$$

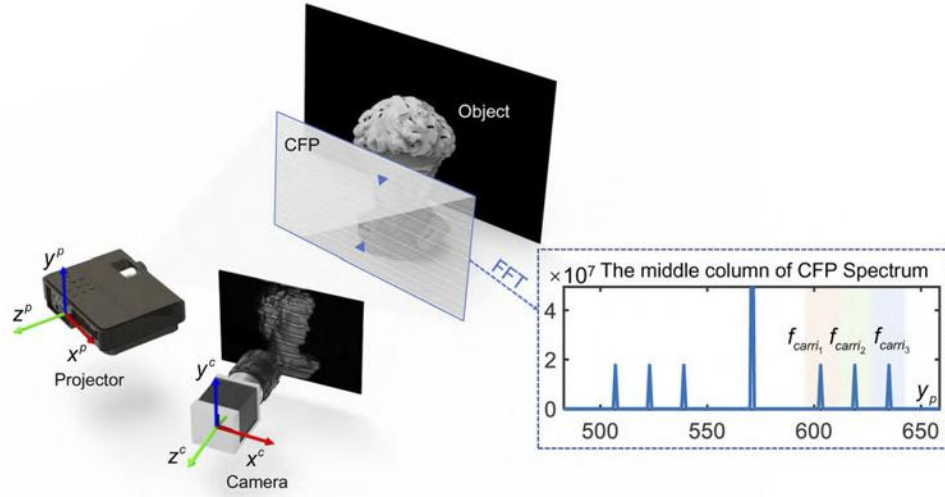
After examining that the pairs of wrapped phase values are unique, the fringe orders  $k_1, k_2$ , and  $k_3$  of the three phase maps can be determined, then we can acquire the high-accuracy absolute phase as part of the high-quality network training datasets.

### 2.3. One-to-three single-shot phase retrieval network

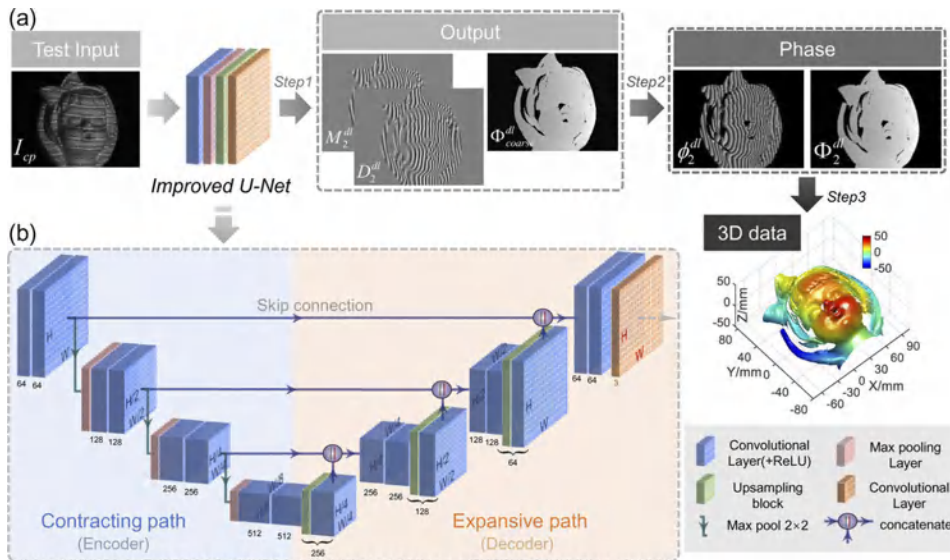
The crucial step of FPP is to retrieve high-precision and unambiguous phase distribution. Ideally, a monocular FPP system (as shown in Fig. 4) can use only one dense fringe image to robustly achieve high-quality phase unwrapping and absolute 3D reconstruction for complex scenes. However, limited by the number of fringe projection patterns required for 3D imaging, the current traditional FPP methods are still unable to robustly complete high-quality phase recovery under the premise of one projection. To this end, inspired by the recent successful applications of deep learning techniques on FPP, we combine a deep convolutional neural network with a single composite fringe image to develop a one-to-three single-shot phase retrieval network. The flow chart of the proposed method is shown in Fig. 5, which mainly includes: data preprocessing and network model construction, phase analysis and phase recovery based on deep convolutional neural network, phase-to-height mapping.

After training and testing different networks, such as ResNet [59], U-Net [60], and U-Net derivative networks (such as MultiResUNet [61], etc.), we finally choose the U-Net network that takes into account versatility and practicability for model training, and make the following fine-tuning of the network based on the prototype structure of the U-Net network: (1) In order to prevent overfitting caused by the bigger network, the designed U-Net network is reduced by one layer, changed from 5 layers to 4 layers, and we use Dropout which is one of the most effective and most commonly used regularization techniques for neural networks to further fight overfitting. (2) We set the network as a one-to-three convolutional neural network with single input and three outputs, the input channel is a composite fringe image designed in Section 2.1, and the three output channels are the numerator term and the denominator term of the wrapped phase arctangent function, and the coarse absolute phase term.

The improved U-Net network architecture is illustrated in Fig. 5(b). It consists of a contracting path (left side) and an expansive path (right side) [60]. The contracting path (also called Encoder)



**Fig. 4.** Hardware system and the middle column distribution of the frequency spectrum generated by the Fourier transform of the designed composite fringe pattern (CFP).



**Fig. 5.** Flowchart of our proposed approach. (a) Input the test data, output the numerator  $M_{dl}$ , denominator  $D_{dl}$  and the low-accuracy absolute phase  $\Phi_{\text{coarse}}$  through the trained network model, then obtain the high-accuracy absolute phase by post-processing and reconstruct the 3D information by the calibration parameters. (b) The improved U-Net network architecture.

follows a typical convolutional network architecture, including two convolutional layers (“SAME” padding) that are repeatedly applied, and each convolution is followed by a rectified linear unit (ReLU) and a  $2 \times 2$  max pooling layer. In each convolution layer of the network, the size of the convolution kernel is  $3 \times 3$ , which is used for feature extraction, and the padding is set to “same” to ensure that the size of the feature map ( $H, W$ ) remains the same after each convolution. After the max pooling layer, the size of the feature map will be downsampled with a stride of 2, the size of the feature map will become  $(H/2, W/2)$ , and the number of feature channels will be doubled. It should be noted that the linear rectification unit (ReLU) in each convolutional layer is one of the important factors to ensure that the deep learning network can be trained, and its operation is as follows:

$$r(\xi) = \max(0, \xi) = \begin{cases} 0, & \text{if } \xi \leq 0 \\ \xi, & \text{otherwise} \end{cases}, \quad (13)$$

where  $\xi$  represents an independent variable. The above process needs to be performed 4 times. Especially in the last execution, max pooling is no longer performed, but the feature map is directly sent to the expansion path.

Each step in the expansion path (also called Decoder) includes an upsampling layer followed by a  $2 \times 2$  convolutional layer, a concatenate, and two  $3 \times 3$  convolutional layers followed by a ReLU. The upsampling layer used for feature mapping halves the number of feature channels and doubles the size of the feature map. A concatenate merges the upper layer and the corresponding feature map from the contracting path through skip connection to retain more dimensional/location information. This critical step will facilitate the subsequent layers to freely choose between shallow and deep features, which is more advantageous for the semantic segmentation task of deep neural networks. At the final layer, a  $1 \times 1$  convolutional layer (“SAME” padding) is used to map the required three 3D tensors. The improved U-Net network has a total of 18 convolutional layers.

Next, we will discuss the specific procedures of our algorithm.

**Step 1:** In order to retrieve high-quality wrapped phase and absolute phase information, we input the three phase-wavelength and three carrier-frequency composite fringe images captured by the camera different test scenarios into the trained improved U-Net network, where the three short wavelengths of the composite fringe pattern are  $\lambda_{\phi_1} = 9$ ,  $\lambda_{\phi_2} = 11$ ,  $\lambda_{\phi_3} = 13$  (satisfying Eq. (11)), and the three carrier frequencies are set to  $f_{carr_1} = 32$ ,  $f_{carr_2} = 48$ ,  $f_{carr_3} = 64$ , respectively. From the perspective of the robustness and accuracy of phase recovery in our algorithm, we finally chose the wrapped phase numerator term  $M_2$ , denominator term  $D_2$ , and absolute phase  $\Phi_2$  corresponding to the second wavelength  $\lambda_{\phi_2}$  as the network label to train our network. Considering the physical model of the traditional phase-shifting algorithm, we choose to predict the numerator and denominator terms instead of directly predicting the wrapped phases. Compared with the network structure that directly connects the fringe pattern to the phase, this strategy bypasses the difficulty of the wrapped phase with the  $2\pi$  phase truncation and effectively removes the influence of the surface reflectivity variations, so as to achieve higher quality phase analysis to predict the high-quality wrapped phase. Inspired by the traditional composite fringe projection profilometry described in Section 2.1, three coprime short-wavelength fringes that can achieve unambiguous phase retrieval in the time domain are combined into one pattern through three carrier frequency. Compared with a single-frequency fringe pattern as the network input, the three-phase-wavelength and three-carrier-frequency composite fringe pattern ensures the unambiguity of the network input, and at the same time ensures that the phase can be unambiguously unwrapped during the absolute phase retrieval process.

**Step 2:** After predicting the numerator  $M_2^{dl}$ , denominator  $D_2^{dl}$ , and coarse absolute phase  $\Phi_{2, coarse}^{dl}$  of the composite fringe image through the trained improved U-Net network, the high-quality

wrapped phase map of the second wavelength  $\phi_2^{dl}$  can be calculated:

$$\phi_2^{dl} = \arctan \frac{M_2^{dl}}{D_2^{dl}}. \quad (14)$$

Then, high-quality absolute phase  $\Phi_2^{dl}$  can be obtained:

$$\Phi_2^{dl} = \phi_2^{dl} + 2\pi \cdot \text{round}[(\Phi_{coarse}^{dl} - \phi_2^{dl}) / (2\pi)], \quad (15)$$

where *round* represents the rounding function. Although the existence of the objects' surface reflectivity  $\alpha$  makes the deep learning training model can only predict "coarse" absolute phase with low-precision, its accuracy is sufficient to provide the correct fringe order of the high-quality wrapped phase. The final high-precision absolute phase can be obtained through the high-quality wrapped phase and the correct fringe order.

**Step 3:** After acquiring the high-accuracy absolute phase, the 3D information of the objects can be reconstructed by utilizing the phase-to-height mapping relationship and the calibration parameters of the FPP system [62]. The relation between the phase and the height coordinates can be written as

$$\begin{cases} x_p = \Phi_2^{dl} W / (2\pi N_{\lambda_2}) \\ Z_w = M_z + N_z / (C x_p + 1) \end{cases}, \quad (16)$$

where  $x_p$  is the projector  $x$  coordinate,  $W$  is the horizontal resolution of the projection pattern,  $N_{\lambda_2}$  is the fringe density,  $M_z$  and  $N_z$ , and  $C$  are the constants derived from calibration parameters.

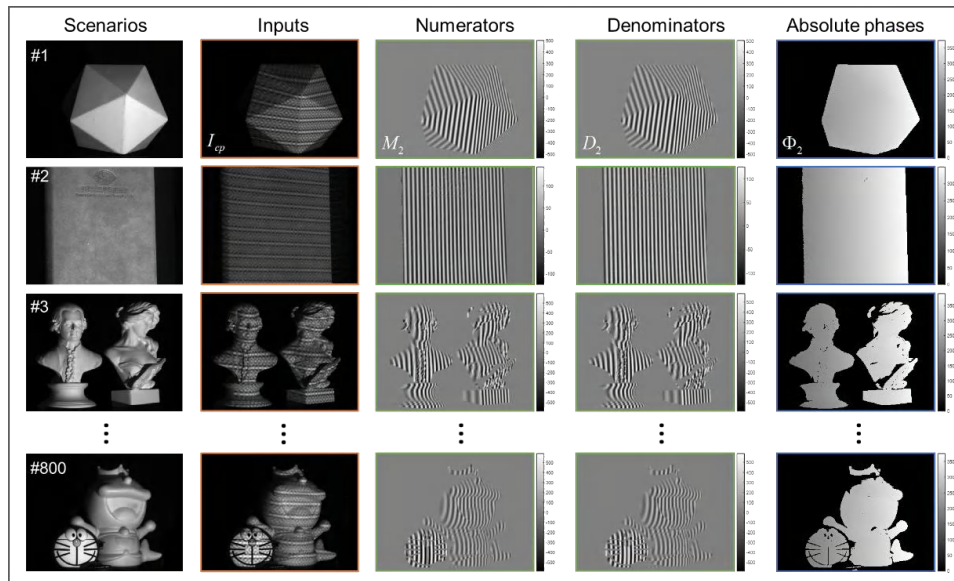
### 3. Experiments

To verify the performance of our method, we construct a monocular fringe projection system (Fig. 4), which consists of a digital light processing (DLP) projector (Texas Instruments DLP LightCrafter 4500) with an WXGA resolution ( $912 \times 1140$ ) DMD and an industrial camera (Basler ace acA640-750  $\mu\text{m}$ ) with  $640 \times 480$  resolution. The camera with the ON Semiconductor PYTHON 300 CMOS sensor delivers 751 fps Frame Rate at VGA resolution. Under the condition of satisfying the above Eq. (11) and Eq. (12), we select  $\{9, 11, 13\}$  wavelength combinations to provide unambiguous phase unwrapping for the whole projection range ( $LCM(9, 11, 13) = 1287 > 912$ ). The field of view (FOV) of the measurement system is about  $210 \text{ mm} \times 160 \text{ mm}$ , and the distance between the camera and the region of interest is 400 mm approximately.

In the supervised training mode, the unambiguous inputs and the corresponding accurately known outputs are required. Figure 6 shows some typical shooting scenes of the training datasets. As mentioned above, a set of input and output network training data includes composite fringe images  $I_{cp}^c$ , as well as the numerators  $M_2$ , denominators  $D_2$  and the absolute phases  $\Phi_2$ , where  $M_2$  and  $D_2$  are calculated by the 12-step PS method, and  $\Phi_2$  is obtained by the three-frequency TPU with PDM method. We collect 1000 sets of data for different scenarios including simple, complex, and isolated objects, and divide them into training sets, validation sets and test sets at a ratio of 8:1:1. The training data sets are used to determine the network weight (Fig. 6); the validation data sets are used to determine when to stop training; after training, the performance is evaluated by test data set that have never been trained.

The constructed neural network is computed on a desktop with Intel Core i7-7800X CPU and a GeForce GTX 1080 Ti GPU (NVIDIA) under the Python deep learning framework Keras with the TensorFlow platform (Google). The optimizer chooses the Adam optimization scheme, which is used to update the network weights with the loss value and achieve better gradient propagation, and its default initial learning rate  $lr$  is set to 0.0001. Batch Normalization is adopted mini-batch





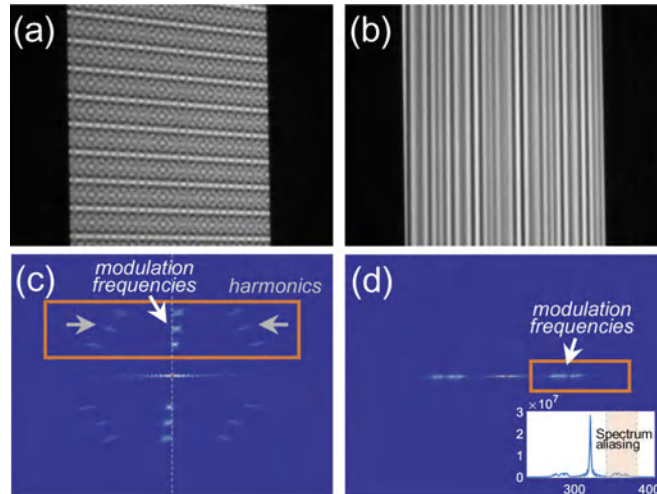
**Fig. 6.** Part of training datasets. Each row shows the network training scenario and labels. The training dataset includes the input image  $I_{cp}$  and three output terms of the second (modulated) frequency fringe: numerator  $M_2$ , denominator  $D_2$ , and absolute phase  $\Phi_2$ . The composite fringe images as input are captured in different scenes, including objects with simple or complex surfaces, continuous surfaces or isolated objects, and objects with different materials. These scenes shown in turn are an icosahedral triangle, a customised notebook with “SCILab” logo, Beethoven and a harp girl, as well as a plastic toy and earphone case. The ground-truth data is calculated by 12-step PS and three-frequency TPU with PDM method.

gradient descent scheme (mini-batch size = 2). The loss function we select in this neural network is mean squared error (MSE), which compares the predicted value with the target value after each batch in each epoch and generates a loss value. At the same time, the root mean squared error (RMSE) is calculated after each epoch to help visually monitor the training process. After 200 epochs were trained on the NVIDIA graphics card, the training loss and validation loss of the network converged. Moreover, due to data enhancement (background removal and normalization of input images) and the improvement of the network structure, the entire training time of our network only takes about 3 hours. We can directly put the captured and processed composite fringe image into the trained network model to retrieve the absolute phase map of target object and complete the offline 3D measurement. The network model prediction speed of our approach is about 15 fps.

### 3.1. Qualitative evaluation

Through “learning” from a large number of data sets, the properly trained neural network can “de-multiplex” high-resolution, spectrum-crosstalk-free phases from the multiplexing composite fringe and directly reconstruct a high-accuracy absolute phase map for single-shot, unambiguous 3D surface imaging. We conducted static and dynamic experiments in several different scenarios to test the trained deep convolutional neural network and verify the superiority of the proposed method over traditional methods.

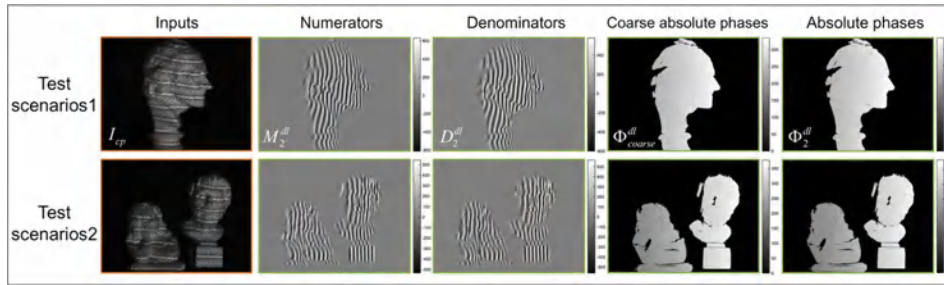
After analyzing different coding schemes of single-shot structured light illumination (see Sec. 1 for detailed analysis), we designed a three-carrier three-frequency composite fringe pattern. Here, we respectively projected the designed composite fringe pattern and the direct composite three-frequency fringe pattern to the scenes. Figure 7 shows the comparison between the designed composite fringe image (Fig. 7(a)) and the direct composite three-frequency fringe image (Fig. 7(b)), and the spectrum distribution of these two kinds of frequency-multiplexing-coded images are shown in Fig. 7(c) and (d), from which we can see that: (1) the directly composite image has serious spectrum aliasing, while our designed composite pattern can separate three close high-frequency information through three carrier frequencies; (2) although the designed fringe pattern avoids spectrum aliasing to some extent, its spectrum is easily affected by the system parameters between the projector and the camera, which results in a slight unknown variation of the three carrier frequencies  $f_{carrier_n}$  in the orthogonal direction. Therefore, it is difficult to demodulate the high-precision phase information through the traditional FT method that uniformly filters three high-frequency channels of the captured composite image by the band-pass filters at the center of  $f_{carrier_n}$ . Our deep learning-based method will solve these obstacles at once through a trained convolutional neural network.



**Fig. 7.** Comparison of two kinds of frequency-multiplexing-coded schemes. (a) Image of a flat plate obtained by projecting the designed three short-wavelength superimposed three carrier-frequency composite fringe image. (b) Image of a flat plate obtained by projecting the direct composite three-frequency fringe image. (c) Spectrum distribution of (a). (d) Spectrum distribution of (b).

To test the performance of the trained neural network, we measured two static scenarios that include single and multiple isolated objects with different surface roughness. The captured row input composite fringe images  $I_{cp}^p(x, y)$  are shown in the first columns of Fig. 8. Note that our neural network has never seen these scenarios during the training phase. After preprocessing these captured composite fringe images, we directly input them into the trained neural network to predict the numerators  $M_2^{dl}$ , denominators  $D_2^{dl}$  and coarse absolute phase  $\Phi_{coarse}^{dl}$  of the input fringe images. The results are shown in the second to fourth columns of Fig. 8, where the estimated numerator and denominator are fed into Eq. (14) to obtain the wrapped phase map, and then the unwrapped phase  $\Phi_2^{dl}$  distribution shown in the fifth column can be retrieved from the calculated wrapped phase and the estimated coarse absolute phase according to Eq. (15). As we can see, the phase ambiguity has been completely eliminated. Furthermore, through the pre-calibrated

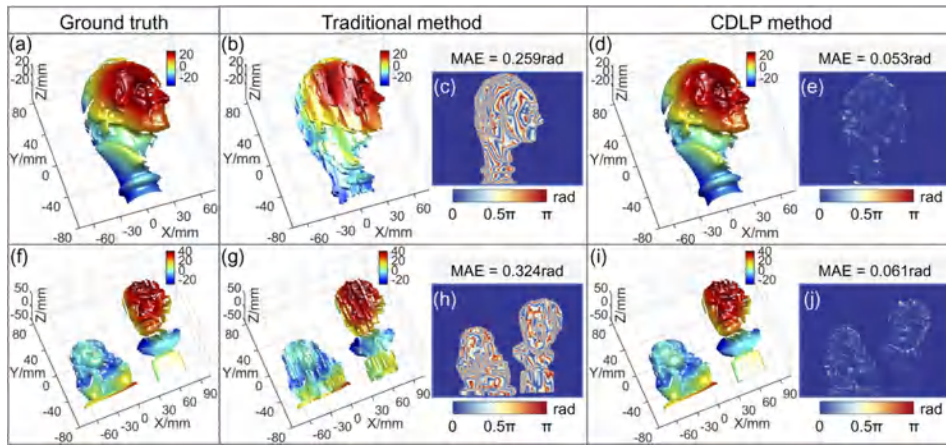
parameters of the camera-projector FPP system and the phase-height mapping (Eq. (16)), we converted the unwrapped phase maps into 3D rendered geometries. In Fig. 9, we compared 3D reconstruction results for the traditional FT method using Guan's coding scheme [36], and our learning-based frequency multiplexing-coded method to ground truth. And to quantitatively analyze the phase quality, Figs. 9(c), (e), (h), and (j) show the corresponding unwrapped phase error maps of the entire measurement area. In the investigation, the phases calculated by the 12-step phase-shifting and three-frequency temporal phase unwrapping with projection distance minimization are serve as ground truth phase maps. Due to the influence of severe spectrum aliasing and frequency shift transform, the phase error of the traditional phase retrieval method using Guan's coding scheme is more obvious than our proposed CDLP method. Specifically, to further quantify this trend, we report the mean absolute error (MAE) of unwrapped phase in Fig. 9. Compared with the tradition FT method, our proposed method reduces the projection mode from three to one without losing the accuracy of phase recovery, improving the time resolution without changing the spatial resolution. Compared with the traditional method with Guan's coding scheme, the proposed method improves the phase recovery accuracy by nearly an order of magnitude.



**Fig. 8.** The prediction results of the two static test scenarios. Each row shows the input composite image  $I_{cp}^p(x, y)$ , the estimated results of numerator  $M_2^{dl}$ , denominator  $D_2^{dl}$ , coarse absolute phase  $\Phi_{coarse}^{dl}$ , and the final absolute/unwrapped phase  $\Phi_2^{dl}$ .

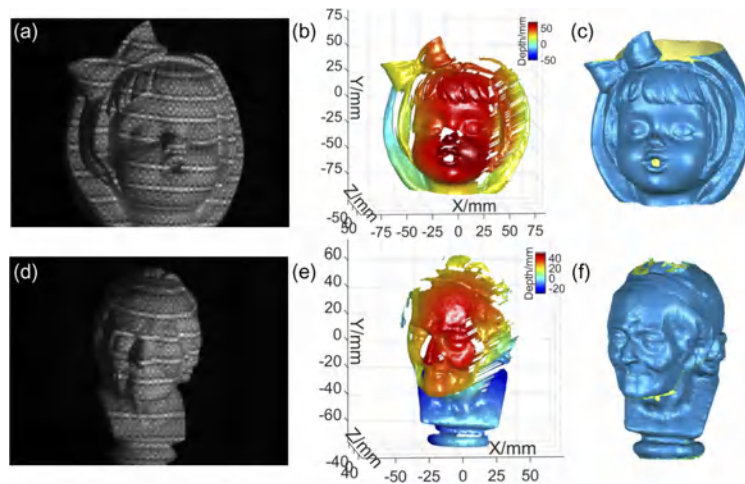
Although Guan's method avoided a significant degree of spectrum loss, but the slight frequency shifts error brought considerable loss to the 3D reconstruction. From the results of our method, it can be seen that the deep learning-based frequency-multiplexing-coded method obtains a higher quality 3D reconstruction, which is almost comparable to the reference 3D model reconstructed by 12-step PS method (the Ground truth). Moreover, our method only needs one composite fringe image to reconstruct absolute 3D information, and the reconstruction efficiency is 36 times higher than that of the reference method. This experiment verified that the deep learning-based frequency-multiplexing-coded method can not only effectively overcome the adverse effects, such as spectrum aliasing, spectrum leakage, and channel crosstalk, but also achieve high-precision absolute phase retrieval and high-quality absolute 3D surface reconstruction from a single-frame fringe image.

In the second experiment, we measured two sets of moving objects. a rotating bow girl model and a moving Voltaire plaster model, to verify the ability of our method in dynamic scenes. Figures 10(a) and (d) respectively show the raw image of a certain frame in the two captured videos, Figs. 10(b) and (e) are the corresponding 3D reconstruction results using our method in the selected moments, and Figs. 10(c) and (f) further show the 360-degree point cloud registration results. During the measurement, a single-frame composite fringe pattern was continuously projected on the surface of the object, meanwhile, a monochrome camera simultaneously captured the gray fringe image of each frame. We can see that our method is fundamentally immune to phase-shifting errors induced by object motion thanks to its single-shot nature. Consequently, it



**Fig. 9.** 3D reconstruction results of Ground truth, traditional method, and our proposed CDLP method in two measurement scenes. (a), (f) 3D reconstruction result of the Ground truth (12-step PS with number-theoretical method). (b) (c), (g) (h) 3D reconstruction result and its corresponding absolute phase error map of traditional method (FT method with Guan's coding scheme). (d) (e), (i) (j) 3D reconstruction result and its corresponding absolute phase error map of CDLP method (composite fringe projection deep learning profilometry).

is suitable for dynamic 3D imaging of rapidly moving objects. The whole measurement process of the rotating statues are shown in Fig. 10 (Multimedia views).



**Fig. 10.** The dynamic 3D measurement results of a rotating bow girl model and a moving Voltaire plaster model. (a), (d) The captured composite fringe images at two different moments. (b), (e) The corresponding 3D results reconstructed by our method. (c), (f) Registration results. (Multimedia views: see [Visualization 1](#) and [Visualization 2](#) for the whole measurement process of these two scenes)

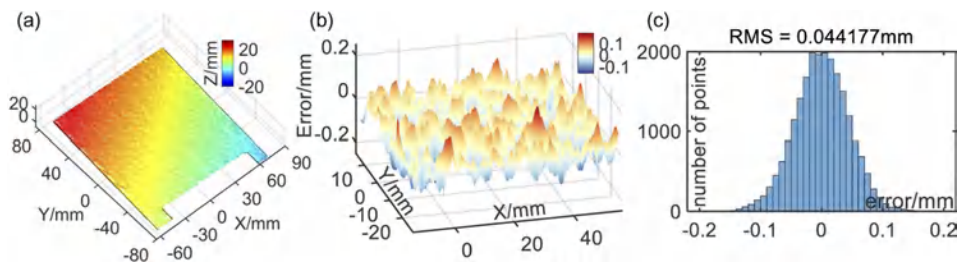


### 3.2. Quantitative evaluation

Last but not least, we respectively measured a standard ceramic plate (Fig. 11) and a pair of standard ceramic spheres (Fig. 12) to quantitatively evaluate the 3D reconstruction precision of the proposed method. It is noted that the structural parameters of the standard ceramic balls have been calibrated by the coordinate measuring machine, and their radii are  $RA = 25.3999$  mm and  $RB = 25.3983$  mm, respectively. The center-to-center distance of the standard ceramic balls is  $D = 100.1563$  mm with an uncertainty of  $1.0 \mu\text{m}$ . We produced the measurement results of the plate and two spheres and performed plane and spherical fitting on the measurement results. Their errors are shown in Figs. 11(b), (c), Figs. 12(c1), (c2), and Figs. 12(d1), (d2). The radii of reconstructed spheres are  $RA_{dl} = 25.5246$  mm and  $RB_{dl} = 25.2901$  mm, with the mean absolute error (MAE) of  $0.0531$  mm and  $0.0506$  mm. The measured center distance is  $D_{dl} = 100.2027$  mm with the deviation of  $\Delta d = 0.0464$  mm. Additionally, the root mean square (RMS) error of sphere A and sphere B are respectively  $0.066$  mm and  $0.062$  mm, as shown in Figs. 12(c2) and (d2). This experiment proves that our method can provide high-quality 3D measurements using only a single fringe image.

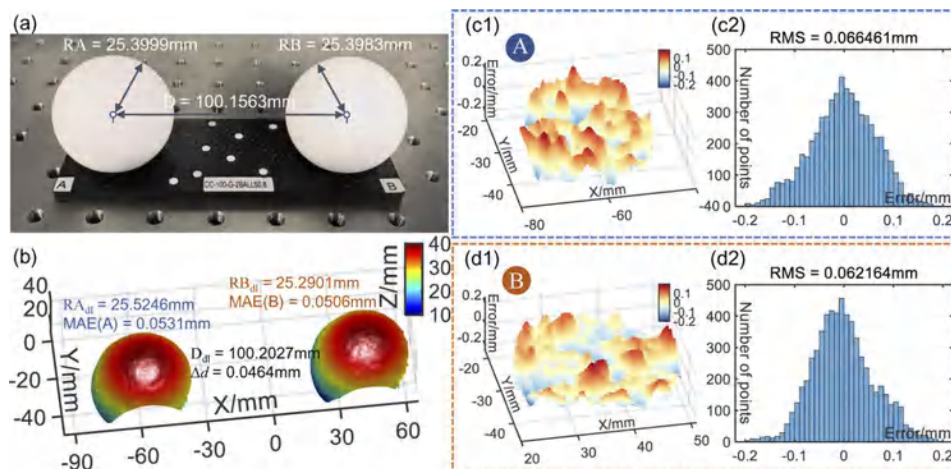
## 4. Conclusions

In this work, we have proposed a deep learning-based single-shot composite fringe projection profilometry (CFPP), which combines a deep learning technology with a specially designed spatial frequency multiplexing coding strategy to achieve single-frame, high-precision, unambiguous 3D shape reconstruction. According to experimental results, this deep learning-based method can perform high-quality 3D shape measurements on discontinuous and/or mutually isolated objects in fast motion. Compared with the existing high-speed 3D imaging method based on multi-frame images, our method is fundamentally immune to motion errors. Compared with the traditional FT and frequency multiplexing FT methods, our approach can effectively overcome the adverse effects, such as spectrum aliasing, spectrum leakage, and channel crosstalk. Using only a single composite fringe pattern, the 3D imaging quality of the proposed method is comparable to the performance of the traditional 12-step PS method. Besides, the trained network model can be fully automatic to achieve high-quality 3D measurement without tuning parameters.



**Fig. 11.** Precision analysis of standard ceramic plate. (a) 3D reconstruction results by our method. (b) Error distribution. (c) RMS error.

Deep learning technology has thoroughly “permeated” into almost all tasks of optical metrology and has delivered some pretty impressive results. This paper intends to point out that with its powerful learning capabilities, deep learning technology can break the limitations of various influencing factors in traditional single-frame 3D imaging algorithms and achieve impressive results for single-shot, instantaneous absolute 3D shape measurement of discontinuous and/or isolated objects. However, the underlying reasons behind these successes of deep learning prediction remain unclear at this stage. Many researchers are still skeptical and maintain a wait-and-see attitude towards its applications in high-risk scenarios, such as industrial inspection and



**Fig. 12.** Precision analysis of a pair of standard ceramic spheres. (a), (b) 3D reconstruction results by our method. (c1), (c2) The error distribution and corresponding RMS error of sphere A. (d1), (d2) The error distribution and corresponding RMS error of sphere B.

medical care. But it can be envisaged that with the further development of artificial intelligence technology, the continuous improvement of computer hardware performance, and the further development of optical information processing techniques, these challenges will gradually be solved in the near future. Deep learning will thus play a more significant role and make a more far-reaching impact in optics and photonics.

**Funding.** National Natural Science Foundation of China (62075096); Leading Technology of Jiangsu Basic Research Plan (BK20192003); Jiangsu Provincial “One belt and one road” innovation cooperation project (BZ2020007); Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX21\_0273); Fundamental Research Funds for the Central Universities (30919011222, 30920032101, 30921011208).

**Disclosures.** The authors declare no conflicts of interest.

**Data availability.** Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

## References

1. J. Geng, “Structured-light 3d surface imaging: a tutorial,” *Adv. Opt. Photonics* **3**(2), 128–160 (2011).
2. S. Feng, C. Zuo, T. Tao, Y. Hu, M. Zhang, Q. Chen, and G. Gu, “Robust dynamic 3-d measurements with motion-compensated phase-shifting profilometry,” *Opt. Lasers Eng.* **103**, 127–138 (2018).
3. B. Pan, H. Xie, Z. Wang, K. Qian, and Z. Wang, “Study on subset size selection in digital image correlation for speckle patterns,” *Opt. Express* **16**(10), 7037–7048 (2008).
4. Y. Hu, Q. Chen, S. Feng, and C. Zuo, “Microscopic fringe projection profilometry: A review,” *Opt. Lasers Eng.* **135**, 106192 (2020).
5. S. S. Gorthi and P. Rastogi, “Fringe projection techniques: whither we are?” *Opt. Lasers Eng.* **48**(2), 133–140 (2010).
6. X. Su and Q. Zhang, “Dynamic 3-d shape measurement method: a review,” *Opt. Lasers Eng.* **48**(2), 191–204 (2010).
7. S. Feng, L. Zhang, C. Zuo, T. Tao, Q. Chen, and G. Gu, “High dynamic range 3d measurements with fringe projection profilometry: a review,” *Meas. Sci. Technol.* **29**(12), 122001 (2018).
8. J. Qian, S. Feng, T. Tao, Y. Hu, K. Liu, S. Wu, Q. Chen, and C. Zuo, “High-resolution real-time 360 3d model reconstruction of a handheld object with fringe projection profilometry,” *Opt. Lett.* **44**(23), 5751–5754 (2019).
9. J. Salvi, S. Fernandez, T. Pribanic, and X. Llado, “A state of the art in structured light patterns for surface profilometry,” *Pattern Recognit.* **43**(8), 2666–2680 (2010).
10. L. Gao, J. Liang, C. Li, and L. V. Wang, “Single-shot compressed ultrafast photography at one hundred billion frames per second,” *Nature* **516**(7529), 74–77 (2014).
11. S. Heist, C. Zhang, K. Reichwald, P. Kühmstedt, G. Notni, and A. Tünnermann, “5d hyperspectral imaging: fast and accurate measurement of surface shape and spectral characteristics using structured light,” *Opt. Express* **26**(18), 23366–23379 (2018).

12. D. Qi, S. Zhang, C. Yang, Y. He, F. Cao, J. Yao, P. Ding, L. Gao, T. Jia, J. Liang, Z. Sun, and L. V. Wang, "Single-shot compressed ultrafast photography: a review," *Adv. Photonics* **2**(1), 014003 (2020).
13. J. Qian, S. Feng, M. Xu, T. Tao, Y. Shang, Q. Chen, and C. Zuo, "High-resolution real-time 360° 3d surface defect inspection with fringe projection profilometry," *Opt. Lasers Eng.* **137**, 106382 (2021).
14. S. Lei and S. Zhang, "Flexible 3-d shape measurement using projector defocusing," *Opt. Lett.* **34**(20), 3080–3082 (2009).
15. S. Zhang and P. S. Huang, "High-resolution, real-time three-dimensional shape measurement," *Opt. Eng.* **45**(12), 123601 (2006).
16. S. Heist, P. Lutzke, I. Schmidt, P. Dietrich, P. Kühmstedt, A. Tünnemann, and G. Notni, "High-speed three-dimensional shape measurement using gobo projection," *Opt. Lasers Eng.* **87**, 90–96 (2016).
17. C. Zuo, T. Tao, S. Feng, L. Huang, A. Asundi, and Q. Chen, "Micro fourier transform profilometry (uftp): 3d shape measurement at 10,000 frames per second," *Opt. Lasers Eng.* **102**, 70–91 (2018).
18. K. Liu, Y. Wang, D. L. Lau, Q. Hao, and L. G. Hassebrook, "Dual-frequency pattern scheme for high-speed 3-d shape measurement," *Opt. Express* **18**(5), 5229–5244 (2010).
19. C. Zuo, Q. Chen, G. Gu, S. Feng, F. Feng, R. Li, and G. Shen, "High-speed three-dimensional shape measurement for dynamic scenes using bi-frequency tripolar pulse-width-modulation fringe projection," *Opt. Lasers Eng.* **51**(8), 953–960 (2013).
20. C. Zuo, Q. Chen, G. Gu, S. Feng, and F. Feng, "High-speed three-dimensional profilometry for multiple objects with complex shapes," *Opt. Express* **20**(17), 19493–19510 (2012).
21. Y. Zhang, Z. Xiong, and F. Wu, "Unambiguous 3d measurement from speckle-embedded fringe," *Applied optics* **52**(32), 7797–7805 (2013).
22. S. Feng, Q. Chen, and C. Zuo, "Graphics processing unit-assisted real-time three-dimensional measurement using speckle-embedded fringe," *Appl. Opt.* **54**(22), 6865–6873 (2015).
23. T. Tao, Q. Chen, J. Da, S. Feng, Y. Hu, and C. Zuo, "Real-time 3-d shape measurement with composite phase-shifting fringes and multi-view system," *Opt. Express* **24**(18), 20253–20269 (2016).
24. S. Heist, P. Kuehmstedt, A. Tuennermann, and G. Notni, "Theoretical considerations on aperiodic sinusoidal fringes in comparison to phase-shifted sinusoidal fringes for high-speed three-dimensional shape measurement," *Appl. Opt.* **54**(35), 10541–10551 (2015).
25. M. Takeda, H. Ina, and S. Kobayashi, "Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry," *J. Opt. Soc. Am.* **72**(1), 156–160 (1982).
26. M. Takeda and K. Mutoh, "Fourier transform profilometry for the automatic measurement of 3-d object shapes," *Appl. Opt.* **22**(24), 3977–3982 (1983).
27. Q. Kemao, "Two-dimensional windowed fourier transform for fringe pattern analysis: principles, applications and implementations," *Opt. Lasers Eng.* **45**(2), 304–317 (2007).
28. Q. Kemao, "Windowed fourier transform for fringe pattern analysis," *Appl. Opt.* **43**(13), 2695–2702 (2004).
29. L. Huang, Q. Kemao, B. Pan, and A. K. Asundi, "Comparison of fourier transform, windowed fourier transform, and wavelet transform methods for phase extraction from a single fringe pattern in fringe projection profilometry," *Opt. Lasers Eng.* **48**(2), 141–148 (2010).
30. Z. Zhang, Z. Jing, Z. Wang, and D. Kuang, "Comparison of fourier transform, windowed fourier transform, and wavelet transform methods for phase calculation at discontinuities in fringe projection profilometry," *Opt. Lasers Eng.* **50**(8), 1152–1160 (2012).
31. J. Zhong and J. Weng, "Spatial carrier-fringe pattern analysis by means of wavelet transform: wavelet transform profilometry," *Appl. Opt.* **43**(26), 4993–4998 (2004).
32. X. Su, W. Chen, Q. Zhang, and Y. Chao, "Dynamic 3-d shape measurement method based on ftp," *Opt. Lasers Eng.* **36**(1), 49–64 (2001).
33. X. Su and W. Chen, "Fourier transform profilometry: a review," *Opt. Lasers Eng.* **35**(5), 263–284 (2001).
34. C. Zuo, L. Huang, M. Zhang, Q. Chen, and A. Asundi, "Temporal phase unwrapping algorithms for fringe projection profilometry: A comparative review," *Opt. Lasers Eng.* **85**, 84–103 (2016).
35. M. Takeda, Q. Gu, M. Kinoshita, H. Takai, and Y. Takahashi, "Frequency-multiplex fourier-transform profilometry: a single-shot three-dimensional shape measurement of objects with large height discontinuities and/or surface isolations," *Appl. Opt.* **36**(22), 5347–5354 (1997).
36. C. Guan, L. Hassebrook, and D. Lau, "Composite structured light pattern for three-dimensional video," *Opt. Express* **11**(5), 406–417 (2003).
37. C. Zuo, S. Feng, L. Huang, T. Tao, W. Yin, and Q. Chen, "Phase shifting algorithms for fringe projection profilometry: A review," *Opt. Lasers Eng.* **109**, 23–59 (2018).
38. H.-M. Yue, X.-Y. Su, and Y.-Z. Liu, "Fourier transform profilometry based on composite structured light pattern," *Opt. Laser Technol.* **39**(6), 1170–1175 (2007).
39. M. Lu, X. Su, Y. Cao, Z. You, and M. Zhong, "Modulation measuring profilometry with cross grating projection and single shot for dynamic 3d shape measurement," *Opt. Lasers Eng.* **87**, 103–110 (2016).
40. J. Pages, J. Salvi, C. Collewet, and J. Forest, "Optimised de bruijn patterns for one-shot shape acquisition," *Image Vis. Comput.* **23**(8), 707–720 (2005).
41. Z. Zhang, D. P. Towers, and C. E. Towers, "Snapshot color fringe projection for absolute three-dimensional metrology of video sequences," *Appl. Opt.* **49**(31), 5947–5953 (2010).

42. G. Sansoni and E. Redaelli, "A 3d vision system based on one-shot projection and phase demodulation for fast profilometry," *Meas. Sci. Technol.* **16**(5), 1109 (2005).
43. G. Barbastathis, A. Ozcan, and G. Situ, "On the use of deep learning for computational imaging," *Optica* **6**(8), 921–943 (2019).
44. S. Feng, C. Zuo, L. Zhang, W. Yin, and Q. Chen, "Generalized framework for non-sinusoidal fringe analysis using deep learning," *Photonics Res.* **9**(6), 1084–1098 (2021).
45. S. Feng, Q. Chen, G. Gu, T. Tao, L. Zhang, Y. Hu, W. Yin, and C. Zuo, "Fringe pattern analysis using deep learning," *Adv. Photonics* **1**(2), 025001 (2019).
46. S. Feng, C. Zuo, Y. Hu, Y. Li, and Q. Chen, "Deep-learning-based fringe-pattern analysis with uncertainty estimation," *Optica* **8**(12), 1507–1510 (2021).
47. J. Shi, X. Zhu, H. Wang, L. Song, and Q. Guo, "Label enhanced and patch based deep learning for phase retrieval from single frame fringe pattern in fringe projection 3d measurement," *Opt. Express* **27**(20), 28929–28943 (2019).
48. S. Feng, C. Zuo, W. Yin, G. Gu, and Q. Chen, "Micro deep learning profilometry for high-speed 3d surface imaging," *Opt. Lasers Eng.* **121**, 416–427 (2019).
49. S. Van der Jeught and J. J. Dirckx, "Deep neural networks for single shot structured light profilometry," *Opt. Express* **27**(12), 17091–17101 (2019).
50. W. Yin, Q. Chen, S. Feng, T. Tao, L. Huang, M. Trusiak, A. Asundi, and C. Zuo, "Temporal phase unwrapping using deep learning," *Sci. Rep.* **9**(1), 20175 (2019).
51. J. Qian, S. Feng, T. Tao, Y. Hu, Y. Li, Q. Chen, and C. Zuo, "Deep-learning-enabled geometric constraints and phase unwrapping for single-shot absolute 3d shape measurement," *APL Photonics* **5**(4), 046105 (2020).
52. J. Qian, S. Feng, Y. Li, T. Tao, J. Han, Q. Chen, and C. Zuo, "Single-shot absolute 3d shape measurement with deep-learning-based color fringe projection profilometry," *Opt. Lett.* **45**(7), 1842–1845 (2020).
53. Z.-W. Li, Y.-S. Shi, C.-J. Wang, D.-H. Qin, and K. Huang, "Complex object 3d measurement based on phase-shifting and a neural network," *Opt. Commun.* **282**(14), 2699–2706 (2009).
54. H. Nguyen, Y. Wang, and Z. Wang, "Single-shot 3d shape reconstruction using structured light and deep convolutional neural networks," *Sensors* **20**(13), 3718 (2020).
55. Y. Zheng, S. Wang, Q. Li, and B. Li, "Fringe projection profilometry by conducting deep learning from its digital twin," *Opt. Express* **28**(24), 36568–36583 (2020).
56. H. Nguyen, N. Dunne, H. Li, Y. Wang, and Z. Wang, "Real-time 3d shape measurement using 3led projection and deep machine learning," *Appl. Opt.* **58**(26), 7100–7109 (2019).
57. T. Pribanić, S. Mrvoš, and J. Salvi, "Efficient multiple phase shift patterns for dense 3d acquisition in structured light scanning," *Image Vis. Comput.* **28**(8), 1255–1266 (2010).
58. H. Li, Y. Hu, T. Tao, S. Feng, M. Zhang, Y. Zhang, and C. Zuo, "Optimal wavelength selection strategy in temporal phase unwrapping with projection distance minimization," *Appl. Opt.* **57**(10), 2352–2360 (2018).
59. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), pp. 770–778.
60. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, (Springer, 2015), pp. 234–241.
61. N. Ibtchaz and M. S. Rahman, "Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation," *Neural Netw.* **121**, 74–87 (2020).
62. S. Feng, C. Zuo, L. Zhang, T. Tao, Y. Hu, W. Yin, J. Qian, and Q. Chen, "Calibration of fringe projection profilometry: A comparative review," *Opt. Lasers Eng.* **143**, 106622 (2021).





# PHOTONICS Research

## Neural-field-assisted transport-of-intensity phase microscopy: partially coherent quantitative phase imaging under unknown defocus distance

YANBO JIN,<sup>1,2,3,†</sup> LINPENG LU,<sup>1,2,3,†</sup> SHUN ZHOU,<sup>1,2,3</sup> JIE ZHOU,<sup>1,2,3</sup> YAO FAN,<sup>1,2,3</sup> AND CHAO ZUO<sup>1,2,3,\*</sup>

<sup>1</sup>Smart Computational Imaging Laboratory (SCILab), School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

<sup>2</sup>Smart Computational Imaging Research Institute (SCIRI) of Nanjing University of Science and Technology, Nanjing 210019, China

<sup>3</sup>Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing 210094, China

<sup>†</sup>These authors contributed equally to this work.

\*Corresponding author: zuochao@njjust.edu.cn

Received 19 February 2024; revised 28 April 2024; accepted 9 May 2024; posted 9 May 2024 (Doc. ID 521056); published 1 July 2024

The transport-of-intensity equation (TIE) enables quantitative phase imaging (QPI) under partially coherent illumination by measuring the through-focus intensities combined with a linearized inverse reconstruction algorithm. However, overcoming its sensitivity to imaging settings remains a challenging problem because of the difficulty in tuning the optical parameters of the imaging system accurately and because of the instability to long-time measurements. To address these limitations, we propose and experimentally validate a solution called neural-field-assisted transport-of-intensity phase microscopy (NFTPM) by introducing a tunable defocus parameter into neural field. Without weak object approximation, NFTPM incorporates the physical prior of partially coherent image formation to constrain the neural field and learns the continuous representation of phase object without the need for training. Simulation and experimental results of HeLa cells demonstrate that NFTPM can achieve accurate, partially coherent QPI under unknown defocus distances, providing new possibilities for extending applications in live cell biology. © 2024 Chinese Laser Press

<https://doi.org/10.1364/PRJ.521056>

### 1. INTRODUCTION

Quantitative phase imaging (QPI) has gained increased interest in optical microscopy research for its capability to quantify optical thickness and morphologies of unlabeled samples [1–3]. The QPI approach can be categorized into iterative and deterministic methods [4–6], where the deterministic method requires the establishment of an analytical expression for the object phase with respect to the measured intensity images. Given that the image formation process in QPI is inherently non-linear, linearization approaches are commonly invoked to facilitate solving for phase as a function of intensity measurements. For example, as a well-established deterministic phase retrieval approach, the transport-of-intensity equation (TIE) applies paraxial approximation and slowly-varied approximation to linearize the phase retrieval problem and can recover the quantitative phase by utilizing intensity images at multiple axially defocused planes [6,7]. Under partially coherent illumination, TIE is expected to achieve improved spatial resolution beyond the coherent diffraction limit [8]. Nevertheless, in a conventional microscope with circular illumination, partial coherence tends to diminish the phase contrast, resulting in compromised imaging resolution [9].

To achieve high-resolution and high-contrast QPI, the annular illumination (AI) matching objective numerical aperture (NA) has been employed in deconvolution-based TIE, referred to as AI-TIE [10]. AI-TIE strongly boosts the phase contrast and significantly improves the practical imaging resolution to a 2-fold objective NA. The strong phase contrast is ultimately transformed to the quantitative phase images by WOTF (weak object transfer function) inversion, yielding high-quality results with enhanced resolution. However, AI-TIE is usually limited to weak scattering samples since it linearizes the image formation model by invoking weak object approximation with ignoring higher-order terms in the complex transmittance of the sample. In addition, WOTF is a function directly related to the light source distribution, objective pupil function, and defocus distance. Once WOTF is determined, AI-TIE is not capable of adaptively adjusting optical parameters such as the defocus distance during the imaging process. Therefore, such TIE-based methods may result in degraded quality of phase retrieval due to the inaccurate inverse reconstruction for nonweak objects or cases where optical parameters are incorrect.

In contrast to the aforementioned physics-based approaches [6,10], data-driven deep learning methods can establish the nonlinear pseudo-inverse mapping relation between the defocused intensity and the object phase [11–14], bypassing the obstacle of “solving nonlinear ill-posed inverse problems.” Essentially, the major reason for the success of deep learning is the abundance of training data and the explicit agnosticism from *a priori* knowledge of how such data are generated [15]. However, high-quality paired data acquisition in experiments requires professional supervision and extensive labor. Furthermore, the lack of data diversity will restrict its generalization to out-of-domain cases with dissimilar optical parameters. Thus, the data-driven deep learning methods tend to fail in situations where it is difficult to obtain a large amount of high-quality paired data from a variety of different imaging systems.

To overcome the above limitations, researchers have developed untrained network approaches by incorporating physical priors into deep neural networks, such as the deep phase decoder [16] and PhysenNet [17]. These methods aim to achieve nonlinear optimization by minimizing the error between the prior model-generated image and the actual measurement. Their superiority lies in introducing neural networks as advanced regularization for automatic tuning. For instance, deep image prior (DIP) method can use a randomly initialized neural network as a prior to solve inverse problems such as pixel super-resolution [18]. Especially, the BlindNet method takes distance uncertainty into account and further addresses the phase retrieval problem with unknown defocus distance [19]. Additionally, we have witnessed the rise of neural field (NF), which has become a prominent self-supervised learning method [20]. NF can represent a three-dimensional (3D) scene as a continuous field, which is parameterized by a lightweight multilayer perceptron (MLP, i.e., fully connected network) and trained without ground truth data. In conjunction with computational imaging techniques, NF typically dispenses with training on a dataset and iterates the MLP network directly on the test data until the desired physical quantities are recovered, similar to physics-driven untrained network approaches. For example, NF can be incorporated into 3D diffraction tomography [21] or two-dimensional (2D) microscopy such as lensless microscopy [22] and Fourier ptychographic imaging [23]. However, these physics-driven deep learning methods only involve coherent imaging and are unsuitable for partially coherent imaging scenarios. In fact, considering partial coherence in phase retrieval helps to yield accurate results thanks to its better alignment with the actual situation [24]. Nevertheless, it needs to introduce additional parameters (such as coherence parameter) to establish a more complete forward model. Consequently, it remains a challenge to achieve stable partially coherent QPI under varying optical parameters.

In this work, we present a partially coherent QPI approach by using a neural field and taking the Abbe imaging model [25] as the physical prior. The proposed method, termed neural-field-assisted transport-of-intensity phase microscopy (NFTPM), is actually a gradient-based iterative algorithm. It drives a coordinate-based MLP through the physical prior to represent the phase distribution as a neural field and optimizes the MLP using the gradient computed by backpropagation in

both the physical model and the MLP model. This framework empowers NFTPM to concurrently adjust the defocus distance of the physical model by introducing a tunable defocus parameter, enabling stable QPI under unknown defocus distance. Moreover, NFTPM is applicable to non-weak phase objects, since the weak object approximation is not applied to the forward image formation. Instead of an image-to-image 2D CNN, NFTPM forms a point-to-point mapping function from spatial coordinates to phase values, which effectively constrains the solution space and renders single-shot QPI possible. Unlike untrained networks based on coherent imaging systems, NFTPM can adapt to various partially coherent illuminations, which is validated by simulations under circular illumination and annular illumination. Furthermore, based on a bright-field microscope equipped with annular NA-matched illumination [26,27] formed by sparsely distributed light-emitting diode (LED) elements, we realize stable QPI of unstained *Henrietta Lacks* (HeLa) cells, demonstrating that NFTPM is a valid approach for adaptive correction of defocus aberration during the long-term phase microscopy. Given the simplicity and effectiveness of the NFTPM method, it promises to advance the integration of partially coherent imaging with physics-driven deep learning and open new possibilities for robust non-interferometric QPI in dynamic optical environments.

## 2. METHODS

### A. Reconstruction Algorithm of NFTPM

The schematic diagram of NFTPM is outlined in Fig. 1(a), and Fig. 1(b) illustrates the image formation process in a partially coherent microscope, which corresponds to the physics prior used to drive NFTPM to perform phase recovery. The pipeline of NFTPM comprises a radial encoding module [21] and a 5-layer MLP ( $\mathbf{W}$  is the weights) that maps 2D spatial coordinate  $\mathbf{r} = (x, y)$  to phase value  $\phi(\mathbf{r})$ , which can finally represent the phase as a neural field  $\Phi_{\mathbf{W}}(\mathbf{r})$ . We adopt  $M \times N$  grid coordinates  $\mathbf{R} = \{(x_i, y_i)\}_{i=0}^{M \times N - 1}$  for the field of view (FOV) of interest (generally  $-1 \leq x \leq 1$ ,  $-1 \leq y \leq 1$ ), and the coordinates correspond to pixels on the image sensor. Initially, we utilize the radial encoding module to map densely distributed two-dimensional coordinate points to sparsely distributed high-dimensional space, thus allowing the MLP to better discriminate between different coordinate positions in order to characterize high-frequency information. For  $\mathbf{r} = (x, y) \in \mathbb{R}^{1 \times 2}$ , radial encoding can be expressed as

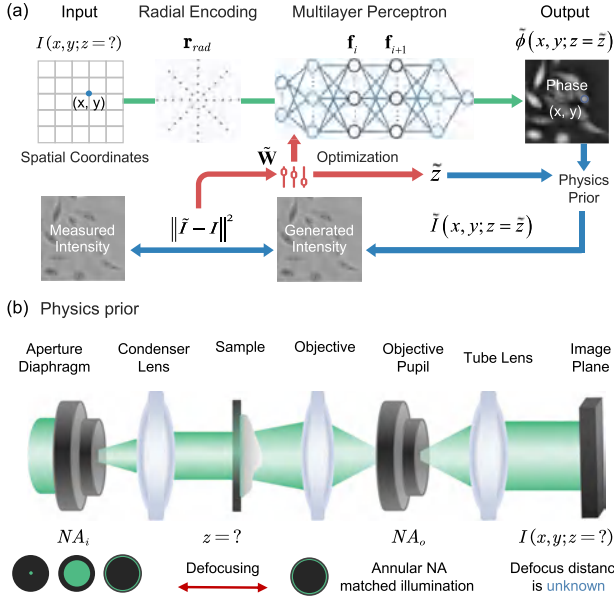
$$\mathbf{r}_{\text{rad}} = T_f \{\cos(\mathbf{T}_L \mathbf{r} \mathbf{T}_R), \sin(\mathbf{T}_L \mathbf{r} \mathbf{T}_R)\}, \quad (1)$$

where  $\mathbf{r}_{\text{rad}}$  is the encoded feature, and  $\mathbf{T}_L$  is the transformation matrix used for frequency expansion, which can be defined as

$$\mathbf{T}_L = [2^0 \pi, 2^1 \pi, \dots, 2^{L-1} \pi]^T, \quad (2)$$

where  $L$  is the number of the expanded frequencies. The purpose of introducing  $L$  frequencies is to characterize features at various scales in the radial positions.  $\mathbf{T}_R$  contains multiple rotation matrices, and it can be specified as

$$\mathbf{T}_R = \begin{bmatrix} 1 & 0 & \dots & \cos \theta_i & \sin \theta_i & \dots \\ 0 & 1 & \dots & -\sin \theta_i & \cos \theta_i & \dots \end{bmatrix}, \quad (3)$$



**Fig. 1.** (a) Schematic diagram of our proposed NFTPM method. (b) The physics prior (forward image formation model) of NFTPM.

where  $\theta_i = 2\pi i/N_\theta$  ( $i = 0, 1, \dots, N_\theta - 1$ ), and  $N_\theta$  is the number of rotation intervals. The rotation  $\theta_i$  further enables the MLP to respond to features at diverse orientations, allowing for better feature representation and avoiding noise [21].  $T_f\{\cdot\}$  is applied to flatten concatenated matrices into a vector, which for matrices  $\mathbf{A}$  and  $\mathbf{B}$  can be defined as

$$T_f\{\mathbf{A}, \mathbf{B}\} = T_f \left\{ \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} & b_{11} & b_{12} & \cdots & b_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} & b_{21} & b_{22} & \cdots & b_{2N} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MN} & b_{M1} & b_{M2} & \cdots & b_{MN} \end{bmatrix} \right\} \quad (4)$$

$$= [a_{11}, a_{12}, \dots, a_{1N}, b_{11}, b_{12}, \dots, b_{1N}, \dots, a_{M1}, a_{M2}, \dots, a_{MN}, b_{M1}, b_{M2}, \dots, b_{MN}].$$

Here, we simplify the radial encoding module by replacing Eq. (3) with

$$\mathbf{T}_R = \begin{bmatrix} 1 & \cdots & \cos \theta_i & \cdots \\ 0 & \cdots & \sin \theta_i & \cdots \end{bmatrix}, \quad (5)$$

and including  $\mathbf{r} = (x, y)$  into the encoded feature. Hence,  $\mathbf{r} = (x, y)$  can be processed like the 2D Fourier series expansion

$$\mathbf{r}_{\text{rad}} = [x, y, \cdots \cos(2\pi(\mathbf{u}_{l,i}x + \mathbf{v}_{l,i}y)), \sin(2\pi(\mathbf{u}_{l,i}x + \mathbf{v}_{l,i}y)), \cdots]_{l:0 \leq l \leq L-1}, \quad (6)$$

where  $\mathbf{r}_{\text{rad}} \in \mathbb{R}^{1 \times (2+2LN_\theta)}$ ,  $\mathbf{u}_{l,i} = [2^{l-1} \cos(\theta_i)]_{i:0 \leq i \leq N_\theta-1}$ , and  $\mathbf{v}_{l,i} = [2^{l-1} \sin(\theta_i)]_{i:0 \leq i \leq N_\theta-1}$ . To retrieve the phase distribution, we input  $\mathbf{r}_{\text{rad}}$  into the  $K$ -layer MLP of the neural field  $\Phi_{\mathbf{W}}$  ( $K = 5$ ), incorporating the following processing modules. (1) Linear module  $\mathbf{W}_0 \in \mathbb{R}^{(2+2LN_\theta) \times C}$ , converting  $\mathbf{r}_{\text{rad}}$  into hidden features with  $C$  channels ( $C = 128$ ). (2) Linear module for hidden features  $\mathbf{W}_i \in \mathbb{R}^{C \times C}$  ( $i = 1, 2, \dots, K-2$ ). (3) The last linear transformation  $\mathbf{W}_{K-1} \in \mathbb{R}^{C \times 1}$ . (4) Leaky rectified linear unit (LeakyReLU)  $\sigma_i$  ( $i = 0, \dots, K-2$ ). (5) Sigmoid activation

function  $\sigma_{K-1}$ . Specifically, let the  $i$ th feature be denoted as  $\mathbf{f}_i$ , and then the  $(i+1)$ th feature is given by

$$\mathbf{f}_{i+1} = \sigma_i(\mathbf{f}_i \mathbf{W}_i), \quad (7)$$

where  $i = 0, \dots, K-1$ , and  $\mathbf{f}_0 = \mathbf{r}_{\text{rad}}$ . The phase value can be represented as  $\phi(\mathbf{r}) = \Phi_{\mathbf{W}}(\mathbf{r}) = 2\pi \mathbf{f}_K$  ( $\mathbf{f}_K$  is the output of the  $K$ -layer MLP), and  $\phi(\mathbf{R})$  can be reshaped as an image of phase distribution. For most biological samples, the complex transmittance can be expressed as  $t(\mathbf{r}) = e^{j\phi(\mathbf{r})}$ . In a typical  $6f$  optical imaging system, the source with distribution  $S_{\text{pc}}(\mathbf{u})$  at the aperture diaphragm plane ( $\mathbf{u}$  corresponds to the 2D coordinates in Fourier space) provides partially coherent illumination, resulting in an image captured at the image plane,

$$I(\mathbf{r}) = \iint T(\mathbf{u}_1) T^*(\mathbf{u}_2) \text{TCC}(\mathbf{u}_1, \mathbf{u}_2) e^{j2\pi \mathbf{r}(\mathbf{u}_1 - \mathbf{u}_2)} d\mathbf{u}_1 d\mathbf{u}_2, \quad (8)$$

where  $T(\mathbf{u})$  is the Fourier transform of  $t(\mathbf{r})$ , and TCC (transmission cross-coefficient) [10,28] satisfies the following relation:

$$\text{TCC}(\mathbf{u}_1, \mathbf{u}_2) = \int S_{\text{pc}}(\mathbf{u}) P(\mathbf{u} + \mathbf{u}_1) P^*(\mathbf{u} + \mathbf{u}_2) d\mathbf{u}, \quad (9)$$

where  $P(\mathbf{u}) = |P(\mathbf{u})| e^{jkz\sqrt{1-\lambda^2|\mathbf{u}|^2}}$  represents the complex pupil function of the imaging system,  $z$  is the defocus distance along the optical axis,  $k$  is the wavenumber, and  $|P(\mathbf{u})|$  is a circular function determined by the objective NA and wavelength  $\lambda$ . The TCC formula is an abstraction for the spectral coupling of a light source to an objective pupil, intrinsically characterizing the imaging system compatible with partially coherent illumination. When the illumination distribution  $S_{\text{pc}}(\mathbf{u})$  of the

imaging system is specified, the captured image is determined by the sample's inherent property (phase delay  $\phi$ ) and the defocus distance  $z$ . Therefore, we can use a function  $H\{\phi, z\}$  to represent the image formation model of  $I$ .

In order to achieve phase retrieval with defocus distance prediction, the uncertain defocus distance can be incorporated into the computational graph as a tunable parameter  $z$  to be optimized along with the MLP. The trade-off in determining the optimal solution of NFTPM is to ensure the accuracy of the



predicted defocus distance while minimizing the error between the generated intensity image and the measurement. Given a captured intensity  $I$ , the spatial coordinates  $\mathbf{R} = \{\mathbf{r}_i\}_{i=0}^{M \times N - 1}$  are fed into  $\Phi_{\mathbf{W}}$  to obtain the phase, which is then processed through the physical model  $H\{\phi, z\}$  to generate intensity  $\tilde{I}$  for comparison with  $I$  using the mean square error (MSE) loss function. The above operations can be abstracted into an optimization problem

$$\mathbf{W}^\dagger, z^\dagger = \arg \min_{\mathbf{W}, z} \sum_{\mathbf{r} \in \mathbf{R}} \left\| H\{\Phi_{\mathbf{W}}(\mathbf{R}), z\} - I \right\|_2^2, \quad (10)$$

where  $\Phi_{\mathbf{W}}(\mathbf{R})$  is the retrieved phase, and  $z^\dagger$  is the predicted defocus distance. The optimization is executed based on back-propagation and the gradient descent algorithm [29], and the specific optimization process is described in Section 7 of Ref. [30]. It is worth mentioning that the samples are assumed as pure phase objects in NFTPM, so the phase contrast provided by a single-shot defocused intensity is sufficient for precise phase recovery based on the principle of deep image prior [18,31].

## B. Experimental Setup

Neural-field-assisted transport-of-intensity phase microscopy can be easily implemented on a commercial inverted bright-field microscope (IX83, Olympus, Japan) assisted by programmable LED array illumination due to the advantage of non-interferometric measurements. The LED array provides quasi-monochromatic illumination with a center wavelength of 525 nm and spectral bandwidth of 20 nm. These LED elements can be controlled to turn on to form point, circle, or annulus patterns by a field-programmable gate array (FPGA) unit (EP4CE10E22C8N, Intel, US). Twelve annularly distributed LED elements were selected in the array, with the center of the circle coinciding with the optical axis to provide matched annular illumination with a maximum illumination NA of 0.4. We utilized a Bertrand lens, positioned in an eyepiece observation tube in place of the normal eyepiece, to examine the rear focal plane of the objective lens. This examination is crucial for confirming that the circular illumination is centered precisely in the field of view or that the annular illumination is accurately inscribed within the objective lens's pupil. A CMOS camera (Hamamatsu ORCA-Flash 4.0 C13440) with a resolution of  $2048 \times 2048$  and a pixel size of  $6.5 \mu\text{m}$  was used to record the intensity information under a detection objective ( $10\times/0.4$  UPLSAPO, Olympus). This study was conducted on a workstation equipped with an Intel i9-10900K 3.70 GHz CPU and an NVIDIA GeForce RTX 3090 GPU. The proposed algorithm was operated by Python 3.7.16 and PyTorch 1.12.1.

## 3. RESULTS

### A. Comparison with TIE, AI-TIE, BlindNet, and GS Algorithms

To validate the effectiveness of NFTPM in partially coherent QPI, we conducted simulations to compare the proposed NFTPM with TIE, AI-TIE, BlindNet [19], and GS algorithms [4,5] under both coherent illumination and circular illumination with a coherence parameter (denoted by  $S$ , illumination NA/objective NA) of 0.85. It is noteworthy that AI-TIE here

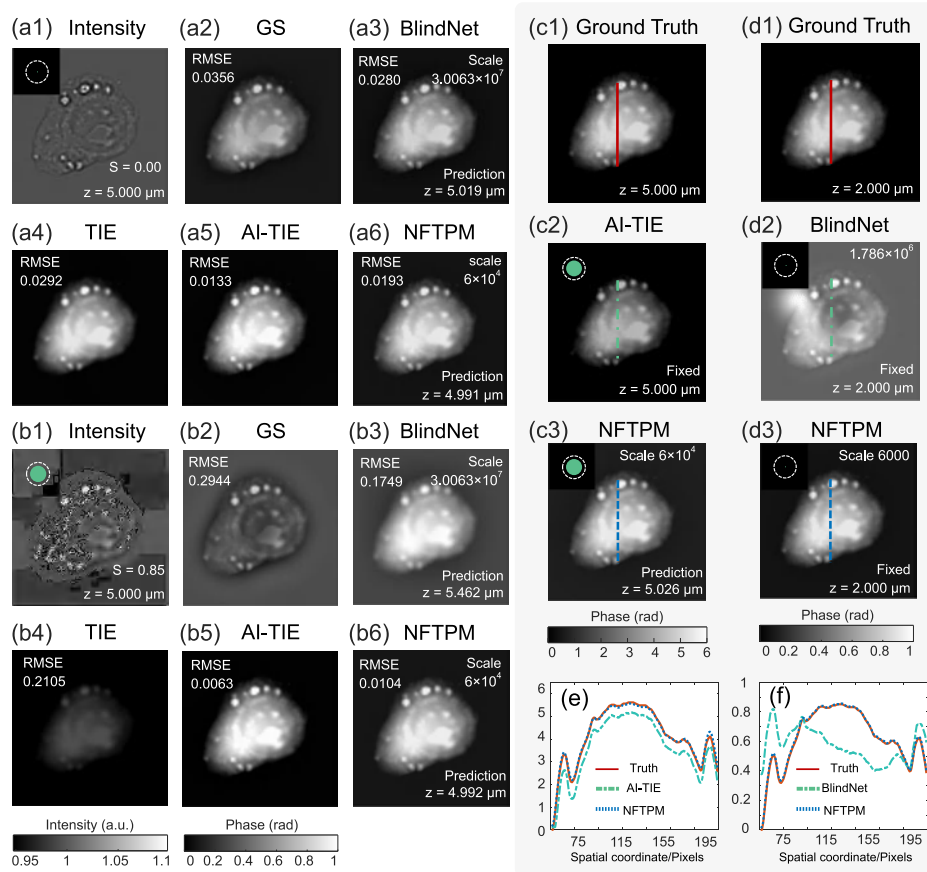
refers to all deconvolution-based TIE methods, adapted to coherent (point), partially coherent illuminations and not limited to annular illumination case. As shown in Figs. 2(a1) and 2(b1), the phase distribution of a HeLa cell (0–1 rad) was used to simulate an intensity image at a defocus distance of  $5 \mu\text{m}$ , defined within a grid of  $256 \times 256$  pixels (pixel size is  $6.5 \mu\text{m}$ ). The objective NA is 0.4 ( $20\times$  magnification), and the wavelength of monochromatic illumination is 550 nm. The comprehensive comparison under coherent illumination is detailed in Figs. 2(a2)–2(a6). Except for BlindNet and NFTPM (randomly given an initial value as the defocus distance, e.g.,  $z = 2 \mu\text{m}$ ), other methods were provided with the correct defocus value ( $z = 5 \mu\text{m}$ ). It can be observed that all methods achieve accurate phase recovery due to the exact match of the physical model and the optical parameters. Specifically, NFTPM and BlindNet show the ability to correctly predict the defocus distance, enabling robust QPI even with incorrect initialization of the defocus parameter. However, since BlindNet, TIE, and GS methods ignore the effect of partial coherence on the forward image formation process, the physical priors used by these methods do not accurately apply to the circular illumination situation, leading to a significant loss of high-frequency information in the partially coherent QPI [Figs. 2(b2)–2(b4)].

In contrast, AI-TIE and NFTPM demonstrate better performance at  $S = 0.85$  [Figs. 2(b5) and 2(b6)], as they establish a nonlinear forward model that conforms to partially coherent illumination by considering illumination distribution in modeling. Besides, we extended the simulated phase range to 0–6 rad (non-weak object) to validate that NFTPM is beyond weak object approximation. The results in Fig. 2(c) show that AI-TIE suffers from low-frequency underestimation, and Fig. 2(e) quantitatively reflects the inaccuracy of the phase image recovered by AI-TIE, while the result of NFTPM is consistent with the ground truth. It is worth noting that NFTPM, with a parameter count of  $6 \times 10^4$  (MLP) and the inference time of  $3 \times 10^{-4}$  s, outperforms BlindNet in terms of speed, which requires a much larger parameter count of  $3.0063 \times 10^7$  (UNet) and a comparatively longer inference time of  $5.8 \times 10^{-3}$  s. Furthermore, when the network size is reduced (fewer channels  $C$  per layer), the representational capacity of 2D UNet ( $1.786 \times 10^6$ ) is significantly weaker, resulting in the deterioration of phase retrieval in BlindNet [Fig. 2(d)] and the mismatch profile [Fig. 2(f)], while NFTPM ( $6 \times 10^3$ ) remains robust. In addition, as shown in Fig. S1 in Ref. [30], we also discuss the impact of hyperparameter tuning ( $K, L, N_\theta$ ) on NFTPM, elucidating the stability of NFTPM against the layer changes in the MLP as well as the significance of radial encoding for the high-frequency characterization (see Section 1 in Ref. [30] for detailed analysis).

### B. Verification of QPI at Unknown Defocus Distances for Different Illuminations

Further simulations are shown to verify that NFTPM can accurately recover the phase without prior knowledge of the defocus distance in different partially coherent illuminations. In Figs. 3(a1)–3(a4), we simulated intensity images at  $z = 7 \mu\text{m}$  under circular illuminations ( $S = 0.10, 0.40, \text{ and } 0.75$ ) and annular NA-matched illumination. We randomly provided





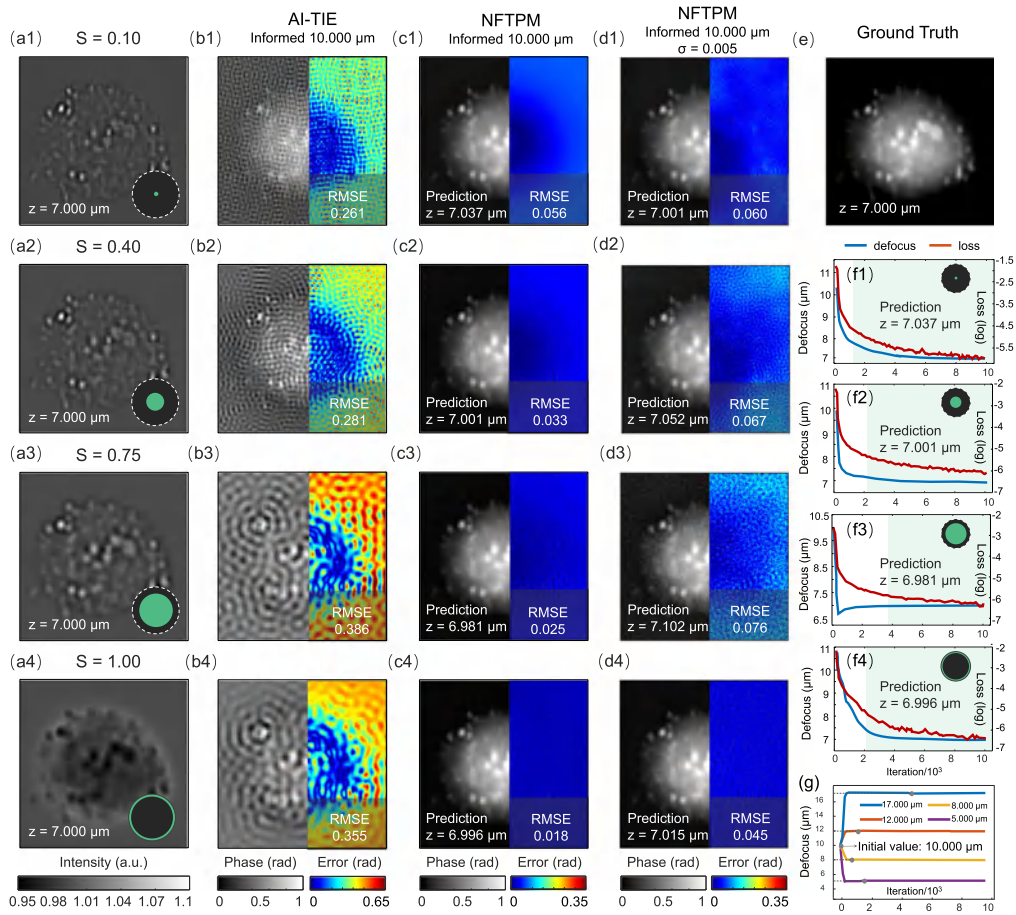
**Fig. 2.** Comparison of phase (0–1 rad) reconstruction results of a simulated cell sample using NFTPM, BlindNet, TIE, AI-TIE, and GS methods under different illumination settings and a defocus distance of 5  $\mu\text{m}$ . (a) Results under coherent illumination. (b) Results under partially coherent illumination ( $S = 0.85$ ). (c) Comparison between AI-TIE and NFTPM for the large phase (0–6 rad) under partially coherent illumination ( $S = 0.85$ ). (d) Comparison between BlindNet and NFTPM after downsizing the network under coherent illumination. (e), (f) Phase line profiles corresponding to (c), (d).

an incorrect initial value of the defocus distance (10  $\mu\text{m}$ ) for the AI-TIE and NFTPM (the robustness of NFTPM to defocus distance initialization is verified in Fig. S2 and Section 2 of Ref. [30]) and evaluated the quality of the retrieved phase using the root mean square error (RMSE). In Figs. 3(b1)–3(b4), severe artifacts appear in the results of AI-TIE, as the phase transfer function (PTF) is mis-estimated due to the uncorrected  $z$  value. In contrast, NFTPM achieves high-precision phase retrieval (RMSE < 0.06) in diverse illuminations and accurately predicts the defocus value based on the tunable defocus parameter  $z$ . We also supplement simulations under other special illuminations (e.g., asymmetric semicircular illumination) to display the adaptability of NFTPM to arbitrary source distribution (see Fig. S3 and Section 3 of Ref. [30] for details). As depicted in Figs. 3(c1)–3(c4), the RMSE of the phase recovered by NFTPM progressively decreases with the increasing maximum illumination angle, demonstrating that NFTPM has higher imaging accuracy at large illumination angles in the absence of noise. However, under noisy conditions (Gaussian noise with a standard deviation of 0.005), as the increase in the illumination angle reduces the response amplitude of the PTF [see Figs. S4(a)–S4(c) in Ref. [30]], the sensitivity to noise

instead leads to an escalation in the RMSE, as illustrated in Figs. 3(d1)–3(d3). Although the annular NA-matched illumination has a larger illumination angle compared to the circular illumination with  $S$  of 0.4 and 0.75, it has a relatively smaller RMSE in the presence of noise [Fig. 3(d4)], owing to its improved spatial frequency response that allows for higher robustness to noise [see Figs. S4(e) and S4(f) in Ref. [30]]. Essentially, NFTPM can be regarded as an iterative process that simultaneously seeks the optimal solutions for defocus distance prediction and phase retrieval. As shown in Figs. 3(f1)–3(f4), the loss function exhibits a steady decline along with a converging trend of defocus distance  $z$  (towards 7  $\mu\text{m}$  in all cases), indicating the parallel optimization of the model parameters and the defocus parameter. Additionally, simulations under annular illumination for various  $z$  (5  $\mu\text{m}$ , 8  $\mu\text{m}$ , 12  $\mu\text{m}$ , and 17  $\mu\text{m}$ ) in Fig. 3(g) validate the stability of NFTPM for defocus distance prediction.

### C. QPI Experiments for Live HeLa Cells

In the actual experiment, the long-time imaging of living HeLa cells [Fig. 4(a)] was performed using the inverted microscope (IX83) without motor drive adjustment or manual correction. Under the influence of temperature fluctuation and other



**Fig. 3.** Comparison between AI-TIE and NFTPM with an incorrect initial value of the defocus distance in various illuminations. (a) Intensity images under circular illuminations ( $S = 0.10, 0.40,$  and  $0.75$ ) and annular NA-matched illumination. (b) The results of AI-TIE. (c) The results of NFTPM. (d) The results of NFTPM in the presence of noise. (e) Ground truth. (f) Convergence curves of the defocus value and loss value (in logarithmic form). (g) The defocus distance prediction process of NFTPM for intensity images simulated at other defocus distances under annular illumination.

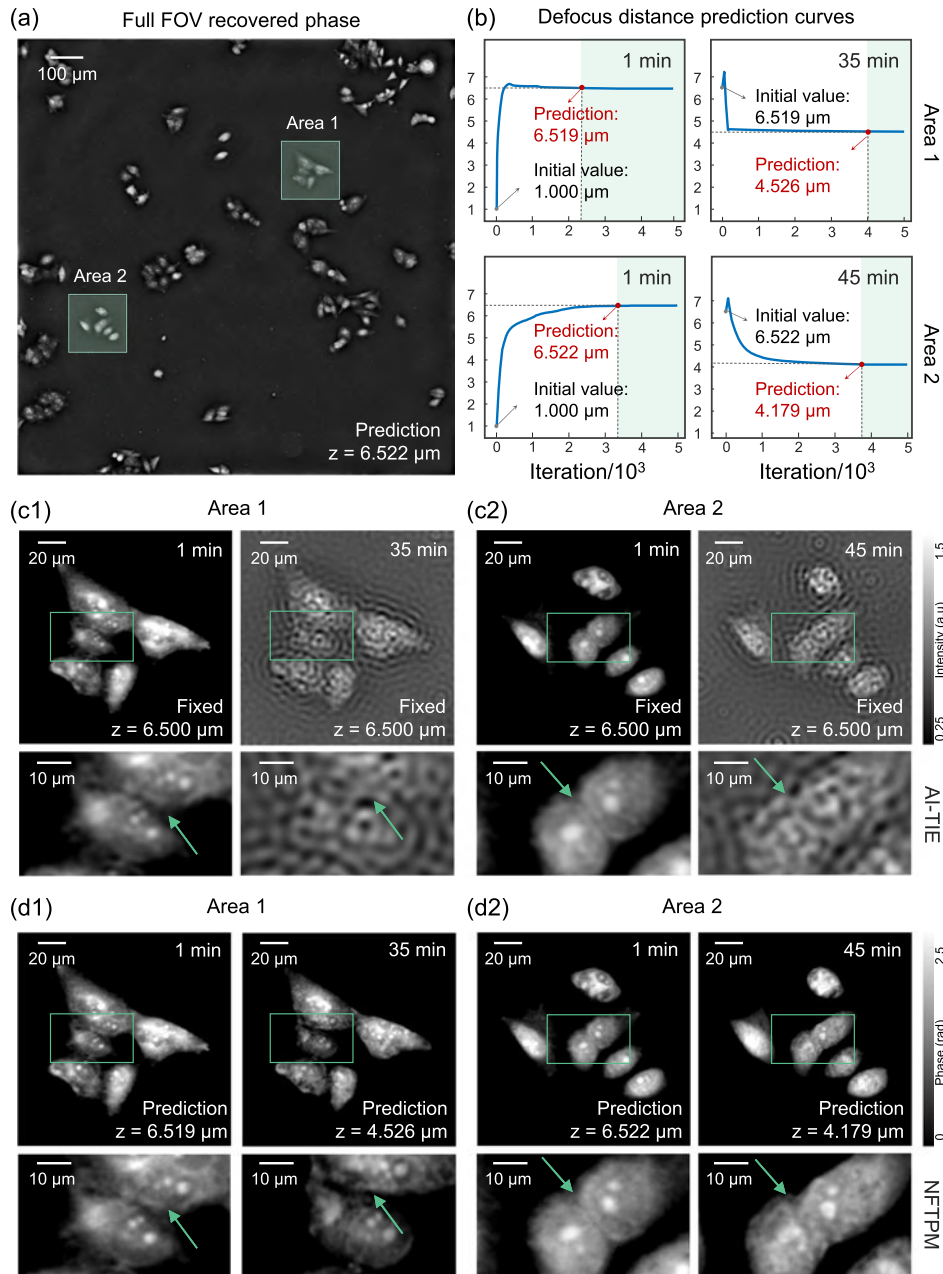
factors that lead to focal drift, NFTPM shows its stability of phase recovery by correctly predicting the unknown defocus distance. Figure 4(b) demonstrates the iterative process of predicting the defocus distance using the NFTPM in Area 1 and Area 2 at different moments. Taking  $1 \mu\text{m}$  as the initial value, we recovered the phase of the first frame measurement by NFTPM with the predicted defocus value of  $6.522 \mu\text{m}$  and adopted  $6.5 \mu\text{m}$  as the defocus distance for AI-TIE. Since the defocus distance was fixed at all moments, the retrieved results of AI-TIE, as depicted in Figs. 4(c1) and 4(c2), gradually deteriorated over time due to the model mismatch induced by the time-varying defocus distance. On the contrary, NFTPM dynamically reconstructed the phase information at different moments through adaptive defocus parameter correction, revealing distinct subcellular details such as nuclei and lipid droplets [Figs. 4(d1) and 4(d2)]. Once NFTPM reconstructs the phase for the measurement at a given moment using random initialization, the QPI of subsequent frames can be accelerated by initializing NFTPM with the MLP model and the predicted  $z$  corresponding to the present frame, which utilizes the correlation between frames in the same FOV. Additionally, the efficiency of full FOV phase retrieval can be improved

fivefold by utilizing the pre-iterated model of the subregion to initialize the neural field.

#### 4. DISCUSSION AND CONCLUSION

In summary, we have proposed a new partially coherent QPI method called NFTPM using the neural field. NFTPM, a single-shot non-interferometric iterative method, employs a straightforward MLP model for continuous phase representation and can accurately predict the defocus distance without prior knowledge, which eliminates the necessity for precise motor drive adjustment or manual correction for focus drift.

The rough defocus distance provided by the focusing device can be manually adjusted to reduce the WOTF error and thus improve the reconstruction results of AI-TIE. But with the limited time and manpower required, AI-TIE still cannot be applied as an effective method for long-term live cell imaging. In contrast, NFTPM replaces the costly manual operation with gradient-based tuning, which is based on the backpropagation algorithm in the prior model and the MLP. Therefore, NFTPM can adaptively obtain reconstruction results without defocus artifacts.



**Fig. 4.** Experimental observation of HeLa cells via NFTPM under annular NA-matched illumination (see Visualization 1). (a) The full FOV of the reconstruction result of NFTPM. (b) The defocus distance prediction process. (c1), (c2) The phase retrieved by AI-TIE. (d1), (d2) The phase retrieved by NFTPM.

For circular illumination, high-resolution reconstruction results from large-angle illumination, providing more high-frequency information. However, this comes at the expense of a steadily diminishing PTF response with increasing illumination angle, making noise more detrimental to the phase reconstruction of NFTPM. In contrast, annular illumination exhibits strong noise immunity due to a uniformly high response in its pass-band over a large illumination angle. To obtain high-quality QPI results, the defocus distance also needs to be selected appropriately. The low-frequency response of the PTF becomes weak when the defocus distance

is too small, which is not conducive to the recovery of low-frequency information. Besides, the PTF obtained at excessive defocus distance has a low response in its pass-band and contains multiple deep dips and zero-crossings, rendering this part of information susceptible to noise (see Figs. S5 and S6 in Section 4 of Ref. [30]). Remarkably, NFTPM can also be applied in pixel-aliasing conditions by additionally introducing pixel binning as a prior. Its capability of pixel super-resolution QPI is validated in Fig. S7 (Section 5 of Ref. [30]) by simulating a pixel-aliased defocused intensity of a USAF resolution test target.



Although our method incorporates partially coherent illumination into the forward image formation model, other optical parameters that are beneficial for improving reconstruction quality are still overlooked. Therefore, in the future, more optical parameters will be considered in the physical model to further promote the quality of phase reconstruction. For instance, it is possible to achieve prediction of unknown illumination by a grid search in a preset series of coherence parameters.

**Funding.** National Natural Science Foundation of China (62227818, 62105151, 62175109, U21B2033); National Key Research and Development Program of China (2022YFA1205002); Leading Technology of Jiangsu Basic Research Plan (BK20192003); Youth Foundation of Jiangsu Province (BK20210338); Biomedical Competition Foundation of Jiangsu Province (BE2022847); Key National Industrial Technology Cooperation Foundation of Jiangsu Province (BZ2022039); Fundamental Research Funds for the Central Universities (30920032101, 30923010206); Fundamental Scientific Research Business Fee Funds for the Central Universities (2023102001); Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging Intelligent Sense (JSGP202105, JSGP202201).

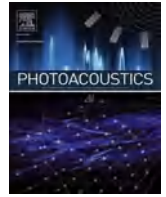
**Disclosures.** The authors declare no conflicts of interest.

**Data Availability.** The data that support the plots and maps within this paper and other findings of this study are available from the corresponding author upon reasonable request.

## REFERENCES

- G. Popescu, *Quantitative Phase Imaging of Cells and Tissues* (McGraw-Hill Education, 2011).
- X. Chang, L. Bian, and J. Zhang, "Large-scale phase retrieval," *eLight* **1**, 4 (2021).
- Z. Huang, P. Memmolo, P. Ferraro, *et al.*, "Dual-plane coupled phase retrieval for non-prior holographic imaging," *PhotonIX* **3**, 3 (2022).
- R. W. Gerchberg, "Phase determination from image and diffraction plane pictures in the electron microscope," *Optik* **34**, 275–284 (1971).
- R. W. Gerchberg, "A practical algorithm for the determination of plane from image and diffraction pictures," *Optik* **35**, 237–246 (1972).
- M. R. Teague, "Deterministic phase retrieval: a Green's function solution," *J. Opt. Soc. Am.* **73**, 1434–1441 (1983).
- C. Zuo, J. Li, J. Sun, *et al.*, "Transport of intensity equation: a tutorial," *Opt. Lasers Eng.* **135**, 106187 (2020).
- E. Barone-Nugent, A. Barty, and K. Nugent, "Quantitative phase-amplitude microscopy I: optical microscopy," *J. Microsc.* **206**, 194–203 (2002).
- L. Lu, J. Li, Y. Shu, *et al.*, "Hybrid brightfield and darkfield transport of intensity approach for high-throughput quantitative phase microscopy," *Adv. Photonics* **4**, 056002 (2022).
- C. Zuo, J. Sun, J. Li, *et al.*, "High-resolution transport-of-intensity quantitative phase microscopy with annular illumination," *Sci. Rep.* **7**, 7654 (2017).
- A. Sinha, J. Lee, S. Li, *et al.*, "Lensless computational imaging through deep learning," *Optica* **4**, 1117–1125 (2017).
- Y. Wu, Y. Rivenson, Y. Zhang, *et al.*, "Extended depth-of-field in holographic imaging using deep-learning-based autofocusing and phase recovery," *Optica* **5**, 704–710 (2018).
- Y. Rivenson, Y. Zhang, H. Günaydn, *et al.*, "Phase recovery and holographic image reconstruction using deep learning in neural networks," *Light Sci. Appl.* **7**, 17141 (2018).
- K. Wang, J. Di, Y. Li, *et al.*, "Transport of intensity equation from a single intensity image via deep learning," *Opt. Lasers Eng.* **134**, 106233 (2020).
- C. Zuo, J. Qian, S. Feng, *et al.*, "Deep learning in optical metrology: a review," *Light Sci. Appl.* **11**, 39 (2022).
- E. Bostan, R. Heckel, M. Chen, *et al.*, "Deep phase decoder: self-calibrating phase microscopy with an untrained deep neural network," *Optica* **7**, 559–562 (2020).
- F. Wang, Y. Bian, H. Wang, *et al.*, "Phase imaging with an untrained neural network," *Light Sci. Appl.* **9**, 77 (2020).
- D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 9446–9454.
- X. Zhang, F. Wang, and G. Situ, "Blindnet: an untrained learning approach toward computational imaging with model uncertainty," *J. Phys. D* **55**, 034001 (2021).
- B. Mildenhall, P. P. Srinivasan, M. Tancik, *et al.*, "NeRF: representing scenes as neural radiance fields for view synthesis," *Commun. ACM* **65**, 99–106 (2021).
- R. Liu, Y. Sun, J. Zhu, *et al.*, "Recovery of continuous 3D refractive index maps from discrete intensity-only measurements using neural fields," *Nat. Mach. Intell.* **4**, 781–791 (2022).
- H. Zhu, Z. Liu, Y. Zhou, *et al.*, "Dnf: diffractive neural field for lensless microscopic imaging," *Opt. Express* **30**, 18168–18178 (2022).
- H. Zhou, B. Y. Feng, H. Guo, *et al.*, "Fourier Ptychographic microscopy image stack reconstruction using implicit neural representations," *Optica* **10**, 1679–1687 (2023).
- L. Lu, Y. Fan, J. Sun, *et al.*, "Accurate quantitative phase imaging by the transport of intensity equation: a mixed-transfer-function approach," *Opt. Lett.* **46**, 1740–1743 (2021).
- E. Abbe, "Beiträge zur theorie des mikroskops und der mikroskopischen wahrnehmung," *Arch. Mikroskopische Anatomie* **9**, 413–468 (1873).
- S. Zhou, J. Li, J. Sun, *et al.*, "Transport-of-intensity Fourier Ptychographic diffraction tomography: defying the matched illumination condition," *Optica* **9**, 1362–1373 (2022).
- Y. Shu, J. Sun, J. Lyu, *et al.*, "Adaptive optical quantitative phase imaging based on annular illumination Fourier Ptychographic microscopy," *PhotonIX* **3**, 24 (2022).
- C. J. Sheppard, "Three-dimensional phase imaging with the intensity transport equation," *Appl. Opt.* **41**, 5951–5955 (2002).
- D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," *arXiv*, arXiv:1412.6980 (2014).
- "Supplemental document for 'Neural-field-assisted transport-of-intensity phase microscopy: partially coherent quantitative phase imaging under unknown defocus distance'," Figshare, 2024, [https://figshare.com/articles/journal\\_contribution/Supplementary\\_document\\_for\\_Neural-field-assisted\\_transport-of-intensity\\_phase\\_microscopy\\_partially\\_coherent\\_quantitative\\_phase\\_imaging\\_under\\_unknown\\_defocus\\_distance/25835650](https://figshare.com/articles/journal_contribution/Supplementary_document_for_Neural-field-assisted_transport-of-intensity_phase_microscopy_partially_coherent_quantitative_phase_imaging_under_unknown_defocus_distance/25835650).
- D. Paganin, S. C. Mayo, T. E. Gureyev, *et al.*, "Simultaneous phase and amplitude extraction from a single defocused image of a homogeneous object," *J. Microsc.* **206**, 33–40 (2002).





## 4D spectral-spatial computational photoacoustic dermoscopy

Yang Gao<sup>a,b,c</sup>, Ting Feng<sup>d,\*</sup>, Haixia Qiu<sup>e</sup>, Ying Gu<sup>e</sup>, Qian Chen<sup>a,c</sup>, Chao Zuo<sup>a,b,c,\*</sup>,  
Haigang Ma<sup>a,b,c,\*</sup>

<sup>a</sup> Nanjing University of Science and Technology, School of Electronic and Optical Engineering, Smart Computational Imaging Laboratory (SCILab), Nanjing 210094, China

<sup>b</sup> Smart Computational Imaging Research Institute (SCIRI) of Nanjing University of Science and Technology, Nanjing 210094, China

<sup>c</sup> Nanjing University of Science and Technology, School of Electronic and Optical Engineering, Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing 210094, China

<sup>d</sup> Fudan University, Academy for Engineering and Technology, Shanghai 200433, China

<sup>e</sup> First Medical Center of PLA General Hospital, Beijing 100853, China

### ARTICLE INFO

#### Keywords:

4D spectral-spatial  
Skin imaging  
Photoacoustic dermoscopy  
Quantitative calculation  
Learning from simulation

### ABSTRACT

Photoacoustic dermoscopy (PAD) is an emerging non-invasive imaging technology aids in the diagnosis of dermatological conditions by obtaining optical absorption information of skin tissues. Despite advances in PAD, it remains unclear how to obtain quantitative accuracy of the reconstructed PAD images according to the optical and acoustic properties of multilayered skin, the wavelength and distribution of excitation light, and the detection performance of ultrasound transducers. In this work, a computing method of four-dimensional (4D) spectral-spatial imaging for PAD is developed to enable quantitative analysis and optimization of structural and functional imaging of skin. This method takes the optical and acoustic properties of heterogeneous skin tissues into account, which can be used to correct the optical field of excitation light, detectable ultrasonic field, and provide accurate single-spectrum analysis or multi-spectral imaging solutions of PAD for multilayered skin tissues. A series of experiments were performed, and simulation datasets obtained from the computational model were used to train neural networks to further improve the imaging quality of the PAD system. All the results demonstrated the method could contribute to the development and optimization of clinical PADs by datasets with multiple variable parameters, and provide clinical predictability of photoacoustic (PA) data for human skin.

### 1. Introduction

Skin is the largest organ of the human body, whose health is closely related to the whole body. Skin diseases, one of the most common ailments among humans, are characterized by structural and functional changes in the tissue components of the skin. Imaging technologies play an essential role in dermatology, providing non-invasive means of observation and diagnosis, and offering valuable information for clinical practitioners [1,2]. Traditional optical dermoscopy, such as confocal microscopy and optical coherence tomography, is limited to observing only the conformation of the epidermis and superficial dermis due to imaging depth constraints. Ultrasound imaging enables visualization of the entire skin structure through deep penetration, but its spatial resolution and ability to visualize microvasculature are poor, and it is unable

to obtain metabolism-related biochemical information [3–5]. Photoacoustic imaging (PAI), an emerging imaging technology with both high spatial resolution and deep tissue imaging capabilities, has received widespread attention from the biomedical research community [6]. PAI is expected to bring more opportunities for the maintenance of skin health and the treatment of diseases.

As an effective application mode of PAI, photoacoustic dermoscopy (PAD), an emerging non-invasive imaging technique, can directly measure the optical absorption characteristics of tissues, thus facilitating the diagnosis of skin diseases [7]. PAD combines the advantages of optical and ultrasound imaging. By illuminating short-pulsed laser onto the skin, and then receiving the ultrasound signals generated by endogenous chromophores (hemoglobin, melanin, lipids, collagen, glucose, etc.), PAD can provide clinical practitioners with high-contrast and

\* Corresponding authors at: Nanjing University of Science and Technology, School of Electronic and Optical Engineering, Smart Computational Imaging Laboratory (SCILab), Nanjing 210094, China.

E-mail addresses: [gaoyang6613@njust.edu.cn](mailto:gaoyang6613@njust.edu.cn) (Y. Gao), [fengting@fudan.edu.cn](mailto:fengting@fudan.edu.cn) (T. Feng), [qiuhref@126.com](mailto:qiuhref@126.com) (H. Qiu), [guyinglaser301@163.com](mailto:guyinglaser301@163.com) (Y. Gu), [chenq@njust.edu.cn](mailto:chenq@njust.edu.cn) (Q. Chen), [zuochao@njust.edu.cn](mailto:zuochao@njust.edu.cn) (C. Zuo), [mahaigang@njust.edu.cn](mailto:mahaigang@njust.edu.cn) (H. Ma).

<https://doi.org/10.1016/j.pacs.2023.100572>

Received 11 September 2023; Received in revised form 16 October 2023; Accepted 9 November 2023

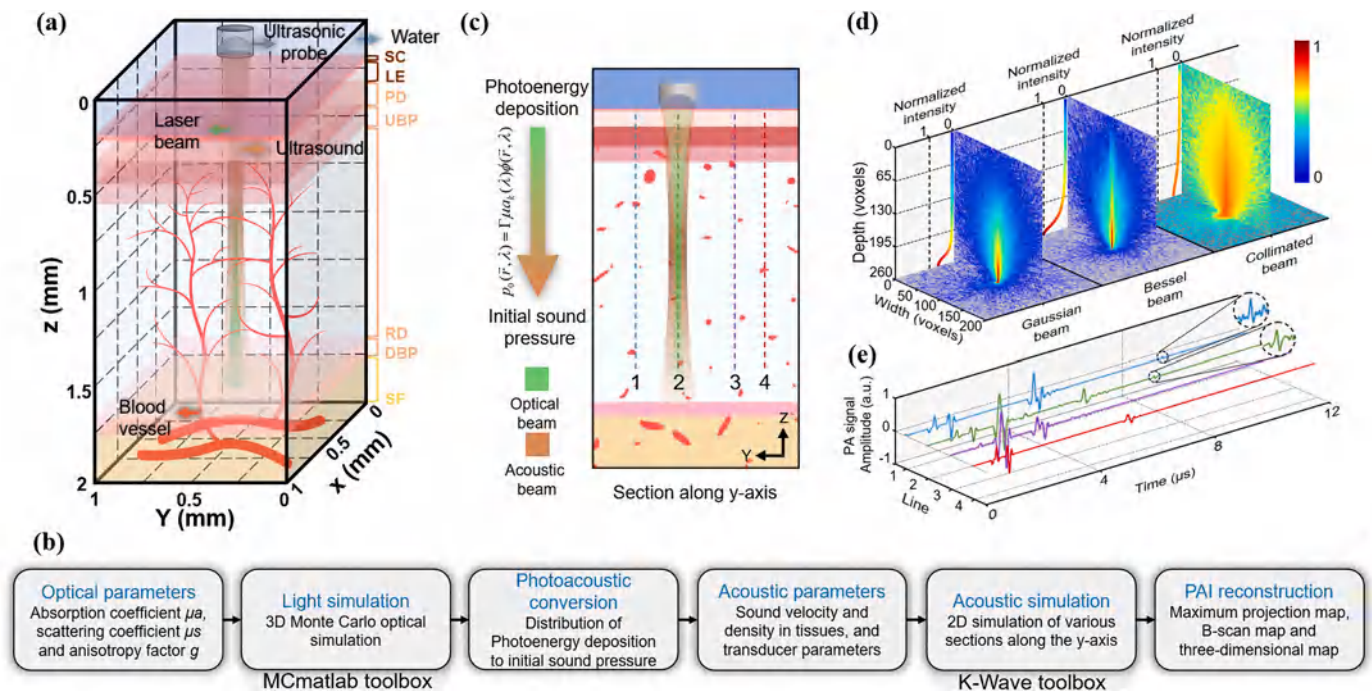
Available online 10 November 2023

2213-5979/© 2023 The Authors.

Published by Elsevier GmbH. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

**Table 1**  
The parameters of skin layers used in the computational model [23,35,43].

Layer	Thickness (mm)	$W_k$ (%)	$B_k$ (%)	$M_k$ (%)	$\mu_s R_k$ ( $\text{cm}^{-1}$ )	Velocity (m/s)	Density ( $\text{kg/m}^3$ )
Stratum corneum	0.01	0	0	1	80	1540	1500
Living epidermis	0.08	60	0	10	80	1720	1190
Papillary dermis	0.1	75	3	1	80	1650	1200
Upper blood plexus	0.08	75	3.8	0	40	1650	1200
Reticular dermis	1.2	75	3	0	40	1790	1200
Deep blood plexus	0.07	75	2.3	0	40	1540	1116
Subcutaneous fat	3	5	2.1	0	42	1450	971



**Fig. 1.** 4D spectral-spatial computational model of skin. (a) Schematic diagram of a seven-layer skin model. SC: stratum corneum, LE: living epidermis, PD: papillary dermis, UBP: upper blood plexus, RD: reticular dermis, DBP: deep blood plexus, SF: subcutaneous fat. (b) Flowchart of key steps. (c) 2D simulation diagram. (d) Focused Gaussian beam, Bessel beam, collimated Gaussian beam photon distribution. (e) Photoacoustic signals along the dotted lines position in (c).

high-resolution morphological, functional, and pathological information of the skin, which has great potential in biomedical research and clinical applications [8]. In recent years, several PAD systems have been developed for imaging melanoma [9,10], café-au-lait macules [11], psoriasis [12], and skin blood vessels [13–15]. Despite the progress made in PAD research, there is still a lack of research on realistic modeling of the optical and acoustic properties of multilayered skin tissues for the quantitative accuracy of reconstructed PAD images caused by the wavelength and distribution of the excitation light, and the acoustic properties of the ultrasound transducer. The reliable computational methods can benefit the optimization design of the optical and acoustic parameters of PAD equipment.

Although deep learning has been used for PAI, most deep learning-based photoacoustic imaging needs thousands pairs of labeled input-output data to train the neural network, especially those applications in clinical skin imaging, which requires even larger amounts of data. It also should be noted that in many cases the ground truth corresponding to the experimental data is inaccessible. In such cases, an efficient “learning from computational model” scheme is urgently needed to obtain matching datasets. In addition, human skin tissues are multi-layered physiopathological structures with variability in optical absorption and acoustic impedance, which requires a rigorous computational model of the physical process of PAI.

The computational model of PAI, which encompasses both optical

and acoustic simulations, plays a vital role. Various methodologies such as the radiative transfer equation [16], the finite element method [17], and the Monte Carlo (MC) method have been employed to simulate the scattering and absorption of light in tissues and to capture the distribution of luminous flux in tissues. Of these, the Monte Carlo method stands out as the gold standard, which is widely used to calculate the propagation of light in complex tissue structures [18–20]. However, many existing MC algorithms are for simpler layered tissue models [21–24]. Most models resort to representing skin as a single or triple-layered homogeneous medium, a simplification that often overlooks the intricate optical characteristics of multi-layered skin structures [25]. For the simulation of acoustics, tools such as the k-Wave toolbox [26] and the finite element method, with platforms like COMSOL, are predominant. The k-Wave, in particular, has gained traction due to its efficiency and simplicity [27–32]. Despite these advances, the current literature has limitations in certain aspects of photoacoustic dermoscopy (PAD) simulations, especially models that incorporate the intricate interactions between excitation optical fields and detectable ultrasonic fields. Furthermore, while multi-spectral PAI computational models have been explored [33,34], there is still molecular information on the depth- and wavelength-dependent multispectral PAD imaging that has not been studied enough for us to make authoritative statements.

In this study, we propose a photoacoustic hybrid 4D spectral-spatial computational model aimed at in-depth analysis of PAD skin structure

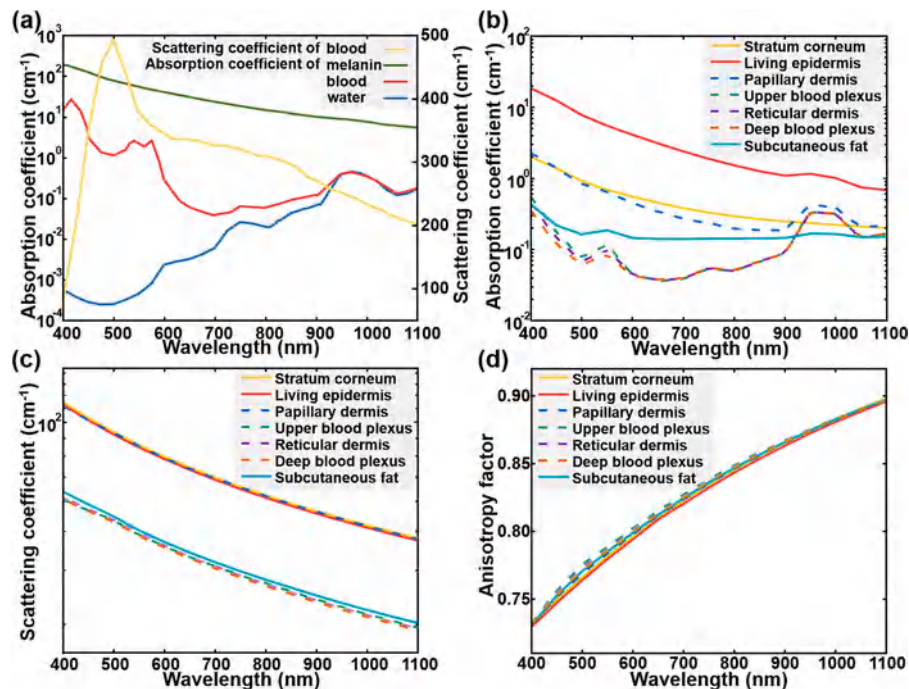


Fig. 2. The optical properties of each layer of the skin. (a) The absorption spectra of blood, melanin, and water, as well as the scattering spectra of blood [36,44–47]. (b) The absorption spectra of each layer of the seven-layer skin model. (c) Scattering spectrum. (d) Anisotropic factors.

and functional imaging for system optimization. Our model prioritizes the simulation of experimental scenarios as much as possible, adopts a point-by-point scanning mode based on photoacoustic microscopy, the model integrates forward propagation of light based on the Monte Carlo method and backward propagation of ultrasound computed based on k-Wave, and takes into account the multilayered heterogeneous structure of the skin as well as a specific vascular model, which makes the photoacoustic physical process of our proposed model on the more accurate. Furthermore, we have also incorporated the spectral dimension, achieving multispectral PAD imaging and unmixing with multiple beam types at varying wavelengths and energy levels. This allows for quantitative measurement of component intensities, which has the potential to greatly aid in disease diagnostic applications [10,34]. In addition, the model helps to accurately calibrate subcutaneous optical and acoustic distributions, providing a precise and optimized solution for imaging multilayered skin tissues using a PAD system. Finally, this study illustrates how the dataset obtained from our computational model can be utilized for neural network training to further break through hardware and biological constraints to improve the imaging quality of PAD experimental images.

The remainder of the paper is organized as follows: Section II describes the structure of the 4D spectral-spatial computational model, including the optical and acoustic properties of the tissue, the workflow of the computational model, the structure of the network model, the PAD experimental system and its quantitative optimization method. Section III presents a series of experimental results related to the reliability of the computational model. Also, the conclusion is summarized in Section IV.

## 2. Methodology

### 2.1. Simulation geometry

Here, the model defines a three-dimensional voxel grid with a size of 1 mm × 1 mm × 2 mm and a voxel count of 200 × 200 × 400 in each direction. A seven-layer skin model based on the anatomical structure of human skin (The skin model consists of seven layers: stratum corneum,

living epidermis, papillary dermis, upper blood plexus, reticular dermis, deep blood plexus, and subcutaneous fat, respectively. The thickness of each layer is shown in Table 1.) is constructed, which is modeled as a multilayered planar medium [35]. The thermal and optical parameters of the multilayered skin tissues do not vary with temperature is assumed. Based on the physical, optical, and physiological properties of the cells, and the pigmentation content, the skin is subdivided into sublayers on the basis of a three-layer skin model [36]. The epidermis can be subdivided into two sublayers: the stratum corneum and the living epidermis. The stratum corneum is thin and flat, composed of dead squamous cells, with a high degree of keratinization, high fat and protein content, and relatively low water content. The living epidermis contains most of the skin pigments, mainly melanin [37]. The dermis is a vascularized layer, with the main absorbers in the visible spectrum being hemoglobin, carotenoids, and bilirubin. It can be subdivided into four layers: the papillary dermis, the upper vascular plexus, the reticular dermis, and the deep vascular plexus [38]. These subdivided layers and the subcutaneous adipose tissue layer constitute the seven-layer model, which is illustrated in Fig. 1a. The epidermis layer contains no blood tissue, and a three-dimensional vascular model publicly available from *Tetteh et al.* [39] is inserted beneath the epidermis layer of the model, as an approximation of the skin vasculature. Fig. 1b shows the flowchart of the key steps of the model, which is based on photoacoustic microscopy to realize PAD imaging, and Fig. 1c shows a two-dimensional (2D) schematic of the scanning process. Fig. 1d illustrates the three types of beams used in the study and their photon distributions, including focused Gaussian beam, Bessel beam, and collimated Gaussian beam. Imaging under the focused beam corresponds to confocal optical-resolution photoacoustic microscopy while imaging under the collimated Gaussian beam corresponds to acoustic-resolution photoacoustic microscopy. Fig. 1e shows the simulated photoacoustic signals along the position of the dashed line in Fig. 1c.

### 2.2. Optical properties of tissue

Skin is a complex multilayered heterogeneous tissue, and the depth and direction of light propagation within the skin are determined by the



optical properties of the various layers of tissue and blood vessels in the skin, which are wavelength dependent and vary according to the random inhomogeneous distribution of various chromophores and pigments [38]. For simplicity, each layer is typically treated as a homogeneous structure in the computational model, and the optical properties vary between layers but remain constant within each layer [28–33]. Typically, the optical properties of each skin layer include the absorption coefficient ( $\mu_a$ ), scattering coefficient ( $\mu_s$ ), anisotropy factor ( $g$ ), and refractive index ( $n$ ). The refractive index does not vary significantly between layers, and therefore, the refractive index can be set to a fixed value of 1.4 for all wavelengths and under all skin layers [40–42].

The absorption coefficient of each skin layer is mainly contributed by three basic chromophores: blood, melanin, and water. The variation of absorption coefficients with wavelength for these three components is illustrated in Fig. 2a. Table 1 lists the thickness of each layer and the relative amount of the three chromophores. The absorption coefficient of each layer  $\mu_{a,k}$  can be calculated by Eq. 1 [23]:

$$\mu_{a,k}(\lambda) = B_k \mu_{a\_blood}(\lambda) + M_k \mu_{a\_melanin}(\lambda) + W_k \mu_{a\_water}(\lambda) + (1 - B_k - M_k - W_k) \mu_{a\_background} \quad (1)$$

Where,  $k$  represents the number of layers,  $\lambda$  is the wavelength at which the absorption coefficient is being calculated,  $B_k$ ,  $W_k$ ,  $M_k$  are the volume fractions of blood, water, and melanin in the layer, respectively.  $\mu_{a\_blood}$ ,  $\mu_{a\_water}$ ,  $\mu_{a\_melanin}$  and  $\mu_{a\_background}$  represent the absorption coefficients of blood, water, melanin, and background tissue, respectively. It can be considered that  $\mu_{a\_background}$  is independent of wavelength and is set as a fixed value of  $0.15 \text{ cm}^{-1}$  in the model. The calculated absorption spectra of each layer are presented in Fig. 2b.

The scattering coefficient of each skin layer in the model is mainly determined by blood. The variation of the scattering coefficient of blood with wavelength is illustrated in Fig. 2a. The scattering coefficient of each layer  $\mu_{s,k}$  can be calculated using Eq. 2:

$$\mu_{s,k}(\lambda) = B_k C_k \mu_{s\_blood}(\lambda) + (1 - B_k) \mu_s T_k(\lambda) \quad (2)$$

Where, the correction coefficient  $C_k$  is related to the diameter of blood vessels, it is assumed that the blood vessels in each skin layer have the same diameter, assuming  $C_k = 0.2$ .  $\mu_{s\_blood}$  represents the scattering coefficient of blood. The scattering coefficient  $\mu_s T_k(\lambda)$  of bloodless tissue varies with wavelength. In this study, Eq. 3 was used to calculate:

$$\mu_s T_k(\lambda) = \mu_s R_k \left( \frac{577 \text{ nm}}{\lambda} \right) \quad (3)$$

Where,  $\mu_s R_k$  is the scattering coefficient at the reference wavelength of 577 nm as shown in Table 1. The calculated scattering spectra of each layer are presented in Fig. 2c. The anisotropy factor  $g_k(\lambda)$  can be expressed as Eq. 4:

$$g_k(\lambda) = \frac{B_k C_k \mu_{s\_blood}(\lambda) g_{blood} + (1 - B_k) \mu_s T_k(\lambda) g T(\lambda)}{\mu_{s,k}(\lambda)} \quad (4)$$

Where,  $g T(\lambda)$  is the anisotropy factor of bloodless tissue, obtained through Eq. 5:

$$g T(\lambda) = 0.7645 + 0.2355 \left[ 1 - \exp \left( - \frac{\lambda - 500 \text{ nm}}{729.1 \text{ nm}} \right) \right] \quad (5)$$

The calculated anisotropy factors of each layer are presented in Fig. 2d.

### 2.3. Computational flowchart

This section describes the workflow of the 4D spectral-spatial computational PAD (Fig. 1b). The first step of the computation is to calculate the forward propagation of light in the tissue and the distribution of light energy deposition. We use the open-source Monte Carlo

toolkit MCmatlab to solve this problem. The input beam is simulated by emitting photon packets and calculating their paths in the simulated body [47]. There are three beams available in the model: focused Gaussian beam, Bessel beam, and collimated Gaussian beam (Fig. 1d). The beam is incident vertically along the Z-axis of the model, when photon packets propagate from one voxel to another, some energy is deposited into the voxel based on its absorption coefficient. The deposited energy is numerically accumulated in a three-dimensional matrix, which is the light energy deposition distribution. The light energy deposition distribution is then converted into an initial pressure distribution matrix using Eq. 6:

$$p_0(r, \lambda) = \Gamma(r) \mu_{a,k}(\lambda) \phi(r, \lambda) \quad (6)$$

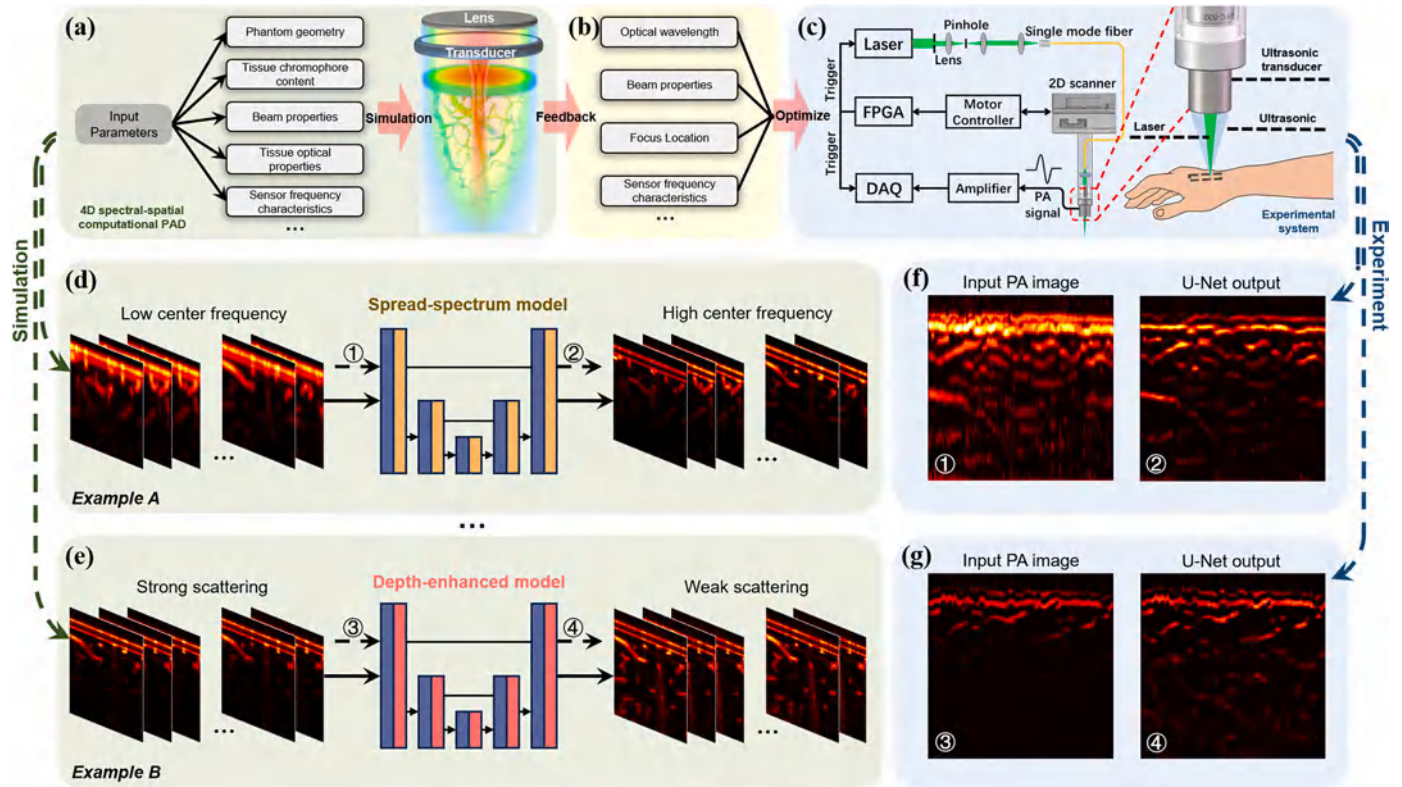
Where,  $\Gamma$  is the Gruneisen parameter, which measures the conversion efficiency from light absorption to sound pressure. In the research conducted in this article, it is assumed that the homogeneity value of  $\Gamma$  is 0.2 [48].  $\mu_{a,k}(\lambda)$  is the absorption coefficient of the corresponding dielectric layer  $k$  at position  $r$  at wavelength  $\lambda$ .  $\phi(r, \lambda)$  is the luminous flux at position  $r$  at wavelength  $\lambda$ . Using the equation, the initial pressure values for each grid position in the model are obtained, and the propagation of photoacoustic waves to the transducer at each grid position is implemented using k-Wave [26].

The speed and density of ultrasound for each skin layer related to the acoustic computation are recorded in Table 1. The attenuation of acoustic waves plays a pivotal role in determining the acoustic properties of tissues. One of the primary contributors to this attenuation is acoustic absorption. This frequency ( $f$ ) dependent attenuation is characterized using a power-law model. In the computational model of this paper, the acoustic attenuation coefficient of the tissue is taken as  $1 f^{1.5} \text{ dB/cm/MHz}^{1.5}$ . This signifies that the acoustic attenuation escalates in proportion to the frequency raised to the power of 1.5 [49,50]. A bowl-shaped focused ultrasound transducer is used in k-Wave with its focusing direction on the same axis as the center of the incident beam, and the center frequency and bandwidth are set accordingly to the experimental needs. Firstly, a 3D Monte Carlo optical computation is performed on the grid points on the scanning plane to obtain the light energy deposition distribution, which is then transformed to obtain the three-dimensional initial pressure distribution. Then, a 2D acoustic computation is performed on the plane of the scanning points along the Y-axis, and this process is repeated to achieve full scanning (Fig. 1c, e). A total of  $160 \times 160$  A-line signals are obtained, resulting in a maximum-intensity projection image. Generating the light energy deposition distribution in the optical model takes approximately 6 s, and collecting the raw photoacoustic signals in the two-dimensional acoustic model takes approximately 2 s. All calculations are performed on an Intel Core i7–10700KF CPU and NVIDIA RTX A2000 GPU. Compared with the three-dimensional acoustic model, the scanning method using the two-dimensional acoustic model is approximately 30 times faster.

### 2.4. Network architecture

U-Net is an encoder-decoder structure network with skip connections, which helps to preserve the detailed information of the image and helps to mitigate the loss of information when recovering the resolution in the decoder stage. U-Net has a relatively small number of parameters and computational complexity, which makes it faster in training and inference and performs well in small sample cases [51]. In order to achieve further optimization of systematic imaging based on computationally generated datasets, a modified U-Net architecture is used in this study (shown in Fig. S1), where the network accepts a  $624 \times 624$  grayscale image of skin blood vessels as input, the first layer contains 32 convolutional filters of size  $3 \times 3$ , and two successive convolution operations are activated by applying a leaky integrator linear unit LReLU layer (slope 0.2), and then convolutional operation is implemented using a  $2 \times 2$  convolutional layers instead of pooling to achieve down-sampling, which allows the model to learn how best to reduce the spatial





**Fig. 3.** The process of using the 4D spectral-spatial computational PAD combined with experiments for dataset acquisition and system optimization for deep learning. (a) Relevant parameters can be set before data acquisition, and the distribution of the model optical field and detector acoustic field under a collimated Gaussian beam in the model is shown. (b) Feedback on relevant performance optimization parameters is provided to the experimental system after simulating calculation. (c) Experimental system. (d) The dataset is used for training the spread-spectrum network model. (e) The dataset is used for training the depth-enhanced network model. (f) The low center frequency detector skin imaging results obtained in the experiment are input into the trained spread-spectrum model to obtain the output image. (g) The skin imaging results under conventional scattering obtained in the experiment are input into the trained depth-enhanced model to obtain the output image.

dimensions rather than relying on fixed operations, while better preserving certain features of the original input. The number of filters is incremented by powers of 2 up to the bottleneck layer, up to a maximum of 512 filters. The output is then upsampled using a  $2 \times 2$  transpose convolution to obtain an output of the same size as the input, and then two successive convolution operations are applied again. The number of channels in each layer is gradually reduced symmetrically. Skip connections are added while the corresponding downsampled layer is upsampled. Finally, the output image is obtained by  $1 \times 1$  convolutional downsampling. The total number of trainable parameters for the network is 8115009. The network is trained using the Adam optimizer's mean-square error (MSE) loss function ( $\beta_1 = 0.8$ ,  $\beta_2 = 0.999$ ), with the learning rate, the number of epochs, and the batch size set to  $3e-4$ , 100, and 2, respectively. During the model training process, we evaluate the model using the validation dataset periodically to avoid overfitting. We evaluated the network performance using mean-square error (MSE), mean-absolute error (MAE), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM) on the simulated test dataset.

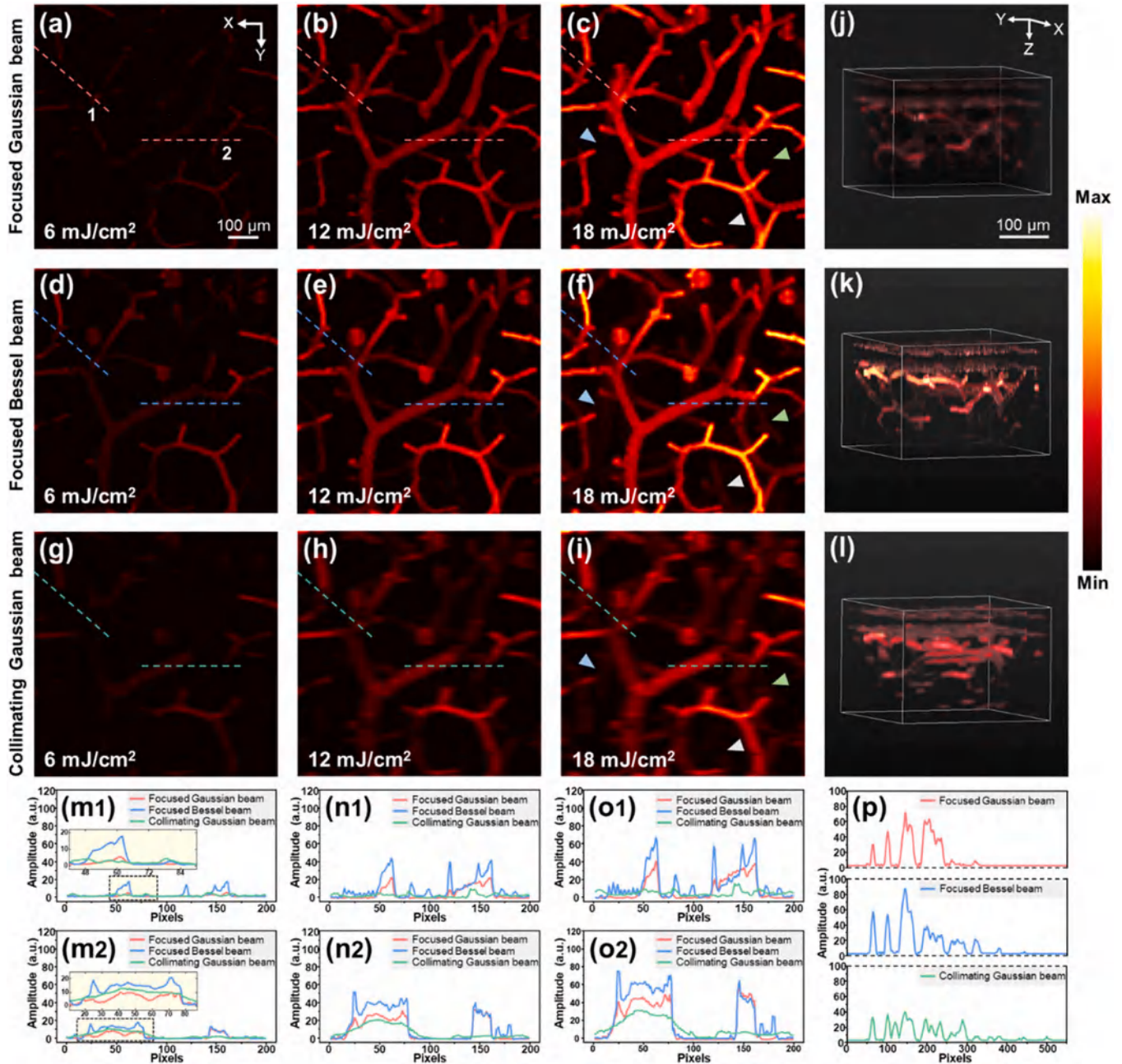
## 2.5. Experimental PAD imaging system

Fig. 3c shows the PAD imaging system employed for the experiment, using a 532 nm Q-switched pulsed laser (Talon 532-40, Spectra-Physics; pulse repetition rate of 10 kHz; pulse width of 20 ns) as a light source to excite the PA signal. The beam was passed through an optical spatial filter system and then coupled into a single-mode fiber with the help of a fiber coupler (PAF-X-7-A, Thorlabs Inc.). The fiber output laser beam was collimated by a fiber collimator (F240FC-532, Thorlab Inc.). Scanning was achieved by a two-dimensional linear motor (LS2-0830, JianCheng Technologies Ltd.) driven by a collimated light focused

through a  $5 \times$  objective lens (S Plan Apo HL 5x/0.13, SIGMA KOKI), and the signal was received by using a homemade hollow-bowl ultrasonic transducer, with a center frequency of 20 MHz and a bandwidth of about 100%. The laser fluence at the tissue surface was about  $18 \text{ mJ/cm}^2$ , which is below the ANSI safety limit of  $20 \text{ mJ/cm}^2$ . During raster scanning, the step between the two A-lines was  $1 \mu\text{m}$ . The acquired PA signals were amplified by a 50 dB low-noise amplifier (LNA-650, RF Bay), and then the amplified PA signals were digitized using a high-speed data acquisition card (M4i.4480, Spectrum). The acquired data were recorded and reconstructed in real-time by a LabVIEW program.

## 2.6. Quantitatively optimal photoacoustic dermoscopy

The corresponding parameters in the 4D spectral-spatial computational model can be set according to the wavelength, energy density and focusing position of the incident beam used in the experimental PAD system as well as the frequency characteristics of the ultrasound transducer (Fig. 3a). The settings in this paper match the experimental system presented in Section 2.5. Adjusting its parameters based on the imaging results calculated by the model, iterating, and eventually feeding back relevant information to update the system's configuration to provide the best system parameters for the current application scenario (Fig. 3b), thus helping aiding in the optimization of the experimental PAD system (Fig. 3c). Meanwhile, the "learning from computational model" schemes are used to break through the limitations of hardware and human body in the experiments to further improve the imaging performance of the PAD system, such as the optimization of imaging resolution and depth. To train the spread-spectrum model (Fig. 3d), we set the center frequencies of the detectors in the k-Wave acoustic simulation to match the 20 MHz of the experimental system and the optimized target 60 MHz,



**Fig. 4.** Computational imaging results of (a-c) Focused Gaussian beam, (d-f) Bessel beam, and (g-i) collimated Gaussian beam when the power densities are 6 mJ/cm<sup>2</sup>, 12 mJ/cm<sup>2</sup>, and 18 mJ/cm<sup>2</sup>. (j-l) The 3D imaging results of focused Gaussian beams, Bessel beams, and collimated Gaussian beams with a power density of 18 mJ/cm<sup>2</sup>. (m1, m2) Profile intensity along dashed lines 1 and 2 in (a, d, g) with a power density of 6 mJ/cm<sup>2</sup>. (n1, n2) Profile intensity along dashed lines 1 and 2 in (b, e, h) with a power density of 12 mJ/cm<sup>2</sup>. (o1, o2) Profile intensity along dashed lines 1 and 2 in (c, f, i) with a power density of 18 mJ/cm<sup>2</sup>. (p) The a-line signal envelope of three different beams at the same position with a power density of 18 mJ/cm<sup>2</sup>.

respectively, and the data enhancement was achieved by horizontally flipping the 900 pairs of images generated by the computation. To ensure the robustness and generalization of the model, we divided the entire dataset into three parts: training, validation, and testing, where 1400 pairs of images were used for training, 200 pairs of images for validation, and 200 pairs of images for testing. For the deep-enhanced model (Fig. 3e), we set the optical scattering coefficient of the dermis as 1% as the ground truth parameter, and computationally obtained 900 pairs of images corresponding to the dermis under strong and weak scattering, which helps to obtain deeper PAD imaging information, and likewise realized the data enhancement by horizontal flipping, whose number of the training, validation, and testing sets are also 1400, 200, 200, respectively. More broadly, the corresponding parameters in the

computational model are set according to the PAD experimental system and the desired experimental results, and the matched datasets are obtained for network training and adjusted within a certain range to increase the diversity of the data in order to obtain a good generalization performance. Finally, by combining the trained network with the PAD experimental system, the optimized imaging results after network processing can be quickly obtained.

### 3. Results and discussions

This section describes experiments on four factors that affect the imaging performance of the 4D spectral-spatial computational model, as well as the acquisition of the dataset and its application in deep learning.



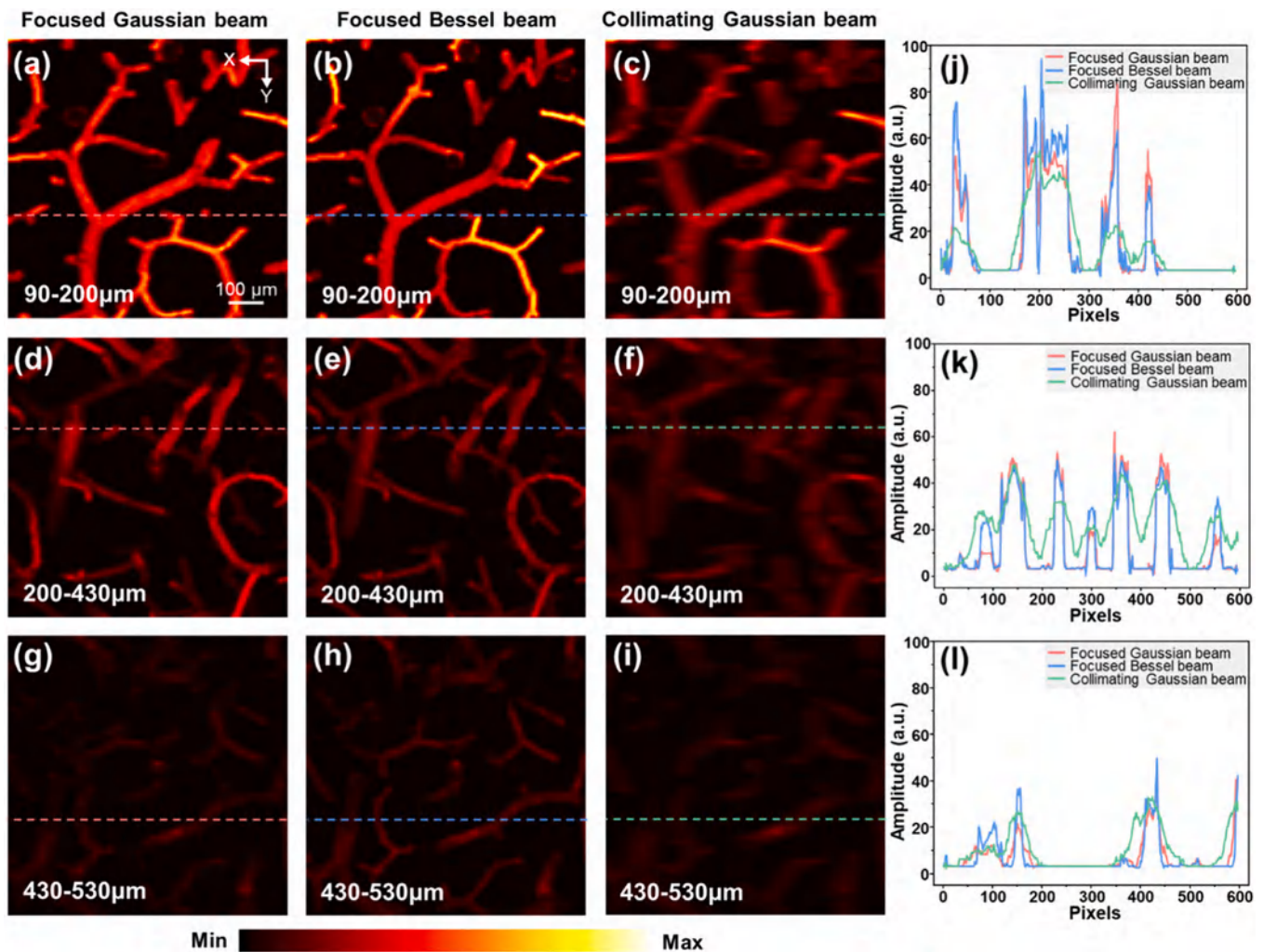


Fig. 5. Maximum intensity projection results of focused Gaussian beam, Bessel beam, and collimated Gaussian beam at different depths with a power density of  $18 \text{ mJ/cm}^2$ . (a-c) Maximum projection image from 90 to 200  $\mu\text{m}$ . (d-f) Maximum projection image from 200 to 430  $\mu\text{m}$ . (g-i) Maximum projection image from 430 to 530  $\mu\text{m}$ . (j) Profile intensity along the dashed lines in (a-c). (k) Profile intensity along the dashed lines in (d-f). (l) Profile intensity along the dashed lines in (g-i).

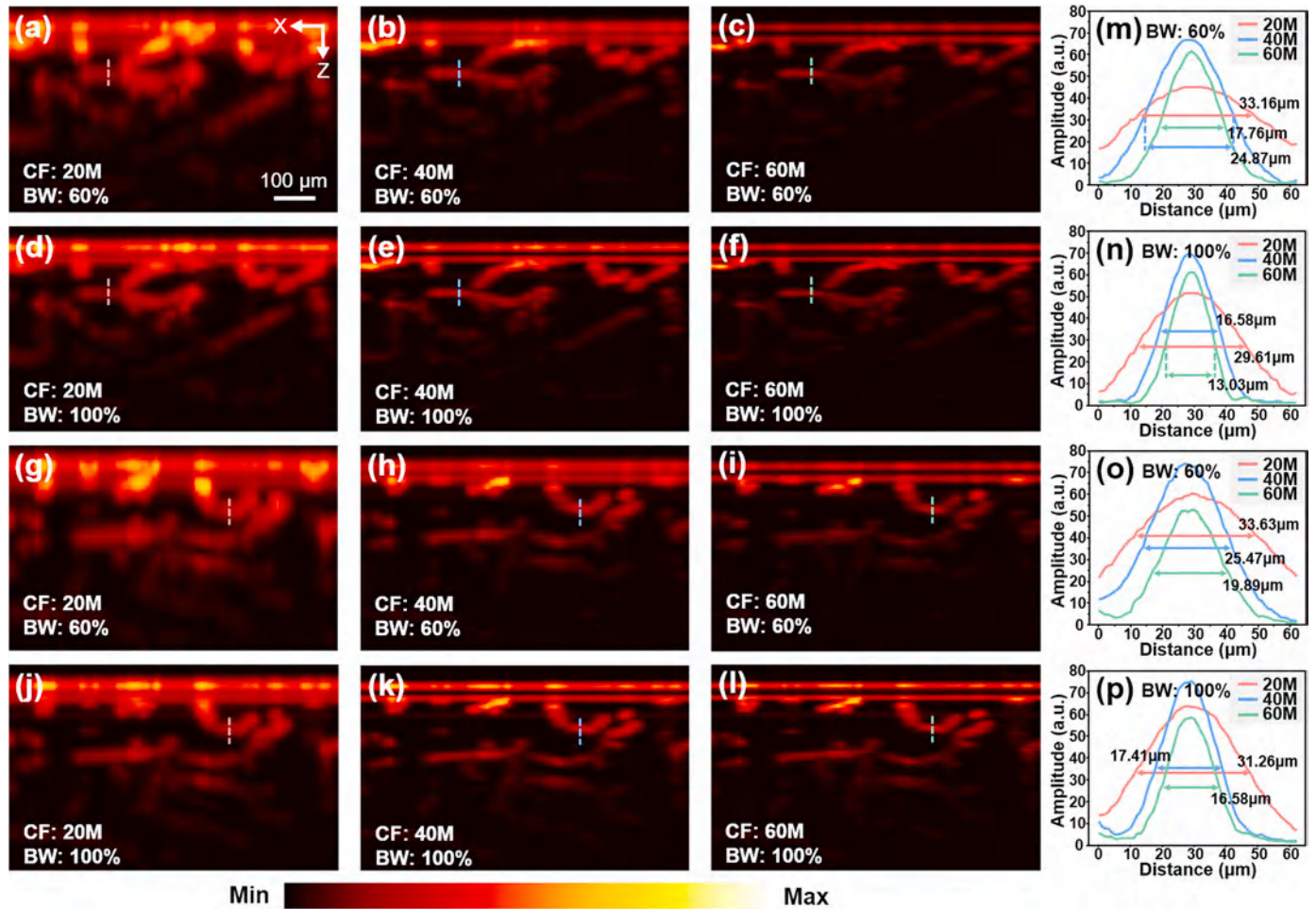
Section 3.1 discusses the influence of the type of incident beam and the power density on the PAD imaging performance. Section 3.2 describes the influence of ultrasound transducers with different center frequencies and bandwidths on the PAD imaging performance. Section 3.3 presents the imaging results of Gaussian beams focused at different depths beneath the skin. Section 3.4 describes the imaging depth under different wavelength beams, demonstrating the multispectral imaging capability of the model. Section 3.5 discusses the feasibility of generating datasets for neural network training using the model.

### 3.1. The influence of the incident beam

The laser parameters used in PAI can greatly affect the imaging results, and since the calculation of the light energy deposition distribution and ultrasonic back propagation of the computational model is performed in a stepwise manner, it can be assumed that the imaging is not affected by the laser pulse width. To validate the influence of beam focusing and energy on imaging performance in the computational model, focused Gaussian, Bessel, and collimated Gaussian beams with varying power densities of incident beams at a wavelength of 532 nm were used for imaging while the optical and acoustic parameters of the model were fixed. The Gaussian and Bessel beams were focused on the boundary between the epidermis and dermis layers of the skin model, and full scanning was performed on the same vascular network for each

type of beam. The ultrasonic transducer with a center frequency of 100 MHz and a bandwidth of 100% was used to receive photoacoustic signals to reduce measurement errors in the computational model. Fig. 4a-c, d-f, and g-i show the maximum intensity projection results of the three types of beams with power densities of  $6 \text{ mJ/cm}^2$ ,  $12 \text{ mJ/cm}^2$ , and  $18 \text{ mJ/cm}^2$ , respectively. Fig. 4j-l show the three-dimensional imaging results of the three types of beams with a power density of  $18 \text{ mJ/cm}^2$ . Fig. 5 shows the maximum intensity projection results at different depths of the three types of beams with a power density of  $18 \text{ mJ/cm}^2$ , in which we could clearly see the difference in imaging depth and resolution between the three beams. The comparisons in Figs. 4m-p and 5 show that, with a fixed power density, Bessel beams and collimated Gaussian beams can achieve greater imaging depth, while Bessel beams have a higher lateral resolution.

In PAI, the intensity of the PA signal is directly proportional to the local optical fluence [52]. Increasing the power density of the incident beam directly increases the optical fluence, which enhances the signal intensity and amplifies small signals in deeper regions, thereby improving the visibility of targets in deeper regions. The imaging resolution is closely related to the size of the optical focus [53]. The collimated Gaussian beams have the lowest imaging quality due to the lack of focus. Both focused Gaussian beams and Bessel beams have excellent spot sizes at the focus point, but Bessel beams achieve better imaging results due to their larger depth of field [54,55]. This result is consistent



**Fig. 6.** Computational imaging results of ultrasonic transducers at different center frequencies and bandwidths. (a-c) Position 1 with center frequencies of 20 MHz, 40 MHz, and 60 MHz and a bandwidth of 60%. (d-f) Position 1 with center frequencies of 20 MHz, 40 MHz, and 60 MHz and a bandwidth of 100%. (g-i) Position 2 with center frequencies of 20 MHz, 40 MHz, 60 MHz, and 60% bandwidth. (j-l) Position 2 with center frequencies of 20 MHz, 40 MHz, 60 MHz, and 100% bandwidth. (m) PA amplitude along the dashed lines in images (a-c). (n) PA amplitude along the dashed lines in images (d-f). (o) PA amplitude along the dashed lines in images (g-i). (p) PA amplitude along the dashed lines in (j-l).

with reality and confirms the reliability of the proposed computational optical model in this work.

### 3.2. The influence of ultrasonic transducer performance

The human skin generates broadband PA signals ranging from a few to hundreds of MHz due to the wide variation in the size of light absorbers. The central frequency and bandwidth of the transducer for detection must be selected based on the size of the target [56]. However, the improper selection of detection bandwidth and central frequency in most PAD studies has resulted in many skin structures being indistinguishable [57].

In this section, under the condition of unchanged optical parameters of the model, in order to reduce the impact of beam focusing on imaging depth, we investigated the effects of changing the center frequency and bandwidth of the ultrasonic transducer on the PAD imaging performance when using a collimated Gaussian beam for illumination of 532 nm. The detection sensitivity of the ultrasonic transducer is higher near the acoustic focus, and its depth of field mainly depends on the center frequency and the numerical aperture of the acoustic lens, typically several hundred micrometers, which is comparable to the depth of field of a Bessel beam. The computational model used in this study employed a bowl-shaped focused ultrasonic transducer consisting of several point detectors on a grid, whose directionality comes from the

spatial average of the pressure field on the detector surface. The depth of field extends along the entire central axis, and it can be assumed that the detection sensitivity of the transducer is uniform within the depth of field range.

In this study, the center frequency of the ultrasonic transducer was chosen as 20 MHz, 40 MHz, and 60 MHz, and the bandwidth was increased from 60% to 100%. The imaging results under different combinations of parameters are shown in Fig. 6a-l, and the profile intensities along the dashed line in the figures are shown in Fig. 6m-p, which show that the axial resolution improves with the increase of the center frequency and bandwidth of the transducer. In addition, due to the correlation between sound attenuation and frequency, as the central frequency of the transducer increases, the visibility of deep blood vessels becomes weaker. These imaging results highlight the benefits of using ultra-broadband ultrasound detectors in PAD.

### 3.3. The influence of laser focusing position

In imaging and diagnostics of PAD, lateral resolution is critical for tissue microstructure studies. Usually, there is a compromise between imaging depth and resolution. Imaging resolution can be improved by beam focusing, but defocusing occurs when the beam is focused at a certain depth under the skin, at which point the imaging resolution deteriorates rapidly. Conversely, if the focusing depth is too shallow, the



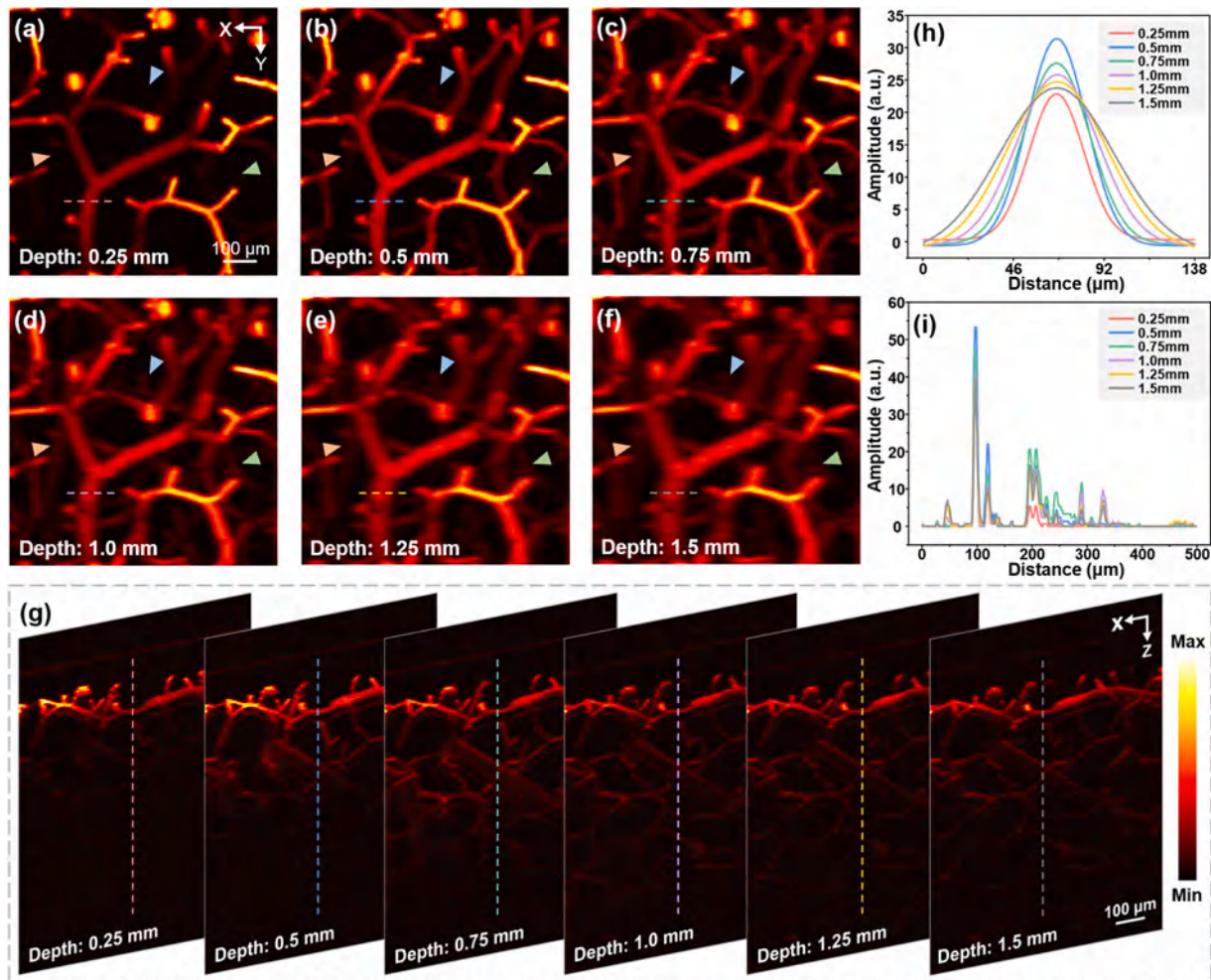


Fig. 7. The imaging results when the Gaussian beam is focused at 0.25 mm, 0.5 mm, 0.75 mm, 1.0 mm, 1.25 mm, and 1.5 mm. (a-f) X-Y maximum projection image. (g) X-Z maximum projection image. (h) PA amplitude along the dashed lines in (a-f). (i) PA amplitude along the dashed lines in (g).

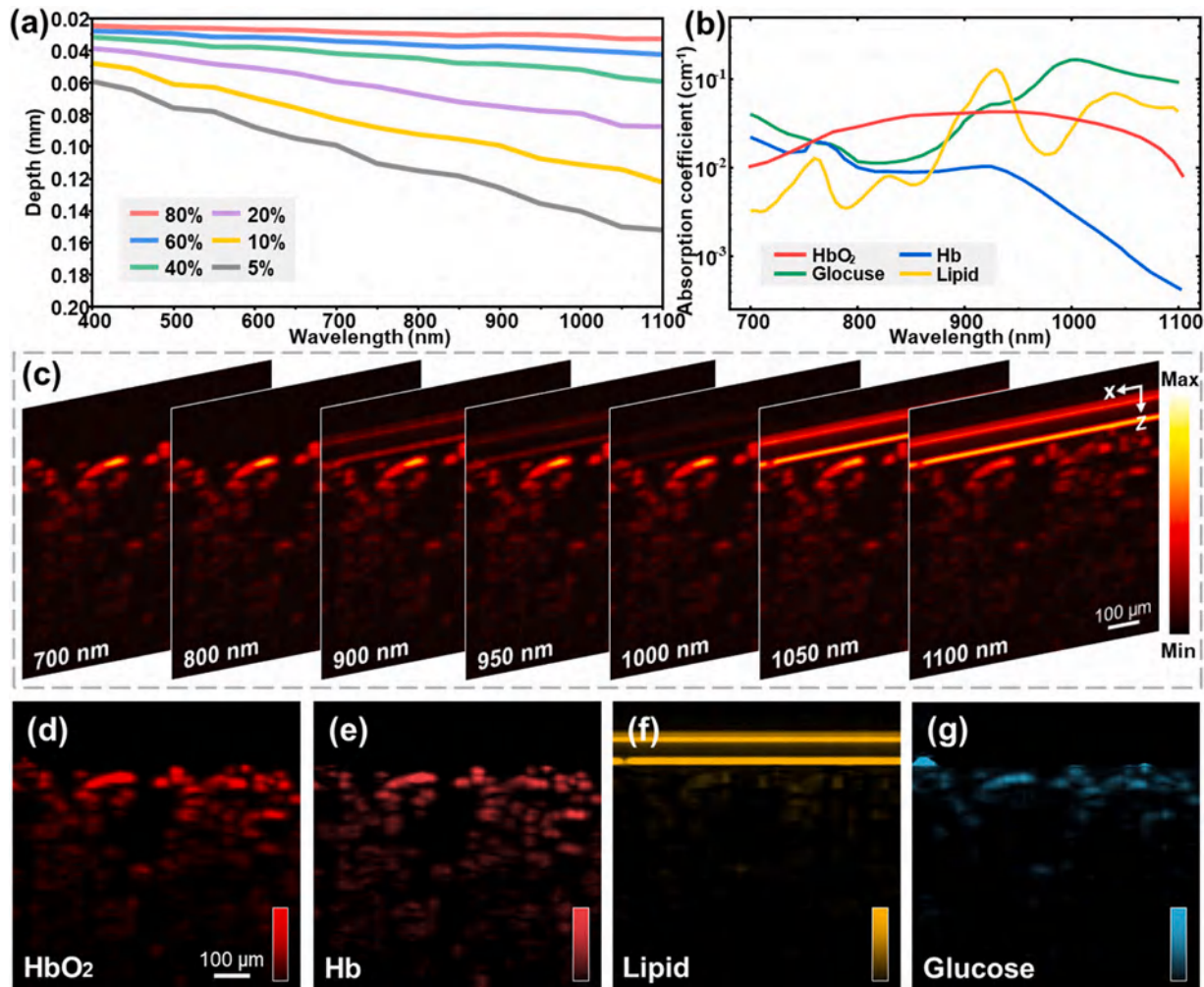
imaging depth may not meet the clinical requirements.

This section discusses the effect of the focusing depth of the incident focused Gaussian beam at 532 nm on imaging. The optical parameters of the skin model remained unchanged, and the incident beam power density was  $18 \text{ mJ/cm}^2$ . The center frequency of the ultrasonic transducer was set to 100 MHz, with a bandwidth of 80%. Full-scan imaging was performed at six different depths where the Gaussian beam was focused at 0.25 mm, 0.5 mm, 0.75 mm, 1.0 mm, 1.25 mm, and 1.5 mm, respectively. Fig. 7a-f show the maximum projection imaging results in the X-Y plane at these six different focusing depths, Fig. 7g shows the maximum projection results in the X-Z plane, and Fig. 7h and i show the normalized profile intensity along the white dashed lines in Fig. 7a-g. As the focusing depth increases, the imaging depth is improved within the maximum range, but the upper vessels are gradually out of focus, leading to a deterioration in the lateral resolution of the imaging. The best imaging results were obtained when the depth of focus was 0.75 mm, which may be because the ultrasound transducer we used in the simulation may have the best focus at this particular depth and the focused Gaussian beam used still maintains a good spot at this depth due to the optical properties of the tissue. However, this observation is not necessarily universally applicable in all cases. Different experimental systems, different samples, and different experimental conditions may affect the determination of the optimal depth of focus.

#### 3.4. Multi-spectral photoacoustic imaging and spectral unmixing

Different skin layers have different optical absorption and scattering properties for different wavelengths, resulting in different penetration depths for each wavelength beam. To further investigate the effect of such wavelengths on the penetration depth, an incident collimated Gaussian beam was used in this section, and computational measurements of imaging depth were made at 50 nm intervals for 15 wavelengths in the wavelength range from 400 nm to 1100 nm. As shown in Fig. 8a, we plotted the fluence rate of incident light at each wavelength up to 80%, 60%, 40%, 20%, 10%, and 5% as a function of penetration depth. As can be seen from the plots, as the wavelength increases, the corresponding penetration depth increases, which is consistent with our expected results. This result suggests that in skin imaging applications, the selection of appropriate wavelengths can realize deeper imaging.

Next, the multi-spectral imaging capability of the proposed model was verified by simulation experiments. Here, we updated the components of the vascular structure in the computational model to a combination of hemoglobin (Hb), oxyhemoglobin ( $\text{HbO}_2$ ), Lipid, and Glucose. The epidermal components only contain lipids, and the variation of optical absorption coefficients of wavelength for the four components is shown in Fig. 8b. B-scan images were acquired at wavelengths ranging from 700 nm to 1100 nm at 25 nm intervals for a total of 17 wavelengths using a collimated Gaussian beam and an ultrasound transducer with a center frequency of 100 MHz and a bandwidth of 100%. Fig. 8c shows the imaging results at seven representative wavelengths: 700 nm,



**Fig. 8.** Multi-spectral imaging and spectral unmixing results. (a)When the fluence rate of incident light reaches 80%, 60%, 40%, 20%, 10%, and 5%, the depth of penetration into the skin varies with the wavelength of the incident light. The fluence rate value is taken from the central column of the output fluence rate grid. (b) Plots of optical absorption coefficients as a function of wavelength for Hb, HbO<sub>2</sub>, Lipid, and Glucose. (c) B-scan images at 700 nm, 800 nm, 900 nm, 950 nm, 1000 nm, 1050 nm, and 1100 nm wavelengths. (d-g) Corresponding spectral unmixing results.

**Table 2**  
Quantitative comparison between ground truth and simulated/U-Net output images on test dataset in the spread-spectrum model. The metrics are represented in the form of mean ± standard deviation.

	MAE	MSE	PSNR	SSIM
Simulated images	0.0543 ± 0.0077	0.1562 ± 0.0040	66.3208 ± 1.0247 dB	0.8115 ± 0.1418
U-Net output images	0.0042 ± 0.0009	0.0001 ± 0.0001	87.4351 ± 1.9440 dB	0.9908 ± 0.0016

**Table 3**  
Quantitative comparison between ground truth and simulated/U-Net output images on test dataset in the depth-enhanced model. The metrics are represented in the form of mean ± standard deviation.

	MAE	MSE	PSNR	SSIM
Simulated images	0.0138 ± 0.0054	0.0019 ± 0.0018	77.0055 ± 3.6093 dB	0.9308 ± 0.0228
U-Net output images	0.0066 ± 0.0029	0.0003 ± 0.0004	84.9128 ± 3.3614 dB	0.9647 ± 0.0239

800 nm, 900 nm, 950 nm, 1000 nm, 1050 nm, and 1100 nm. It can be seen that at the wavelengths of 700 nm and 800 nm, lipids are in the absorption valley, which does not show up clearly in the resulting images, and near 900 nm and 950 nm, HbO<sub>2</sub> is at the absorption peak, which absorbs light more strongly and is shown with higher contrast in the image. The multi-wavelength PA data were utilized to decompose the absorption spectra of the mixed targets by a non-negative constrained least-squares algorithm [58] to obtain unmixed images of the four components, Hb, HbO<sub>2</sub>, Lipid, and Glucose, as shown in Fig. 8d-g. The unmixing results show a decreasing trend with increasing depth because the intensity of the photoacoustic signals produced by each component decreases gradually at the depth of the tissue due to the consideration of the fluence heterogeneity in the optical forward simulation, especially for lower concentrations of glucose. The unmixing results were in close agreement with the modeled components. The multispectral imaging and unmixing capabilities of the skin computational model are useful for the development and optimization of multiwavelength PAD systems, and the analysis of biochemical components in conjunction with spectral unmixing algorithms can help to more accurately diagnose skin diseases.

**3.5. Dataset acquisition and assistance in system optimization**

Deep learning is being widely researched for medical image analysis



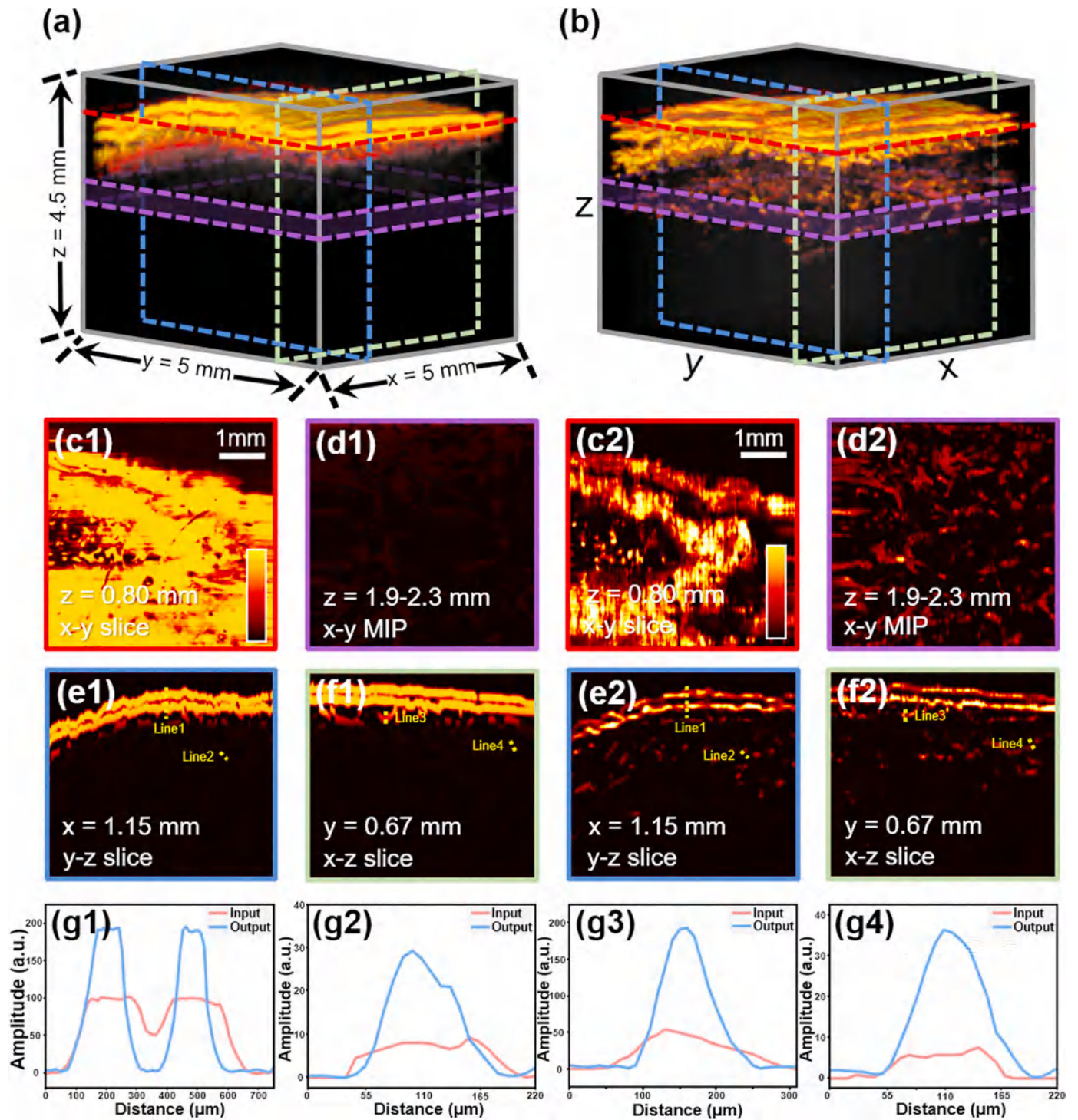


Fig. 9. Network generalization test results. (a) 3D PA image of palm skin. (b) 3D PA image obtained after the spread-spectrum network and the depth-enhanced network processing. (c1, c2) X-Y slice images of red dashed line position in the 3D images. (d1, d2) X-Y maximum intensity projection (MIP) of the purple dashed range in the 3D images. (e1, e2) Y-Z slice images of blue dashed line position in the 3D images. (f1, f2) X-Z slice images of green dashed line position in the 3D images. (g1-g4) Profile intensity along the yellow dashed lines in the slice images.

and processing [59–64]. Most of the deep learning techniques currently used in photoacoustic imaging belong to supervised learning. To train the network, it is essential to establish a matched dataset that pairs ground truth with corresponding measurements. Ideally, datasets should be collected through physical experiments based on the same imaging system. However, in many cases, it is difficult to obtain the ground truth corresponding to the experimental data. In such cases, matching datasets can be obtained through “learning from computational model”

schemes [65–67]. The 4D spectral-spatial computational model in this paper considers wavelength-dependent optical scattering and can calculate the optical and acoustic characteristics of real heterogeneous skin tissue, generating multi-spectral photoacoustic skin imaging datasets under various physical conditions.

Data acquisition requires setting the relevant parameters according to the expected ground truth, including phantom geometry, tissue chromophore content, beam properties, tissue optical properties, sensor

frequency characteristics, etc. This process is shown in Fig. 3a. The optimal system parameters are obtained based on the feedback during the calculation (Fig. 3b), which can assist in optimizing the experimental system (Fig. 3c). Here are two examples illustrating how to implement the “learning from computational model” schemes.

For photoacoustic skin vascular imaging, ultrasound transducers with large center frequencies and wide bandwidths are able to capture fine structural information of blood vessels and obtain higher axial resolution. However, the price of ultrasound transducers is closely related to their performance, and in order to reduce the cost of system construction and data collection, the image spread spectrum can be realized by training the network with simulated data (Fig. 3d). Here, the training of the spread-spectrum network was implemented based on U-Net and the performance of the network on the test set is shown in Table 2. Then it was tested on the experimental PAI data. Again, the experimental system used an ultrasonic detector with a center frequency of 20 MHz, and the scanned PA images of the skin on the back of the author’s hand were fed into the trained spectral spreading network, and the axial resolution of the output image and the completeness of the image information were significantly improved (Fig. 3f).

In addition, since photoacoustic imaging depth is largely affected by optical scattering and optical removal of human skin is difficult to achieve, it is of great significance to train the network to enhance the imaging depth by calculated data. The same network architecture was used to train the depth-enhanced network (Fig. 3e) to help obtain deeper information about PAD imaging. The evaluated parameter values on the test dataset are shown in Table 3. The test results on a priori known leaf sample are shown in Fig. S3. Further, the depth-enhanced network was tested with skin vascularization experimental datas, and the visibility of deeper information in the output image was greatly improved (Fig. 3g).

To examine the generalization of the trained neural network model, here the experimentally acquired 3D skin photoacoustic imaging data of  $5\text{ mm} \times 5\text{ mm} \times 4.5\text{ mm}$  were sequentially fed into the spread spectrum network and depth-enhanced network trained in the above examples with the form of X-Z slices, which have different epidermal shapes and vascular structures. The slices obtained from the network output were reconstructed into 3D images, and Fig. 9 shows the imaging results before and after network processing. It can be clearly seen that the axial resolution of the image is improved and deeper blood vessels are shown. The above experimental results show that the simulated data obtained using the PAD computational model proposed in this paper can obtain results similar to those of actual skin imaging, and can be used to train the network well, thus significantly reducing the training cost.

#### 4. Conclusion

In this paper, we proposed and validated a hybrid computational method of 4D of spectral-spatial imaging for quantitative PAD, which enables structural and molecular computational imaging of blood vessels. The method fully considers wavelength-dependent optical scattering and can calculate the optical and acoustic properties of heterogeneous skin tissue. The computational model integrates two open-source toolboxes: MCmatlab for forward propagation Monte Carlo model of light and calculation of light flux at each grid position, and k-Wave for computation of ultrasonic propagation and reception. By adjusting the types of incident beams, the optical focusing depth, the center frequency, and the bandwidth of the ultrasound transducer, imaging experiments were performed to demonstrate that the computational model can calculate the effects of actual optical and acoustic parameters on photoacoustic imaging and validate its 3D imaging capability. The molecular information obtained for depth- and wavelength-dependent multi-spectral PAD imaging through multi-spectral imaging of Hb, HbO<sub>2</sub>, Lipid, Glucose, and least-squares unmixing of mixed components. In addition to the above experiments, the phantom structure, optical and acoustic parameters can be flexibly set according to the needs to realize the calculated imaging under

various possible variations of multi-scenarios, including skin color, skin thickness, blood vessel number, size and shape, and tissue thickness. The feasibility of simulated datasets generated by computational modeling for neural network training was also demonstrated, helping to solve the major challenge of deep learning techniques in photoacoustic skin imaging that cannot obtain ground truth in many cases, with the potential to further improve the imaging quality of the PAD system through image reconstruction, information processing, and artificial intelligence methods.

In summary, this study provided a comprehensive investigation of the photoacoustic mechanisms in skin tissue, laying a theoretical foundation for the application of photoacoustic imaging detection technology in skin disease diagnosis and treatment, providing a reference for the improvement of treatment protocols, which is crucial for understanding the photoacoustic properties of dermatological diseases and subcutaneous tissues, interpreting and quantifying the diagnostic data as well as evaluating therapeutic and surgical protocols, and also providing a powerful tool for the performance of PAD devices in preclinical and clinical applications. In addition, the PAD computational model proposed in this work can simulate skin tissue, and imaging for specific applications and generate corresponding datasets on a large scale, contributing to the artificial intelligence applications in the PAD field.

Finally, it should be mentioned that, although the “learning from computational model” schemes remove the reliance on large amounts of labeled experimental data, the inconsistency between the image formation model and the actual experimental conditions leads to additional “domain adaptation” challenges. Although in this work we have demonstrated how the model can be utilized to obtain ground truth datasets that are difficult to access in experiments for neural network training, in practice, extending the computational model can be challenging due to a limited understanding of experimental perturbations, such as various noises, aberrations, vibrations, and motion artifacts, and the challenge of not being able to realistically and comprehensively reflect the real experimental system still exists. In future work, we will strive to address the approximate modeling related to these factors and extend the application scenarios of computational models.

#### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Data Availability

Code underlying the results presented in this paper may be obtained from the corresponding author upon reasonable request.

#### Acknowledgements

This work was supported by National Natural Science Foundation of China (62275121, 12204239, 61835015, 1237040502), Youth Foundation of Jiangsu Province (BK20220946), Fundamental Research Funds for the Central Universities (30923011024).

#### Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.pacs.2023.100572](https://doi.org/10.1016/j.pacs.2023.100572).

#### References

- [1] E. Berardesca, H. Maibach, K. Wilhelm. *Non-Invasive Diagnostic Techniques in Clinical Dermatology*, Springer Science & Business Media, Berlin, 2013.
- [2] P. Beard, Biomedical photoacoustic imaging, 602-31, *Interface Focus* 1 (4) (2011), <https://doi.org/10.1098/rsfs.2011.0028>.



- [3] H. Kittler, H. Pehamberger, K. Wolff, Diagnostic accuracy of dermoscopy, *Lancet Oncol.* 3 (3) (2002) 159–165, [https://doi.org/10.1016/S1470-2045\(02\)00679-4](https://doi.org/10.1016/S1470-2045(02)00679-4).
- [4] J. Kim, E. Park, W. Choi, B. Park, K.J. Lee, C. Kim, Clinical photoacoustic/ultrasound imaging: systems and applications, *TM3B.5, Clin. Trans. Biophot.* 01 (2020), <https://doi.org/10.1364/TRANSLATIONAL.2020.TM3B.5>.
- [5] A.J. Deegan, J. Lu, R. Sharma, S.P. Mandell, R.K. Wang, Imaging human skin autograft integration with optical coherence tomography, *Quant. Imaging Med. Surg.* 11 (2) (2021) 784–796, <https://doi.org/10.21037/qims-20-750>.
- [6] L.V. Wang, S. Hu, Photoacoustic tomography: in vivo imaging from organelles to organs, *Science* 335 (6075) (2012) 1458–1462, <https://doi.org/10.1126/science.1216210>.
- [7] H.F. Zhang, K.I. Maslov, L.V. Wang, Automatic algorithm for skin profile detection in photoacoustic microscopy, *J. Biomed. Opt.* 14 (2) (2009), 024050, <https://doi.org/10.1117/1.3122362>.
- [8] I. Steinberg, D.M. Hulandad, O. Vermesh, H.E. Frostig, W.S. Tummers, S. S. Gambhir, Photoacoustic clinical imaging, *Photoacoustics* 14 (2019) 77–98, <https://doi.org/10.1016/j.pacs.2019.05.001>.
- [9] P. Hai, Y. Li, L. Zhu, L. Shmuylovich, L.A. Cornelius, L.V. Wang, Label-free high-throughput photoacoustic tomography of suspected circulating melanoma tumor cells in patients in vivo, *J. Biomed. Opt.* 25 (3) (2020), 036002, <https://doi.org/10.1117/1.JBO.25.3.036002>.
- [10] J. Kim, Y.H. Kim, B. Park, H.M. Seo, C.H. Bang, G.S. Park, Y.M. Park, J.W. Rhie, J. H. Lee, C. Kim, Br.J. Dermatol, Multispectral ex vivo photoacoustic imaging of cutaneous melanoma for better selection of the excision margin, *Br. J. Dermatol.* 179 (3) (2018) 780–782, <https://doi.org/10.1111/bjd.16677>.
- [11] H.G. Ma, Z.Y. Wang, Z.W. Cheng, G. He, T. Feng, C. Zuo, H.X. Qiu, Multiscale confocal photoacoustic dermoscopy to evaluate skin health, *Quant. Imaging Med. Surg.* 12 (5) (2023) 2696–2708, <https://doi.org/10.21037/qims-21-878>.
- [12] J. Aguirre, M. Schwarz, N. Garzorz, M. Omar, A. Buehler, K. Eyerich, V. Ntziachristos, Precision assessment of label-free psoriasis biomarkers with ultra-broadband photoacoustic mesoscopy, *Nat. Biomed. Eng.* 1 (2017) 0068, <https://doi.org/10.1038/s41551-017-0068>.
- [13] H.G. Ma, Z.W. Cheng, Z.Y. Wang, W.Y. Zhang, S.H. Yang, Switchable optical and acoustic resolution photoacoustic dermoscopy dedicated into in vivo biopsy-like of human skin, *Appl. Phys. Lett.* 1116 (7) (2020), 073703, <https://doi.org/10.1063/1.5143155>.
- [14] H.G. Ma, Z.W. Cheng, Z.Y. Wang, H.X. Qiu, T.D. Shen, D. Xing, Y. Gu, S.H. Yang, Quantitative and anatomical imaging of dermal angiopathy by noninvasive photoacoustic microscopic biopsy, *Biomed. Opt. Express* 12 (2021) 6300–6316, <https://doi.org/10.1364/BOE.439625>.
- [15] J. Ahn, J.Y. Kim, W. Choi, C. Kim, High-resolution functional photoacoustic monitoring of vascular dynamics in human fingers, *Photoacoustics* 23 (2021), 100282, <https://doi.org/10.1016/j.pacs.2021.100282>.
- [16] M.A. Mastanduno, S.S. Gambhir, Quantitative photoacoustic image reconstruction improves accuracy in deep tissue structures, *Biomed. Opt. Express* 7 (10) (2016) 3811–3825, <https://doi.org/10.1364/BOE.7.003811>.
- [17] J. Zeng, R. Wang, A. Teng, X. Song, Research on photoacoustic effect of picosecond laser pulse with tissue based on finite element method, *Proc. SPIE* 11844 (2021), 1184416, <https://doi.org/10.1117/12.2601380>.
- [18] S.L. Jacques, Coupling 3D Monte Carlo light transport in optically heterogeneous tissues to photoacoustic signal generation, *Photoacoustics* 2 (2014) 137–142, <https://doi.org/10.1016/j.pacs.2014.09.001>.
- [19] Y.S. Yuan, S. Yan, Q.Q. Fang, Light transport modeling in highly complex tissues using the implicit mesh-based Monte Carlo algorithm, *Biomed. Opt. Express* 12 (2021) 147–161, <https://doi.org/10.1364/BOE.411898>.
- [20] V. Periyasamy, M. Pramanik, Advances in monte carlo simulation for light propagation in tissue, *IEEE Rev. Biomed. Eng.* 10 (2017) 122–135, <https://doi.org/10.1109/RBME.2017.2739801>.
- [21] Y.Q. Tang, J.J. Yao, 3D Monte Carlo simulation of light distribution in mouse brain in quantitative photoacoustic computed tomography, *Quant. Imaging Med. Surg.* 11 (3) (2021) 1046–1059, <https://doi.org/10.21037/qims-20-815>.
- [22] X. Shu, W.Z. Liu, H.F. Zhang, Monte Carlo investigation on quantifying the retinal pigment epithelium melanin concentration by photoacoustic ophthalmoscopy, *J. Biomed. Opt.* 20 (10) (2020), 106005, <https://doi.org/10.1117/1.JBO.20.10.106005>.
- [23] A.N. Bashkatov, E.A. Genina, V.V. Tuchin, G.B. Altschuler, I.V. Yaroslavsky, Monte Carlo study of skin optical clearing to enhance light penetration in the tissue: implications for photodynamic therapy of acne vulgaris, *Adv. Laser Technol.* 2007 (2008), 702209, <https://doi.org/10.1117/12.803909>.
- [24] Y.E. Hajji, E.H.E. Rhaleb, Melanin effect on light beam intensity distribution in skin as a function of wavelength and depth from 200 to 1000 nm using Monte Carlo simulation, *J. Quant. Spectrosc. Radiat. Transf.* 295 (2023), 108411, <https://doi.org/10.1016/j.jqsrt.2022.108411>.
- [25] T. Maeda, N. Arakawa, M. Takahashi, Monte Carlo simulation of spectral reflectance using a multilayered skin tissue model, *Opt. Rev.* 17 (2010) 223–229, <https://doi.org/10.1007/s10043-010-0040-5>.
- [26] B.E. Treeby, B.T. Cox, k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields, *J. Biomed. Opt.* 15 (2) (2010), 021314, <https://doi.org/10.1117/1.3360308>.
- [27] X.L. Song, G.Y. Chen, A.J. Zhao, X.Y. Liu, J.H. Zeng, Virtual optical-resolution photoacoustic microscopy using the k-Wave method, *Appl. Opt.* 60 (2021) 11241–11246, <https://doi.org/10.1364/AO.444106>.
- [28] A. Sharma, M. Pramanik, Convolutional neural network for resolution enhancement and noise reduction in acoustic resolution photoacoustic microscopy, *Biomed. Opt. Express* 11 (2020) 6826–6839, <https://doi.org/10.1364/BOE.411257>.
- [29] N. Akhlaghi, T.J. Pfefer, K.A. Wear, B.S. Garra, W.C. Vogt, Multidomain computational modeling of photoacoustic imaging: verification, validation, and image quality prediction, *Jr. Biom. Opt.* 24 (12) (2019), 121910, <https://doi.org/10.1117/1.JBO.24.12.121910>.
- [30] M. Heijblom, D. Piras, E. Maartens, E.J.J. Huisman, F.M. van den Engh, J. M. Klaase, W. Steenbergen, S. Manohar, Appearance of breast cysts in planar geometry photoacoustic mammography using 1064-nm excitation, *126009-126009, J. Biomed. Opt.* 18 (12) (2013), <https://doi.org/10.1117/1.JBO.18.12.126009>.
- [31] P. Valeriya, K. Daria, K. Aleksandr, K. Mikhail, Combined Monte Carlo and k-wave simulations for reconstruction of blood oxygen saturation in optoacoustics: a pilot study, *J. Biomed. Photonics Eng.* 8 (4) (2022) 40511, <https://doi.org/10.18287/JBPE22.08.040511>.
- [32] B.A. Kaplan, J. Buchmann, S. Prohaska, J. Laufer, Monte-Carlo-based inversion scheme for 3D quantitative photoacoustic tomography, *Photons Ultrasound: Imaging Sens.* 10064 (2017) 802–814, <https://doi.org/10.1117/12.2251945>.
- [33] T. Feng, Y.X. Ge, Y.J. Xie, W.Y. Xie, C.C. Liu, L. Li, D. Ta, Q. Jiang, Q. Cheng, Detection of collagen by multi-wavelength photoacoustic analysis as a biomarker for bone health assessment, *Photoacoustics* 24 (2021), 100296, <https://doi.org/10.1016/j.pacs.2021.100296>.
- [34] B. Park, C.H. Bang, C. Lee, J.H. Han, W. Choi, J. Kim, G.S. Park, J.W. Rhie, J.H. Lee, C. Kim, 3D wide-field multispectral photoacoustic imaging of human melanomas in vivo: a pilot study, *J. Eur. Acad. Dermatol. Venereol.* 35 (2021) 669–676, <https://doi.org/10.1111/jdv.16985>.
- [35] N. Cao, Y.H. Li, R.Y. Zhang, S.B. Liu, Y.P. Xiong, H. Cao, Theoretical analysis of photoacoustic effects in a multilayered skin tissue model, *AIP Adv.* 13 (3) (2023), 035007, <https://doi.org/10.1063/5.0136208>.
- [36] I.R.M. Barnard, P. Tierney, C.L. Campbell, L. McMillan, H. Moseley, E. Eadie, C.T. A. Brown, K. Wood, Quantifying direct DNA damage in the basal layer of skin exposed to UV radiation from sunbeds, *Photochem. Photobiol.* 94 (2018) 1017–1025, <https://doi.org/10.1111/php.12935>.
- [37] A.E. Karsten, J.E. Smit, Modeling and verification of melanin concentration on human skin type, *Photochem. Photobiol.* 88 (2012) 469–474, <https://doi.org/10.1111/j.1751-1097.2011.01044.x>.
- [38] T. Lister, P.A. Wright, P.H. Chappell, Optical properties of human skin, *J. Biomed. Opt.* 17 (2012) 0909011, <https://doi.org/10.1117/1.JBO.17.9.0909011>.
- [39] G. Tetteh, V. Eftremov, N.D. Forkert, M. Schneider, J. Kirschke, B. Weber, Deepvesselet: vessel segmentation, centerline prediction, and bifurcation detection in 3d angiographic volumes, *Front. Neurosci.* 14 (2020), <https://doi.org/10.3389/fnins.2020.592352>.
- [40] I.V. Meglinski, S.J. Matcher, Quantitative assessment of skin layers absorption and skin reflectance spectra simulation in the visible and near-infrared spectral regions, *Physiol. Meas.* 23 (2002) 741–753, <https://doi.org/10.1088/0967-3334/23/4/312>.
- [41] M. Sand, T. Gambichler, G. Moussa, F.G. Bechara, D. Sand, P. Altmeyer, K. Hoffmann, Evaluation of the epidermal refractive index measured by optical coherence tomography, *Ski. Res. Technol.* 12 (2006) 114–118, <https://doi.org/10.1111/j.0909-752X.2006.00144.x>.
- [42] G. Altschuler, M. Smirnov, I. Yaroslavsky, Lattice of optical islets: a novel treatment modality in photomedicine, *J. Phys. D: Appl. Phys.* 38 (2005) 2732–2747, <https://doi.org/10.1088/0022-3727/38/15/027>.
- [43] I.V. Meglinski, A.N. Bashkatov, E.A. Genina, D. Yu. Churmakov, V.V. Tuchin, Study of the possibility of increasing the probing depth by the method of reflection confocal microscopy upon immersion clearing of nearsurface human skin layers, *Quantum Electron.* 32 (2002) 875–882, <https://doi.org/10.1070/QE2002v032n10ABEH002309>.
- [44] R.C. Smith, K.S. Baker, Optical properties of the clearest natural waters (200–800 nm), *Appl. Opt.* 20 (1981) 177–184, <https://doi.org/10.1364/AO.20.000177>.
- [45] K.F. Palmer, D. Williams, Optical properties of water in the near infrared, *J. Opt. Soc. Am. A* 64 (1974) 1107–1110, <https://doi.org/10.1364/JOSA.64.001107>.
- [46] S.L. Jacques, Optical properties of biological tissues: a review, *Phys. Med. Biol.* 58 (2013) 5007, <https://doi.org/10.1088/0031-9155/58/14/5007>.
- [47] D. Marti, R.N. Aasbjerg, P.E. Andersen, A.K. Hansen, MCmatlab: an open-source, user-friendly, MATLAB-integrated three-dimensional Monte Carlo light transport solver with heat diffusion and tissue damage, *J. Biomed. Opt.* 23 (12) (2018), 121622, <https://doi.org/10.1117/1.JBO.23.12.121622>.
- [48] D.K. Yao, C. Zhang, K.I. Maslov, L.V. Wang, Photoacoustic measurement of the Grüneisen parameter of tissue, *Jr. Biom. Opt.* 19 (1) (2014), 017007, <https://doi.org/10.1117/1.JBO.19.1.017007>.
- [49] B.E. Treeby, B.T. Cox, Modeling power law absorption and dispersion for acoustic propagation using the fractional Laplacian, *J. Acoust. Soc. Am.* 127 (5) (2010) 2741–2748, <https://doi.org/10.1121/1.3377056>.
- [50] K.K. Shung, R.A. Sigelmann, J.M. Reid, Scattering of ultrasound by blood, *IEEE Trans. Biomed. Eng.* 6 (1976) 460–467, <https://doi.org/10.1109/TBME.1976.324604>.
- [51] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.* 9351 (2015) 234–241, [https://doi.org/10.1007/978-3-319-24574-4\\_28](https://doi.org/10.1007/978-3-319-24574-4_28).
- [52] K. Maslov, H. Zhang, S. Hu, L.V. Wang, Optical-resolution photoacoustic microscopy for in vivo imaging of single capillaries, *Opt. Lett.* 33 (2008) 929–931, <https://doi.org/10.1364/OL.33.000929>.
- [53] L.V. Wang, Multiscale photoacoustic microscopy and computed tomography, *Nat. Photon* 3 (2009) 503–509, <https://doi.org/10.1038/nphoton.2009.157>.

- [54] B. Park, Reflection-mode switchable subwavelength Bessel-beam and Gaussian-beam photoacoustic microscopy in vivo, *J. Biophotonics* 12 (2019), e201800215, <https://doi.org/10.1002/jbio.201800215>.
- [55] B. Jiang, X. Yang, Q. Luo, Reflection-mode Bessel-beam photoacoustic microscopy for in vivo imaging of cerebral capillaries, *Opt. Express* 24 (2016) 20167–20176, <https://doi.org/10.1364/OE.24.020167>.
- [56] E.M. Strohm, E.S. Berndt, M.C. Kolios, High frequency label-free photoacoustic microscopy of single cells, *Photoacoustics* 1 (3–4) (2013) 49–53, <https://doi.org/10.1016/j.pacs.2013.08.003>.
- [57] M. Schwarz, D. Soliman, M. Omar, A. Buehler, S.V. Ovsepian, J. Aguirre, V. Ntziachristos, Optoacoustic dermoscopy of the human skin: tuning excitation energy for optimal detection bandwidth with fast and deep imaging in vivo, *IEEE Tr. Med. Imag.* 36 (6) (2017) 1287–1296, <https://doi.org/10.1109/TMI.2017.2664142>.
- [58] T. Feng, Y. J. Xie, W.Y. Xie, Y.N. Chen, P. Wang, L. Li, J. Han, D. Ta, L.M. Cheng, Q. Cheng, Characterization of multi-biomarkers for bone health assessment based on photoacoustic physicochemical analysis method, *Photoacoustics* 25 (2022), 100320, <https://doi.org/10.1016/j.pacs.2021.100320>.
- [59] E.M.A. Anas, H.K. Zhang, J. Kang, E. Boctor, Enabling fast and high quality LED photoacoustic imaging: a recurrent neural networks based approach, *Biomed. Opt. Express* 9 (8) (2018) 3852–3866, <https://doi.org/10.1364/BOE.9.003852>.
- [60] P. Farnia, E. Najafzadeh, A. Hariri, S.N. Lavasani, B. Makkiabadi, A. Ahmadian, J. V. Jokerst, Dictionary learning technique enhances signal in LED-based photoacoustic imaging, *Biomed. Opt. Express* 11 (5) (2020) 2533–2547, <https://doi.org/10.1364/BOE.387364>.
- [61] A. Hariri, K. Alipour, Y. Mantri, J.P. Schulze, J.V. Jokerst, Deep learning improves contrast in low-fluence photoacoustic imaging, *Biomed. Opt. Express* 11 (6) (2020) 3360–3373, <https://doi.org/10.1364/BOE.395683>.
- [62] S. Guan, A.A. Khan, S. Sikdar, P.V. Chitnis, Fully dense unet for 2-D sparse photoacoustic tomography artifact removal, *IEEE J. Biomed. Health Inform.* 24 (2) (2020) 568–576, <https://doi.org/10.1109/JBHI.2019.2912935>.
- [63] C.C. Yang, H.R. Lan, F. Gao, F. Gao, Review of deep learning for photoacoustic imaging, *Photoacoustics* 21 (2021), 100215, <https://doi.org/10.1016/j.pacs.2020.100215>.
- [64] Z.L. Wu, I. Kang, Y.D. Yao, Y. Jiang, J.J. Deng, J. Klug, S. Vogt, G. Barbastathis, Three-dimensional nanoscale reduced-angle ptycho-tomographic imaging with deep learning (RAPID), *eLight* 3 (2023) 7, <https://doi.org/10.1186/s43593-022-00037-9>.
- [65] C. Zuo, J.M. Qian, S.J. Feng, W. Yin, Y.X. Li, P.F. Fan, J. Han, K.M. Qian, Q. Chen, Deep learning in optical metrology: a review, *Light Sci. Appl.* 11 (2022) 39, <https://doi.org/10.1038/s41377-022-00714-x>.
- [66] Y. Fan, J.J. Li, L.P. Lu, J.S. Sun, Y. Hu, J.L. Zhang, Z.S. Li, Q. Shen, B.W. Wang, R. N. Zhang, Q. Chen, C. Zuo, Smart computational light microscopes (SCLMs) of smart computational imaging laboratory (SCILab), *PhotonIX* 2 (2021) 19, <https://doi.org/10.1186/s43074-021-00040-2>.
- [67] J.M. Qian, Y. Cao, Y. Bi, H.J. Wu, Y.T. Liu, Q. Chen, C. Zuo, Structured illumination microscopy based on principal component analysis, *eLight* 3 (2023) 4, <https://doi.org/10.1186/s43593-022-00035-x>.



**Haixia Qiu** is the director of the First Medical Center Laser medicine Department, deputy chief physician, master tutor. Good at strong laser treatment of pigment, vascular and proliferative diseases, fatigue marks, etc. She is currently the executive editorial Board member of the Chinese Journal of Laser Medicine, the Standing Committee member of the Laser Medicine Branch of the Beijing Medical Association, the member of the board of directors of the Chinese Optical Society, and national natural science evaluation expert.



**Ying Gu** is an academician of the Chinese Academy of Sciences, chief physician, professor, director of Laser Medicine Department of the First Medical Center of PLA General Hospital, Director of Laser Medicine Center of Hainan Hospital. Permanent member of the International Federation of Laser Medicine, Chairman of the Chinese Medical Association Laser Medicine Branch, chairman of the Chinese Optical Society. She presided over the development of China's first laser medicine clinical technology operating standards and diagnosis and treatment guidelines.



**Qian Chen** as a leading expert in the National Key Discipline of "Optical Engineering" at Nanjing University of Science and Technology. As the primary contributor, he has won a second-class State Technological Invention Award, a second-class State Scientific and Technological Progress Award and five first-class provincial and ministerial-level science and technology awards. As the first inventor, he has obtained 74 granted invention patents, 16 PCT international patents, and 6 U.S. patents. He has authored three books and 374 SCI papers, among which 27 have been featured on the cover. Currently, he serves as a Fellow and Executive Director of the Chinese Society of Optical Engineering and Executive Director of the Chinese Institute of Electronics.



**Chao Zuo** is a professor in optical engineering, Nanjing University of Science and Technology (NJUST), China. He leads the Smart Computational Imaging Laboratory (SCILab: [www.scilaboratory.com](http://www.scilaboratory.com)) at the School of Electronic and Optical Engineering, NJUST. He has long been engaged in the development of novel Computational Optical Imaging and Measurement technologies, with a focus on Phase Measuring Imaging Metrology such as Holographic Interferometric Microscopy, Noninterferometric Quantitative Phase Imaging (QPI), Fringe Projection Profilometry (FPP), and Structured Illumination Microscopy (SIM). He has authored > 200 peer-reviewed publications in prestigious journals with over 11,000 citations.



**Haigang Ma** is an associate Professor, Nanjing University of Science and Technology. In the past five years, he has published more than 20 SCI papers on photoacoustic imaging systems, detection methods and application research. Meanwhile, actively participated in promoting the instrumenting of photoacoustic imaging technology, applied for 25 national invention patents in China, and developed the first photoacoustic microscopic imaging instrument applied to clinical skin detection in China, and obtained the registration certificate for medical device of the People's Republic of China.



**Yang Gao** is a master student from Nanjing University of Science and Technology. He is now in the third year of his master's degree and his current research focuses on applications of photoacoustic imaging in biomedicine.



**Ting Feng** received her bachelor's degree, master's degree and Ph.D. degree from Nanjing University in 2010, 2012 and 2016, respectively. She is currently working at Fudan university in China. She was the visiting scholar at the University of Michigan in 2018 and 2019, and she was the joint-Ph.D. student at the University of Michigan in 2013–2015. Her current research interest includes photoacoustic imaging and measurements. A major part of her research is clinical application of photoacoustic techniques for bone health assessment.

# Deep Learning-Enabled Pixel-Super-Resolved Quantitative Phase Microscopy from Single-Shot Aliased Intensity Measurement

Jie Zhou, Yanbo Jin, Linpeng Lu, Shun Zhou, Habib Ullah, Jiasong Sun, Qian Chen, Ran Ye,\* Jiaji Li,\* and Chao Zuo\*

A new technique of deep learning-based pixel-super-resolved quantitative phase microscopy (DL-SRQPI) is proposed, achieving rapid wide-field high-resolution and high-throughput quantitative phase imaging (QPI) from single-shot low-resolution intensity measurement. By training a neural network with sufficiently paired low-resolution intensity and high-resolution phase data, the network is empowered with the capability to robustly reconstruct high-quality phase information from a single frame of an aliased intensity image. As a graphics processing units-accelerated computational method with minimal data requirement, DL-SRQPI is well-suited for live-cell imaging and accomplishes high-throughput long-term dynamic phase reconstruction. The effectiveness and feasibility of DL-SRQPI have been significantly demonstrated by comparing it with other traditional and learning-based phase retrieval methods. The proposed method has been successfully implemented into the quantitative phase reconstruction of biological samples under bright-field microscopes, overcoming pixel aliasing and improving the spatial-bandwidth product significantly. The generalization ability of DL-SRQPI is illustrated by phase reconstruction of Henrietta Lacks cells at various defocus distances and illumination patterns, and its high-throughput anti-aliased phase imaging performance is further experimentally validated. Given its capability of achieving pixel super-resolved QPI from single-shot intensity measurement over conventional bright-field microscope hardware, the proposed approach is expected to be widely adopted in life science and biomedical workflows.

## 1. Introduction

Optical microscopy has undergone continuous development since its invention in the 17th century and has gradually become an essential tool for visualizing cellular and subcellular features of biological samples, driven by the increasing demand for biomedical research.<sup>[1]</sup> However, generating sufficient contrast in most biological samples is challenging due to their low absorption or weak-scattering characteristic.<sup>[1,2]</sup> To obtain their precise and detailed phase information, extensive research has been conducted for decades. Fluorescence microscopy is one of the most far-reaching developments for weak absorption object visualization. It labels the specimen with fluorescent molecules to provide targeted morphological and biochemical information. With the emergence of new fluorescent molecular probes and novel optical imaging techniques, advanced super-resolution fluorescence microscopy further enables super-resolution subcellular detail observation at the nano-scale well beyond the diffraction limit, such as structured illumination microscopy

J. Zhou, Y. Jin, L. Lu, S. Zhou, H. Ullah, J. Sun, Q. Chen, R. Ye, J. Li, C. Zuo  
Smart Computational Imaging (SCI) Laboratory  
Nanjing University of Science and Technology  
Nanjing, Jiangsu Province 210094, China  
E-mail: ran.ye@nju.edu.cn; jiajili@njust.edu.cn; zuochao@njust.edu.cn

J. Zhou, Y. Jin, L. Lu, S. Zhou, H. Ullah, J. Sun, Q. Chen, J. Li, C. Zuo  
Smart Computational Imaging Research Institute (SCIRI) of Nanjing  
University of Science and Technology  
Nanjing, Jiangsu Province 210019, China

J. Zhou, Y. Jin, L. Lu, S. Zhou, H. Ullah, J. Sun, Q. Chen, J. Li, C. Zuo  
Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense  
Nanjing University of Science and Technology  
Nanjing, Jiangsu Province 210094, China

R. Ye  
School of Computer and Electronic Information  
Nanjing Normal University  
Nanjing, Jiangsu Province 210023, China

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/lpor.202300488>

DOI: 10.1002/lpor.202300488



(SIM),<sup>[3]</sup> stimulated emission depletion microscopy (STED),<sup>[4,5]</sup> photo-activated localization microscopy (PALM), and stochastic optical reconstruction microscopy (STORM).<sup>[6]</sup> However, the utilization of exogenous agents may introduce photo-toxicity and photo-bleaching issues, which hinder the long-term imaging of living cells. Furthermore, the use of fluorescent dyes and proteins as bio-markers inevitably limits certain non-fluorescent applications where biological samples cannot be easily tagged with fluorescent markers.<sup>[7,8]</sup>

In recent years, the technique of computational microscopy, including interferometric<sup>[9–11]</sup> and non-interferometric<sup>[12–14]</sup> manners for both quantitative phase imaging (QPI)<sup>[15–19]</sup> and 3D refractive index (3D RI),<sup>[20,21]</sup> has been proved to be an invaluable tool regarding its distinctive capability to quantify the phase delay of unlabeled biological specimens in a non-destructive way. As two representative QPI approaches, transport of intensity equation (TIE)<sup>[22]</sup> and Fourier ptychographic microscopy (FPM)<sup>[23]</sup> have gained wide attention in the application of biomedicine. With a simple optical implementation of an off-the-shelf bright-field microscope, the phase distribution of specimen can be simply reconstructed by TIE using intensity measurements at multiple axially displaced planes. Nevertheless, the achievable imaging resolution of TIE is restricted to the incoherent diffraction limit under partially coherent illumination, and the spatial bandwidth product (SBP) of TIE is fundamentally restrained by the optical system, resulting in a trade-off between imaging resolution and field-of-view (FOV).<sup>[24,25]</sup> FPM is a recently developed computational imaging technique that could circumvent the imaging resolution-FOV trade-off and improve the throughput of the imaging system.<sup>[23,26]</sup> FPM maintains high imaging resolution and wide FOV simultaneously by stitching together a series of variously illuminated low-resolution but large-FOV intensity images in Fourier space. However, FPM requires a large amount of data redundancy, which leads to a cumbersome data acquisition process. Additionally, the iterative strategy used by FPM limits its recovery efficiency, preventing its application in high-speed cell imaging.

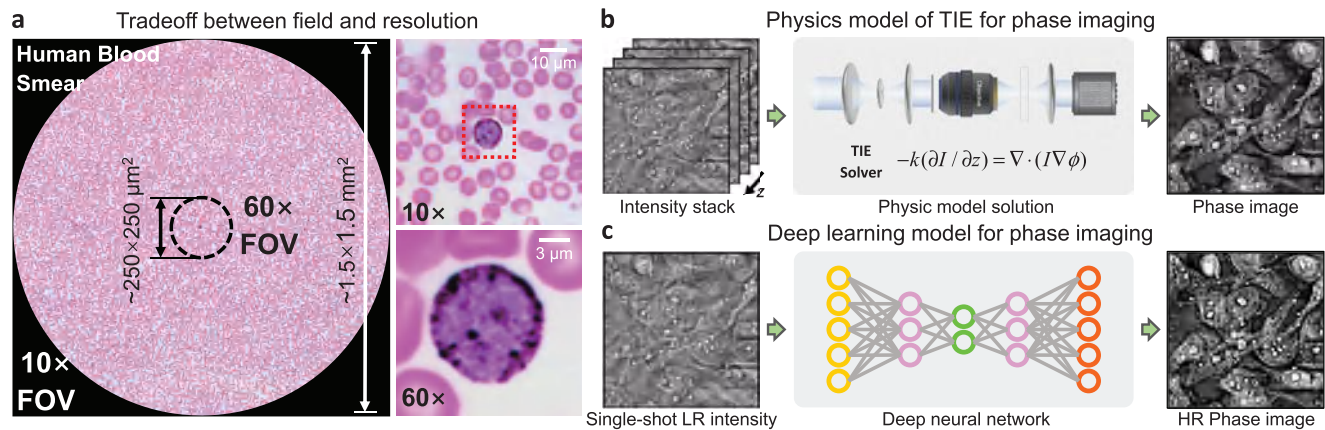
On the other hand, high-throughput QPI faces another major obstacle posed by pixel-aliasing.<sup>[27]</sup> In optical systems, detectors are used to collect intensity information and are typically designed with large pixel sizes to accommodate high photosensitivity and large FOVs for high-throughput imaging. However, large pixel sizes may lead to inadequate sampling or digitization of the transmitted intensity, resulting in low pixel resolution and even leading to the infamous pixel-aliasing/undersampling problem. Although deploying magnification camera adapters or using image sensors with smaller pixel sizes could mitigate the pixel-aliasing problem, it comes at the cost of the FOV. Therefore, this trade-off between pixel resolution and FOV leads to sub-optimal use of SBP of the imaging system. Several QPI techniques with anti-aliased ability have been proposed. For instance, the pixel-aliasing in differential phase contrast (DPC) could be alleviated by the iterative de-multiplexing algorithm. However, its efficacy is still restrained by the elaborate illumination scheme, the requirement of multiple intensity images and the iterative strategy.

Benefiting from the accelerating development in computer science and technology, coupled with exponential growth in processing power, the past few years have witnessed rapid progress

in deep learning, where high-dimensional representations can be learned directly from captured data based on neural networks. With its unique data-driven methodology, deep learning has solved many tasks in computer vision and computer-aided diagnosis with unprecedented performance.<sup>[28]</sup> In the field of computational microscopy, deep learning has led to rapid growth in algorithms and methods for solving various ill-posed inverse problems, transcending the limitations of traditional microscopy.<sup>[29]</sup> For example, deep learning enables super-resolution imaging and reveals microscopic biological details with higher precision.<sup>[30]</sup> It has also been proved that a conventional microscope aided by deep learning could even enable the observation of nano-scale subcellular details well beyond the diffraction limit, reaching the image resolution of STED.<sup>[31]</sup> Deep learning also realizes cross-modality imaging of biomedical samples, such as the digital staining technique that accommodates the generation of quantitative phase images for virtual histological staining, therefore circumventing the procedure of laborious and time-consuming sample staining.<sup>[32]</sup> Thanks to its powerful non-linear ability, deep learning has also been utilized to enhance the phase information acquisition capability of conventional computational imaging techniques by constructing precise mapping relationships between intensity and phase distributions. With a well-trained neural network, FPM can reduce the number of required intensity images from hundreds to five, eliminating the tedious image acquisition process and time-consuming iterations, while maintaining the quality of the reconstructed phase images.<sup>[29,33]</sup> However, these methods still require multiple input images for phase recovery. The end-to-end capability of deep learning implies that the data redundancy requirement for phase retrieval can be further minimized to single frame. Based on Gerchberg and Saxton (GS)'s iterative phase retrieval algorithm, a deep neural network has successfully obtained accurate amplitude and phase information from a single coaxial hologram amidst the interference of twin images and object artifacts.<sup>[34]</sup> Another deep learning-assisted method achieves TIE-based phase retrieval from a single intensity image.<sup>[35]</sup> Nevertheless, though the data efficiency has been improved, the image resolution is still limited, and phase recovery for a large population of cells remains to be investigated. Besides, the problem of pixel-aliasing requires further exploration. Consequently, high-throughput quantitative phase imaging with both wide FOV and pixel super-resolution from single-frame intensity image in bright-field optical implementation has not been proposed yet.

In this work, we present a novel quantitative phase imaging technique, termed deep learning-based single-frame super-resolution quantitative phase imaging (DL-SRQPI). Our method combines deep learning with quantitative phase imaging and achieves high-throughput, high-accuracy phase retrieval in a computational manner without any additional hardware design. After proper training, a neural network identifies the mapping relationship between low-resolution intensity image and high-resolution phase image, with which DL-SRQPI alleviates the pixel-aliasing problem and improves the space-bandwidth product since the inherent large FOV of the low-resolution intensity image is exploited. DL-SRQPI maximizes the data efficiency by reducing the intensity image redundancy requirement to only one frame, and the phase reconstruction speed is greatly accelerated by utilizing the graphics processing unit (GPU). The





**Figure 1.** Illustration of the trade-off between field of view and resolution, and comparison of TIE and DL-SRQPI phase retrieval methods. a) Comparison of resolution and field of view of human blood smear microscopic images under 10x and 60x objectives. b) Standard TIE phase retrieval workflow using an axial defocus intensity image stack as input to solve the TIE equation and obtain phase images. c) DL-SRQPI phase retrieval workflow using a single-frame defocused intensity image as input of a well-trained neural network, and outputs a super-resolved phase image.

effectiveness and feasibility of DL-SRQPI has been illustrated by the comparison with other traditional or network-aided phase retrieval methods, and the robustness of DL-SRQPI is also proved by the phase reconstruction of intensity images at various defocus distances and illumination conditions. To demonstrate its strong capability, we use DL-SRQPI to rapidly convert hundreds of frames of  $512 \times 512$  pixels simulated intensity images into the corresponding  $2048 \times 2048$  pixels phase images with high accuracy. We further validate the capability of DL-SRQPI with experimentally acquired intensity images. For a  $647 \times 490$  pixels intensity image obtained by an off-the-shelf bright-field optical microscope, DL-SRQPI precisely retrieves its phase result at a resolution of  $2588 \times 1960$  pixels while maintaining the original FOV, revealing abundant subcellular details that are once embedded in the aliased pixels. With the large-SBP phase reconstruction capability of DL-SRQPI, we provide long-term high-throughput time-lapse videos of Henrietta Lacks (HeLa) cells undergoing division. These superior performances indicate that the proposed DL-SRQPI is a promising tool for achieving high-throughput dynamic quantitative phase imaging of biological cells.

## 2. Principle and Methods

### 2.1. High-Throughput QPI via Single-Shot Intensity Measurement

High-throughput microscopy permits access to high-throughput quantitative analysis for multiple events in a large population of cells.<sup>[36,37]</sup> However, the achievable SBP of conventional microscopy is fundamentally limited by the optical system, leading to an inevitable trade-off between FOV and imaging resolution. This limitation can be intuitively illustrated by **Figure 1a**. A commercial objective lens with low magnitude (UPlanSApo 10x, 0.4 NA, Olympus) allows the observation of tens of thousands of red blood cells across the FOV of  $\approx 2.25 \text{ mm}^2$ , but the spatial resolution is insufficient to distinguish detailed structures. In contrast, an alternative objective lens with higher magnitude (UPlanSApo 60x, 1.35 NA, Olympus) enables analysis for high-resolution cellular structures and details, such as the sharp boundaries of red

blood cells and the white blood cells' internal particles. However, compared with the large FOV of the 10x objective lens, the achievable FOV of the 60x objective lens shrinks to  $\approx 0.06 \text{ mm}^2$ , where only hundreds of cells could be observed. Hence, it is difficult to take into account large FOV and high resolution simultaneously in conventional microscopic imaging systems.<sup>[38]</sup>

To decouple FOV and resolution from each other in a microscope, considerable research has been conducted, such as imaging stitching,<sup>[39]</sup> synthetic aperture microscopy,<sup>[10,40]</sup> lensless on-chip microscopy,<sup>[41–43]</sup> and FPM,<sup>[23,26,44–46]</sup> achieving high-throughput microscopic imaging with spatial-domain or frequency-domain methods. Image stitching is a simple and widely used approach that mitigates the trade-off between FOV and resolution by scanning the field with a high numerical aperture (NA) objective lens and then stitching the high-resolution segments in the spatial domain. However, the cost of an image stitching system is usually expensive due to the pricey high-NA objective lens and the high-precision electric scanners used. Besides, the necessary mechanical scanning, refocusing, and registration procedures also induce extra computation, resulting in a restriction of space-bandwidth-time production. In contrast, as mentioned above, FPM is a novel QPI technique that uses a low-NA objective lens to take advantage of its innate large FOV and stitches together images in the frequency domain. By varying the illumination angle, FPM shifts different high spatial frequency components of the object spectrum into the passband of the low-NA objective lens, and realizes high-throughput phase imaging using a low-cost system. Nevertheless, the basic strategy of the above-mentioned techniques is to trade numerous data measurements for high system throughput. This data reliance often requires sophisticated optical setups or elaborate illumination schemes, leading to time-consuming data acquisition and severe storage burden problems. On the other side, TIE is a well-established deterministic QPI approach that simply utilizes intensity measurements at multiple axially displaced planes to obtain the axial intensity derivative and reconstruct the quantitative phase (Figure 1b).<sup>[22]</sup> Thanks to its Köhler illumination compatibility within an off-the-shelf bright-field microscope, TIE eliminates the need for elaborate illumination schemes and optical

setups. Additionally, TIE recovers phase in a non-iterative manner with a requirement of only a few intensity measurements, which improves the data efficiency, reduces the storage burden and brings higher imaging speed. Nevertheless, despite its high efficiency, the throughput of phase reconstruction is still fundamentally constrained by the optical system, since TIE is always limited by BF illumination, and the maximum attainable imaging resolution is restrained to the incoherent diffraction limit when matched annular illumination is used.<sup>[24,25,47]</sup> Consequently, a computational QPI technique for wide-field high-resolution and high-throughput phase reconstruction from single-shot intensity measurement is yet to be developed.

The proposed DL-SRQPI has the ability to retrieve high-resolution phase images from low-resolution intensity image using a well-trained neural network (Figure 1c). Thanks to the unique end-to-end mapping mechanism and powerful high-dimensional feature extraction capability of deep learning, DL-SRQPI minimizes the data acquisition requirement to a single intensity measurement. With the help of GPU, this hybrid approach of deep learning and QPI addresses the above-mentioned data reliance limitation, improving the data efficiency and speeding up the phase imaging simultaneously. Meanwhile, DL-SRQPI provides a significant improvement in SBP and realizes high-throughput QPI by enhancing the image resolution without sacrificing the FOV, alleviating the resolution-FOV trade-off in a computational manner. Besides, DL-SRQPI allows for simple and straightforward implementation on a conventional bright-field microscope at a low cost, giving the possibility for its wide application.

## 2.2. Image Preprocessing and Dataset Construction

As an approach based on the end-to-end supervised-learning strategy, DL-SRQPI gains its capability from the training datasets consisting of well-paired low-resolution (LR) intensities and high-resolution (HR) phases.<sup>[28]</sup> We constructed simulated and experimental datasets separately to train DL-SRQPI progressively.

In the simulated dataset, we used numerical propagation methods, including angular spectral propagation and Abbe superposition, to accurately and efficiently generate low-resolution intensity images at different defocus distances and illumination patterns from ground truth high-resolution phases. The propagated complex field utilizing the angular spectrum method, which models the propagation of a wave field by using an analytic formula, can be calculated by Equation (1)

$$U(\mathbf{x}, z) = \mathcal{F}^{-1} \left\{ \hat{U}(u_x, u_y, 0) \exp \left[ j \frac{2\pi}{\lambda} z \sqrt{1 - (\lambda u_x)^2 - (\lambda u_y)^2} \right] \right\} \quad (1)$$

where  $\mathbf{x}$  represents the 2D spatial coordinate  $(x, y)$  in the real space, the scalar coherent field  $U(\mathbf{x}, 0)$  (assuming  $z = 0$ ) is decomposed into the coherent superposition of the angular spectrum (plane wave) components  $\hat{U}(u_x, u_y, 0) = \mathcal{F}\{U(\mathbf{x}, 0)\}$ , and  $\exp[j \frac{2\pi}{\lambda} z \sqrt{1 - (\lambda u_x)^2 - (\lambda u_y)^2}]$  is a phase delay factor, which is also known as the angular spectrum transfer function. Then, the

intensity  $I(\mathbf{x})$  at propagation distance  $z$  can be calculated by multiplying the complex field  $U(\mathbf{x}, z)$  and its conjugate  $U^*(\mathbf{x}, z)$ .<sup>[48,49]</sup>

For the generation of intensities of partially coherent field, we utilized Abbe's superposition method, which describes the formation model of intensity images under different illumination conditions. The Abbe's method could be described by Equation (2)

$$I(\mathbf{x}) = \int S(\mathbf{u}) I_{\mathbf{u}}(\mathbf{x}) d\mathbf{u} \quad (2)$$

where  $S(\mathbf{u})$  is the Fourier transform of the source intensity distribution,  $I_{\mathbf{u}}(\mathbf{x})$  is the coherent partial image arising from the point of the incoherent source. Equation (2) implies that a partially coherent intensity image can be represented as an incoherent superposition of all intensities  $I_{\mathbf{u}}(\mathbf{x})$  generated by all light source points at the condenser aperture plane.<sup>[49]</sup> By this means, we can generate intensity images under various illumination conditions.

The simulated datasets consist of FPM phase images as labels and their calculated LR intensity images as inputs. The label phase images are from our previously published paper.<sup>[50]</sup> The accurately matched intensity images were generated from the ground truth phase images by the above-mentioned numerical propagation methods. With the angular spectrum method, we digitally back-propagated a ground truth phase image to 13 intensity images at various defocus distances, within a range of  $z = (+1 \mu\text{m}, +13 \mu\text{m})$  with  $\Delta z = 1 \mu\text{m}$  increments. With the Abbe's superposition method, we also generated defocused intensity images ( $z = +4 \mu\text{m}$ ) under different coherent parameters within a range of  $S = [0, 0.4]$  with  $S = 0.1$  increments. These generated intensity images were then corrupted by Gaussian noise with a standard deviation of 0.01 to simulate the noise effect. Subsequently, we downsampled the intensity images to simulate the increase of pixel size and introduce the artificial pixel aliasing.<sup>[27]</sup> After the operation of  $4\times$  pixel binning, the pixel resolution of the intensity images was reduced from  $2048 \times 2048$  pixels to  $512 \times 512$  pixels, while the FOV remained unchanged. So far, we had obtained 17 LR intensity images on 13 defocus distances and four coherent parameters. With these images, we constructed 17 simulated datasets, each comprising the original phase image and one intensity image on a specific defocus distance and a coherence parameter generated from the phase. To reduce the memory demand of the computer and speed up the process of network training, we divided the full-field LR intensity images and the corresponding HR phase images into paired image patches, and the patches without valid information were excluded via edge detection algorithm. For each simulated dataset, the full-field intensity image and phase image were segmented into  $64 \times 64$  pixels and  $256 \times 256$  pixels image patches, respectively. With further augmentation by mirroring and rotation, each dataset eventually contains 1480 LR intensity and HR phase image pairs. Out of these images, 1300 pairs were randomly selected to be used as the training dataset, while 80 pairs were used as the validation dataset for validating the network performance and selecting the optimal model, and the remaining 100 pairs formed the testing dataset to blindly quantify the average performance of the final network. To ensure fairness across networks, though an individual dataset was randomly divided, the division of each dataset was identical.

The experimental dataset consists of real experimental intensity images, which are from our previous work.<sup>[51]</sup> The differences in exposure time during intensity image acquisition and the inhomogeneous light absorption of sample areas resulted in non-uniform field brightness, which had a negative impact on the accuracy of phase retrieval. This influence of spatial variability of the light intensity can be corrected with the algorithm  $C_{ij} = R_{ij} \frac{\bar{S}}{\bar{R}}$ , where  $C$  is the corrected image,  $R$  is the original image,  $\bar{R}$  is the average gray level of the image  $R$ ,  $\bar{S}$  is the average gray level of all 270 intensity images, and the subscripts  $i$  and  $j$  indicate that the correction is performed on the  $i$ th and  $j$ th pixel of the image. The phase images can be recovered by TIE, which is given by Equation (3)

$$-k \frac{\partial I(\mathbf{x})}{\partial z} = \nabla \cdot [I(\mathbf{x}) \nabla \phi(\mathbf{x})] \quad (3)$$

where  $\nabla$  denotes the 2D gradient operator with respect to  $x$  and  $y$ , and  $k = 2\pi/\lambda$  is the wave number. Then the phase  $\phi(\mathbf{x})$  can be extracted by solving the equation. The left hand of TIE is the spatial derivative of intensity at the in-focus plane along the  $z$ -axis.<sup>[52]</sup> The right hand of TIE is a second-order elliptic partial differential equation, and we treat it as a Poisson equation, ideally, which can be easily solved with fast-Fourier transform (FFT).<sup>[53]</sup>

We corrected the field brightness of the experimentally acquired 135 intensity image stacks and then reconstructed their corresponding phase images via TIE. To introduce noticeable pixel aliasing effect, we downsampled the resolution of the intensity images from  $2588 \times 1960$  pixels to  $647 \times 490$  pixels by performing  $4\times$  pixel binning. This can be regarded as an  $4\times$  enlargement of the pixel size, resulting in a significant loss of detailed information. The TIE-retrieved phase images and the low-resolution intensity images were used as input and labels for the experimental dataset. Same as the simulated datasets, we constructed the experimental dataset by extracting paired  $128 \times 128$  pixels and  $512 \times 512$  pixels image patches from a pair of experimentally captured full-field LR intensity image and its HR TIE phase image, and augmented the dataset to 433 pairs by rotating and mirroring. 400 pairs were randomly selected as the training dataset, eight pairs were selected as the validation dataset, and the remaining 25 pairs as the testing dataset.

### 2.3. Network Architecture and Training

DL-SRQPI adopts U-Net (Figure 2a) as the neural network to achieve high-speed high-throughput QPI. U-net is a remarkable CNN-based network with excellent performance in biomedical image processing.<sup>[35,54,55]</sup> In DL-SRQPI, U-Net transforms previously fuzzy inferior intensity images with pixel aliasing to clear, superior, alias-free phase images. As shown in Figure 2a, the network consists of an interpolation operation at the input, an encoder branch for feature extraction, and a decoder branch for feature reconstruction, with skip connections combining the features from two branches. The interpolation operation is performed to align the resolution of the intensity image and the phase image, so that the network can identify the complex mapping relationship of super-resolution phase retrieval. The en-

coder branch consists of four identical downsampling modules, each including a  $2 \times 2$  max pooling layer and two convolutional layers with a  $3 \times 3$  kernel and stride of 2. The skip connection path connects the extracted features of each stage of the encoder branch to the corresponding feature layer of the decoder branch. The decoder branch consists of four identical upsampling modules, each consisting of an upsampling convolutional layer concatenating with the corresponding feature map in the encoder branch by skip connection, and two convolutional layers with a  $3 \times 3$  kernel. All convolutional layers are followed by a batch normalization module (BN) and a rectified linear unit (ReLU) to achieve faster training and enhance the nonlinear ability. During training (Figure 2b), the mean-square-error (MSE) between the output phase image and the label phase image was calculated as the loss function, and was back-propagated to the network for optimization. For an image with a size of  $M \times N$  pixels, this loss function over a mini-batch at size of  $K$  is calculated by Equation (4)

$$\text{Loss}(\Theta) = \frac{1}{K} \sum_{k=1}^K \frac{1}{M \times N} \sum_{m=1}^M \sum_{n=1}^N \|Y_{m,n,k}^{\Theta} - Y_{m,n,k}^{\text{GT}}\|^2 \quad (4)$$

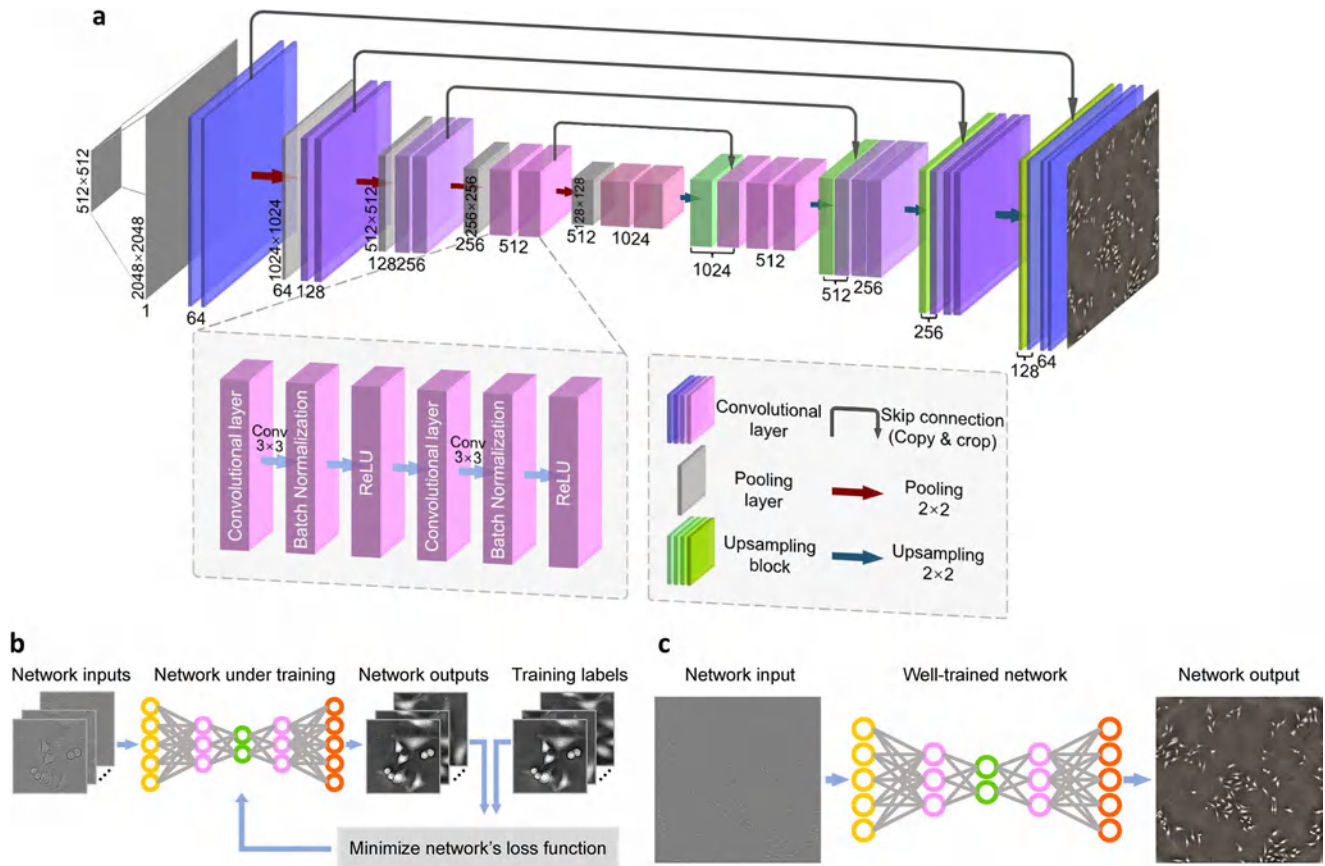
where  $k$  is the  $k$ th image patch among the mini-batch,  $Y_{m,n,k}^{\Theta}$  denotes the  $m$ th and  $n$ th pixel of network output phase image, and  $Y_{m,n,k}^{\text{GT}}$  denotes the  $m$ th and  $n$ th pixel of the training labels (i.e., ground truth). The network's parameter space (e.g., kernels, biases, and weights) is defined by  $\Theta$  and its output is given by  $Y^{\Theta} = F(X_{\text{input}}; \Theta)$ , where  $F$  defines the deep neural network's operation on the network input intensity  $X_{\text{input}}$ . The adaptive moment estimation (ADAM) optimization algorithm with a learning rate of 0.01 is utilized to minimize the MSE and tune the network parameters. After sufficient training, the network has established the mapping relationship between the LR intensity image and the HR phase image. It is worth mentioning that due to the translation invariance of the convolutional neural network, the network can output full-field phase images from full-field intensity images, despite only patches being used during training. The metric used to measure the accuracy of phase retrieval is given by the structural similarity index (SSIM), which comprehensively evaluates luminance ( $l(x, y)$ ), contrast ( $c(x, y)$ ), and image structure ( $s(x, y)$ ), to quantify the similarity of the ideal phase and the retrieved phase. It can be calculated by Equation (5)

$$\text{SSIM}(x, y) = [l(x, y)]^{\alpha} \cdot [c(x, y)]^{\beta} \cdot [s(x, y)]^{\gamma} \quad (5)$$

where

$$\begin{cases} l(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \\ c(x, y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \\ s(x, y) = \frac{\sigma_{x,y} + c_3}{\sigma_x\sigma_y + c_3} \end{cases} \quad (6)$$





**Figure 2.** Network architecture and schematics of network training and testing. a) The U-Net structure is depicted, where each block represents a multi-channel feature map. The number of channels is indicated at the bottom of each block, while the size is denoted in the lower left corner. b) The network training workflow involves utilizing low-resolution intensity image patches as inputs and corresponding high-resolution phase image patches as training labels. The network optimizes its parameters by minimizing the loss function (MSE) between the network outputs and the training labels. c) During network testing, a full-FOV low-resolution intensity image is presented as the input, and the well-trained network generates a full-FOV high-resolution phase image as the output.

where  $\mu_x$ ,  $\mu_y$ ,  $\sigma_x$ ,  $\sigma_y$ , and  $\sigma_{xy}$  are the local means, standard deviations, and mutual covariances of the images  $x$  and  $y$ . When  $\alpha = \beta = \gamma = 1$  and  $c_3 = c_2/2$ , the SSIM index simplifies to

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{x,y} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (7)$$

The SSIM index varies between 0 and 1, where 1 can be achieved if predicted and ground truth images are identical to each other. The fixed network gains the capability to blindly output full-field high-resolution phase images at a high reconstruction speed (Figure 2c), providing a great enhancement of space-bandwidth-time product for the QPI system.

### 3. Results

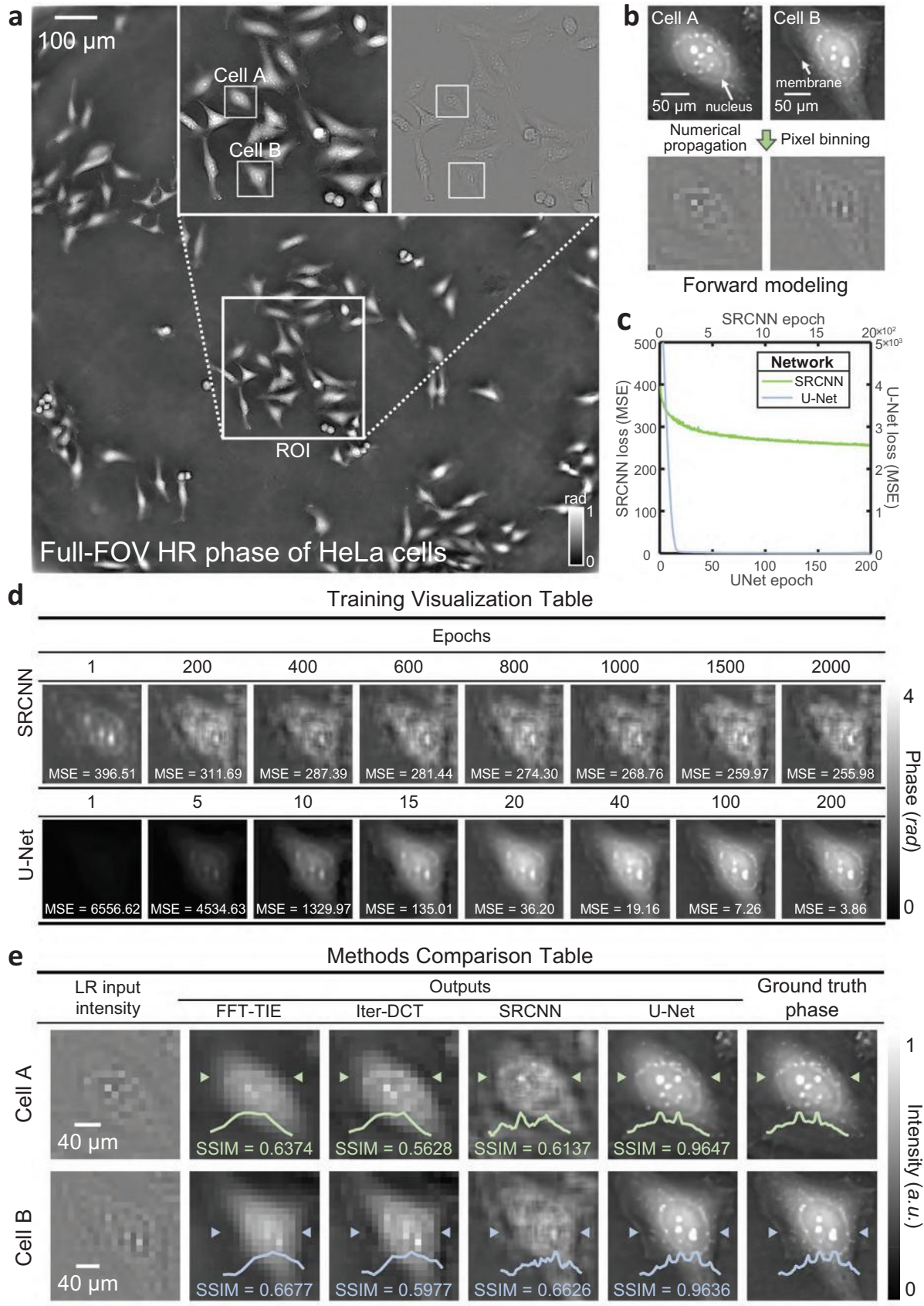
#### 3.1. Benchmarking of DL-SRQPI

To illustrate the applicability of our proposed method, we compared DL-SRQPI with two traditional TIE-based methods (FFT-TIE,<sup>[22]</sup> iterative DCT<sup>[56]</sup>) and a classic super-resolution neural

network SRCNN.<sup>[57]</sup> The full-field FPM phase image was used as the ground truth (Figure 3a), with high quality and abundant detail information. As mentioned in Section 2.2, the LR defocused intensity image used for phase reconstruction was generated following the forward model consisting of angular propagation and pixel binning (Figure 3b), and was corrupted by Gaussian noise to simulate the noise effect with a standard deviation of 0.01.

To make a fair comparison between SRCNN and U-Net, both networks are fully trained with the same dataset and loss function (MSELoss). In Figure 3c, we show how the MSE values of two networks decrease. SRCNN has a low initial loss value of  $\approx 400$ , but it decreases slowly and converges to  $\approx 250$  after 2000 epochs of training. In contrast, though the initial loss value of U-Net is as high as  $\approx 6500$ , the loss function sharply decreases to less than 200 within 15 epochs, and eventually converges to  $\approx 4$ , exhibiting much stronger mapping ability. More intuitively, the outputs of two networks during the training process are shown in Figure 3d. During training of SRCNN, the output images show little improvement. The poor response at low-spatial frequencies leads to severe noise and artifacts in the output, resulting in difficulty distinguishing the cell boundary and inner structure. On the other hand, as U-Net training proceeds and the loss function steadily





decreases, the precise phase gradually emerges from the dark background. The phase reconstruction of DL-SRQPI achieves an SSIM index of 0.96 with respect to the ground truth, indicating its high precision. In the output of DL-SRQPI, the optically thick nucleus, the cell membrane, and some cytoplasmic organelles are shown with high contrast and clarity. Compared with other methods, the high-quality reconstructions of DL-SRQPI exhibit abundant subcellular features in high resolution, unraveling the aliased pixels in the input images.

In Figure 3e, we compare the phase retrieval of these methods for two HeLa cells. After 4× pixel binning, the intensity images become fuzzy and inferior. Due to its low contrast, the intensity images cannot provide much cellular information, and are further exacerbated by the severe pixel aliasing problem. The high frequency detail is almost completely lost in the under-resolved intensity image, preventing the observation of detailed subcellular structures and making phase retrieval very challenging. Since the FFT-TIE and Iter-DCT methods do not possess super-resolution capability, their retrieved phases are blurry and the detailed information is still buried in oversized pixels. Only optically thick cellular structures in the phase results can be visualized, such as the nucleus, while the high-frequency details are completely lost. In Figure 3e, we demonstrate the line profiles across the nucleus and the cell membrane of each output. The line profiles of FFT-TIE and Iter-DCT only show the general trend of phase changes and are unable to present the detailed phase variation of the internal cellular structures. This result reveals the gap in resolution between TIE-based and deep learning methods. Notably, two intensity images at different defocus distances are required by FFT-TIE and Iter-DCT due to their data redundancy requirement, while the deep learning methods only require a single frame of intensity image as input for its end-to-end mechanism.

### 3.2. Generalization Capability Analysis for Axial Defocusing and Illumination Condition

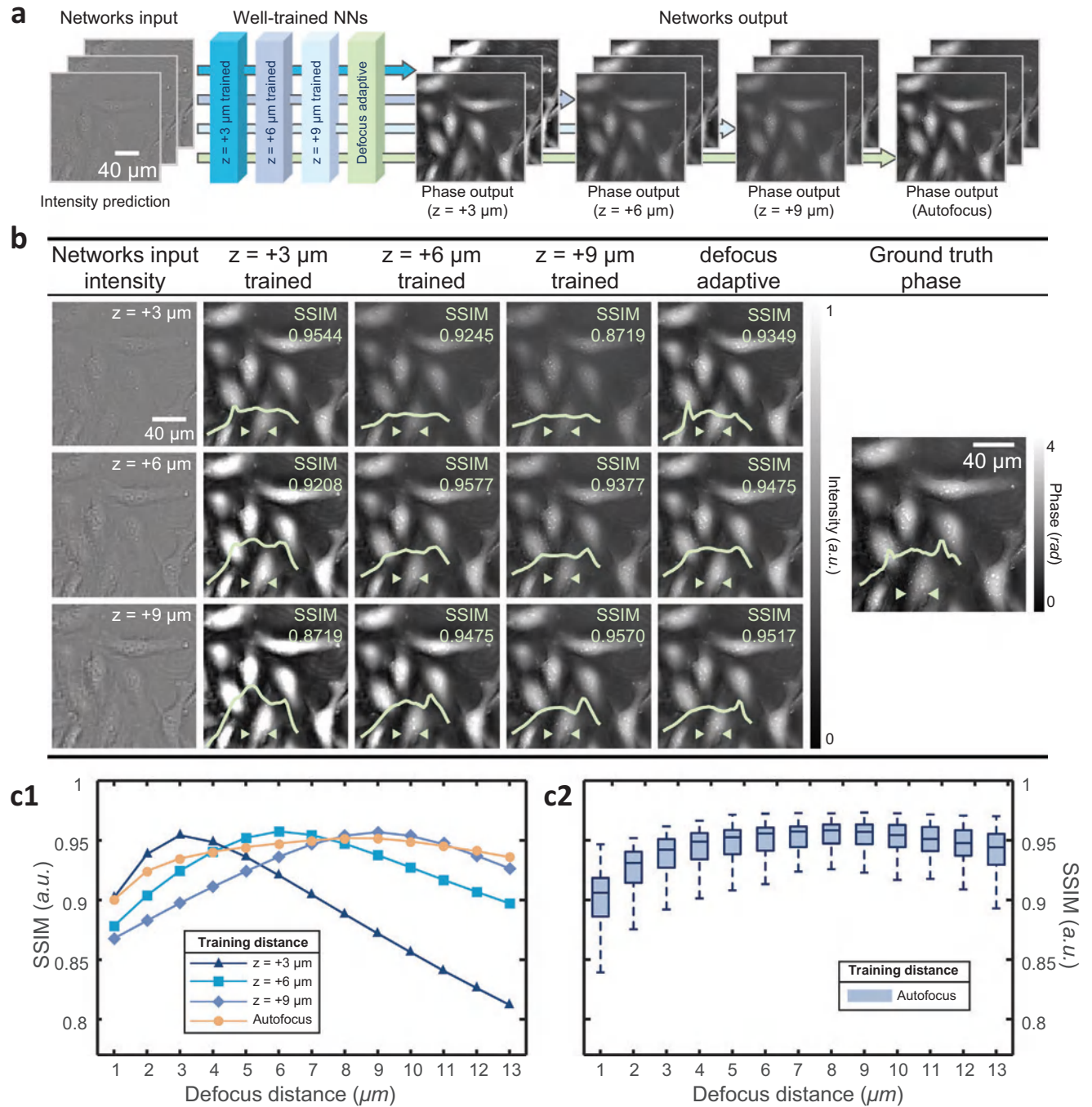
Axial intensity derivative estimation is a key issue in TIE-based QPI methods. TIE requires multiple defocused intensity measurements to achieve phase retrieval. The defocus distance has to be large enough to ensure an adequate SNR,<sup>[49]</sup> but too large a defocus distance tends to introduce phase blurring effect. This noise-resolution trade-off requires a strict and precise choice of defocus distance for TIE-based phase retrieval. Therefore, DL-SRQPI is expected to possess the capability to recover high-quality phase robustly from a single frame of intensity image at a random defocus distance. To initially analyze the generalization capability of DL-SRQPI for axial defocusing, we used the 13 datasets that contain variously defocused intensity images ( $z = [+1 \mu\text{m}, +13 \mu\text{m}]$ ,  $\Delta z = 1 \mu\text{m}$ ) to train 13 U-Net networks independently, which are mentioned in Section 2.2. Following the for-

ward model in Figure 4a, after adequate training, we blindly fed 13 sets of testing intensity images at different defocus distances into each network, each testing dataset containing 100 identically defocused LR intensity images. Note that the image pairs used in the training and test datasets are completely different. All 13 networks responded rapidly to each set of testing images, and the HR phase outputs are shown in Figure 4b. The line profiles show the phase variation along the same region. Notably, with the decrease of testing defocus distance, the phase results become dimmer; while the testing defocus distance becomes higher, the phase results become brighter and even get overexposed. This can be regarded as proof that the neural network has mastered the mapping relationship between intensity and phase: according to Equation (3), too large an estimate of the defocus distance  $z$  results in low phase values of the reconstruction result. To quantify the accuracy and the quality of the phase reconstructions, the average SSIM index for the outputs of each testing dataset was evaluated with respect to the ground truth, and the SSIM indexes are shown in Figure 4c. When the defocus distance of the testing images matches the training images, the phase results are of the highest quality with the average SSIM index reaching nearly 0.96. As the testing defocus distance differs from the training defocus distance, the accuracy and the SSIM indexes of the phase results drop slowly. Figure 4 illustrates that though each network was trained with intensity images at one particular defocus distance, DL-SRQPI still has the ability to reconstruct HR phase images from intensity images at a range of defocus distances with a negligible drop in reconstruction quality.

To maximize the generalization capability of DL-SRQPI for axial defocusing, we used 13 sets of intensity images at various defocus distances to construct a training dataset, each containing 100 identically defocused intensity images. After proper training, DL-SRQPI adapts the variation in defocus distance. We blindly tested the network with the above testing dataset and calculated the average SSIM indexes of the phase outputs from each set of intensities at a certain defocus distance. All these average indexes are higher than 0.9, and the overall average SSIM index reaches nearly 0.95, representing the network is adaptive to axial defocusing. So far, we have demonstrated that DL-SRQPI is robust to the variation of defocus distance, and the proposed defocus adaptive DL-SRQPI could further enhance the generalization ability on axial defocusing.

The spatial coherence of the illumination is also an essential factor for phase retrieval. It can be quantified by a normalized factor  $S = NA_{\text{cond}}/NA_{\text{obj}}$  (so-called coherent parameter), where  $NA_{\text{cond}}$  is the numerical aperture of the condenser lens and  $NA_{\text{obj}}$  is the numerical aperture of the objective lens. Similar to the noise-resolution trade-off in the choice of defocus distance, the spatial coherence also brings a compromise between the contrast and the resolution of phase recovery. Reducing  $NA_{\text{cond}}$  can effectively improve the phase contrast, but at the same time reduce the imaging resolution of the system. The generalization capability

**Figure 3.** Methods comparison and training visualization. a) The full-FOV high-resolution ground truth phase image of HeLa cells in vitro, along with the phase of a ROI (region of interest) and its intensity image obtained through numerical propagation. b) The defocused low-resolution intensity images of cell A and cell B were generated by numerical propagation from their high-resolution phase images and subsequently downsampling them through a 4× pixel binning. c) The plot of the Mean Squared Error loss function for SRCNN and U-Net. d) Phase reconstruction images and corresponding loss function values from the outputs of SRCNN and U-Net at different epochs during network training. e) Phase reconstruction results of FFT-TIE, Iter-DCT, SRCNN, and U-Net using the low-resolution intensity images of cell A and cell B as inputs, along with the line profiles of the cells.



**Figure 4.** Analysis of axial defocusing generalization capability. a) The forward model of DL-SRQPI for intensity images with different defocus distances. Each neural network is trained with intensity images at a specific defocus distance and tested blindly with intensity images at different defocus distances. b) The phase reconstructions of testing defocused intensity images ( $z = +3/+6/+9 \mu\text{m}$ ), output by three well-trained networks trained on intensity images at defocus distances of  $z = +3/+6/+9 \mu\text{m}$ , along with their SSIM indexes compared to the ground truth phase image, and their line profiles showing the subcellular features. c1) The average SSIM index curves depicting the similarity between the phase images obtained from intensity images at different defocus distances ( $z = +1$  to  $+13 \mu\text{m}$ ) and the ground truth phase, using the defocus adaptive network and three networks trained on three sets of defocused intensity images ( $z = +3/+6/+9 \mu\text{m}$ ). c2) The boxplot of the SSIM index of the phase images obtained from the defocus adaptive network and three networks trained on three sets of defocused intensity images using 13 groups of testing datasets, and the ground truth phase images.



of DL-SRQPI on coherent parameters has been verified. We used the five datasets mentioned in Section 2.2 to train five U-Net networks independently, each containing intensity images at a certain parameter ( $S = [0, 0.4]$ ,  $\Delta S = 0.1$ ). After training, each well-trained network was blindly tested with five sets of testing images at different coherent parameters, each set containing 100 LR intensity images (Figure 5a). Note that the image pairs used in the training and test datasets are completely different. The results are shown in Figure 5b and the average SSIM indexes of each set of outputs were evaluated with respect to the ground truth (Figure 5c). The phase results and the line profiles in Figure 5b show that when the coherent parameter of the testing intensity images matches the coherent parameter of the training intensity images, the quality of the phase results reaches the highest, with an average SSIM index of 0.95, which shows the high similarity between ground truth and network output images. As the coherent parameter of the testing image gets lower, the phase imaging contrast reduces, bringing a blurry effect on the internal structures and edges of cells, causing a significant loss of high-frequency detail information. With the increase of testing coherent parameter, the phase imaging contrast becomes excessively high, resulting in a sharpening-like effect in the phase reconstruction, which added difficulty to the cell morphology analysis. The box plot of the SSIM indices for each set of outputs from the  $S = 0.4$  trained network shows a small variance in the accuracy of the outputs, indicating the stability of the DL-SRQPI. By far, the robustness of the phase retrieval framework of DL-SRQPI is critically estimated through the above testing experiments.

### 3.3. Quantitative and Generalizable Characterization of DL-SRQPI

We conducted an experiment to demonstrate the quantitative property of DL-SRQPI using a microlens array as the test subject. For the establishment of the ground truth high-resolution phase image, the TIE algorithm was utilized to reconstruct the accurate phase image of the microlens array. Subsequently, we performed  $4\times$  pixel binning on the original defocused intensity image to generate a low-resolution intensity image. This process enlarged the pixel size to  $8.8\ \mu\text{m}$  and decreased the pixel resolution of the defocused intensity image from  $1280 \times 960$  pixels to  $320 \times 240$  pixels. With the LR defocused intensity image and ground truth HR phase image, we created a dataset comprising high-resolution phase and low-resolution intensity image pairs of the microlens array using the same method described in Section 2.2. We trained a U-Net network and blindly fed it with a new defocused low-resolution intensity image of the microlens (Figure 6 a2). The network rapidly generated a high-resolution phase reconstruction (Figure 6 a3), and the heights of the microlens were deduced from the output phase image (Figure 6 a4). In Figure 6 a5, we compare the line profiles of the DL-SRQPI output phase, the ground truth high-resolution phase, and the low-resolution phase for a single lens. It is evident that the three profiles almost completely overlap. As seen in the zoomed area, the discrete pixels of LR phase exhibit a noticeable stair-step pattern due to the pixel binning, indicating a large pixel size and a low resolution. Conversely, the DL-SRQPI achieves a high resolution comparable to the ground truth HR phase. Figure 6 a6 presents

the difference between the ground truth phase and the output phase of DL-SRQPI. The difference between the output phase and the ground truth phase is only 1 to 2 gray levels, which is equivalent to 0 to  $0.2\ \mu\text{m}$  in height, showing the high accuracy of DL-SRQPI. These results confirm the precise quantitative characteristics of DL-SRQPI.

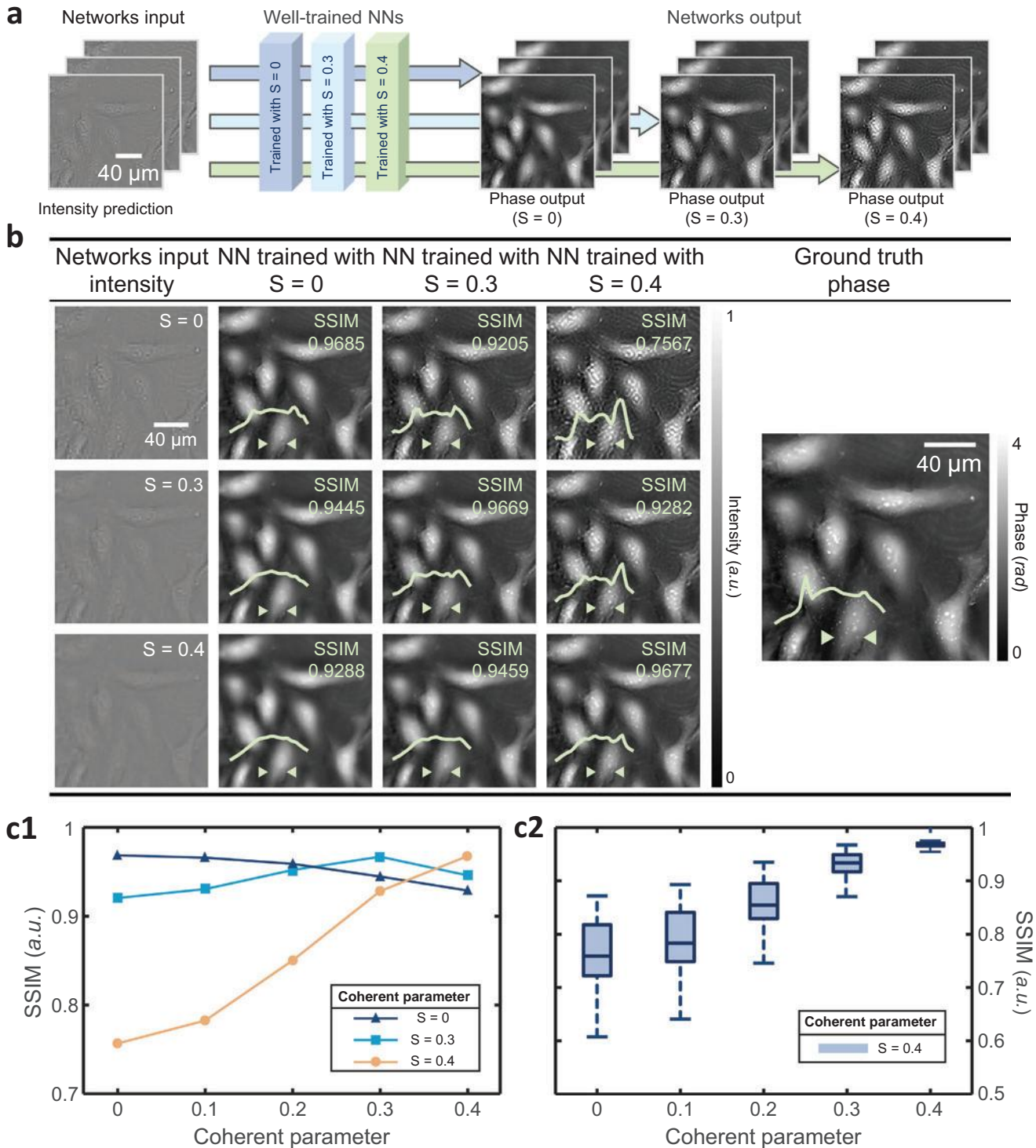
To further demonstrate the generalizability of DL-SRQPI, we utilized experimentally acquired differently defocused intensity images of HeLa cells. Following the same method described in Section 2.2, we constructed three experimental datasets for HeLa cells, each consisting of HR phase images and LR intensity images captured at different defocus distances ( $z = +3/+6/+9\ \mu\text{m}$ ). Subsequently, we trained three separate U-Net networks using these datasets. Figure 6 b1–b3 displays three LR intensity images of a single HeLa cell captured at the corresponding defocus distances of  $z = +3/+6/+9\ \mu\text{m}$ . These LR intensity images were input into their respective networks, which rapidly generated four-fold pixel super-resolved phase images (Figure 6 b4–b6). Importantly, the output phase images obtained from the LR intensity images at different defocus distances all exhibited high SSIM values when compared to the ground truth (Figure 6 b7). This experimental evidence clearly demonstrates the strong generalization capability of DL-SRQPI.

### 3.4. Evaluation on the Simulation Dataset Using DL-SRQPI

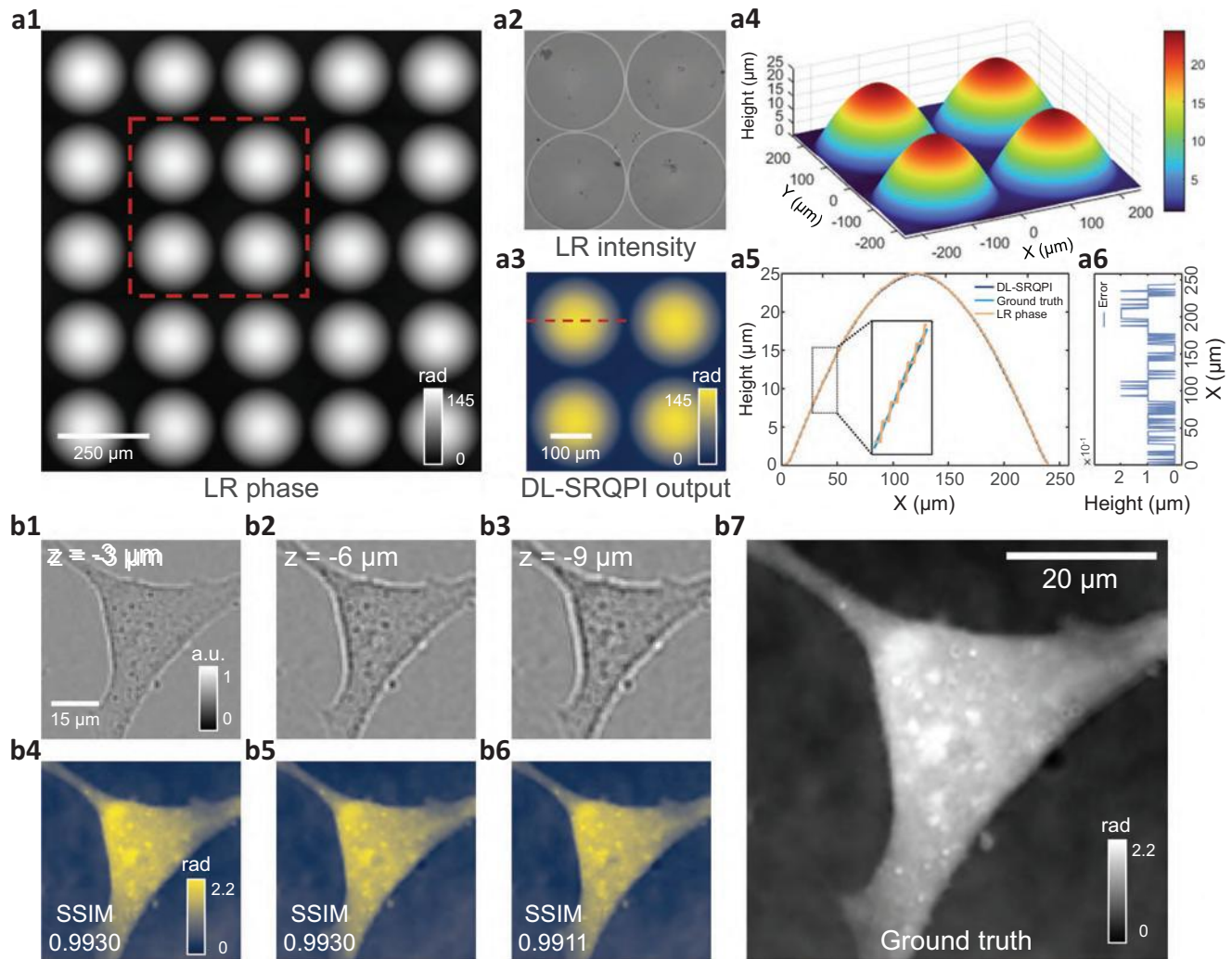
We verified the feasibility of DL-SRQPI to achieve rapid high-throughput single-shot QPI using the simulation dataset of in vitro HeLa cells. Figure 7 shows the full FOV phase prediction for the simulated dataset of HeLa cells in vitro. The simulated input LR intensity image has a wide FOV of  $\approx 1.77\ \text{mm}^2$ , matching the FOV size of the objective lens with  $10\times$  magnification. The input intensity image has a resolution of  $512 \times 512$ -pixel with an effective pixel size of  $2.6\ \mu\text{m}$ . The high-throughput phase reconstruction is displayed in Figure 7a, which shows that our DL-SRQPI is able to achieve a four fold enhancement in the pixel resolution from  $512 \times 512$  pixels to  $2048 \times 2048$  pixels while maintaining the large FOV size of  $\approx 1.77\ \text{mm}^2$ , as the effective pixel size improves to  $0.65\ \mu\text{m}$ . The comparison between LR intensity and the predicted HR phase of two ROIs is shown in Figure 7a. Compared to the low-contrast aliased intensities, the recovered phases display improved overall contrast of cell organelles and highlight high-spatial-frequency subcellular details.

Additionally, thanks to the non-invasive and nontoxic properties of DL-SRQPI, it can also serve as a practical tool for visualizing the morphological dynamics of living HeLa cells. We used DL-SRQPI to reconstruct phase imaging videos with large SBP (Videos S1 and S2, Supporting Information). As shown in Figure 7b, cells A and B are enlarged to present different typical mitosis phases and the morphological evolution of cells during the mitotic cycle. The subcellular features of both cells, such as cytoplasmic vesicles and pseudopodium, and their sub-pixel organelle motions, such as plasmid migration, are demonstrated in the video. Since each HR phase image was reconstructed within only 0.3 s, which can be further accelerated with higher performance hardware, all the retracting, extending, reorganizing, migrating, and maturing processes of cells could be recovered accurately avoiding motion blur, which lays the foundation for the





**Figure 5.** Analysis of generalization capability for illumination condition. a) The forward model of DL-SRQPI for intensity images with various coherent parameters. Each network is trained with intensity images at a specific coherent parameter and tested blindly with intensity images at different coherent parameters. b) The phase reconstructions of testing intensity images ( $S = 0/0.3/0.4$ ), output by three well-trained networks trained on intensity images at illumination condition of  $S = 0/0.3/0.4$ , along with their SSIM indexes compared to the ground truth phase image, and their line profiles showing the subcellular features. c1) The average SSIM index curves depicting the similarity between the phase images obtained from intensity images at different coherent parameter ( $S = 0/0.3/0.4$ ) and the ground truth phase, using three networks trained on three sets of intensity images at the same coherent parameter. c2) The boxplot of the SSIM index of the phase images obtained from the well-trained network ( $S = 0.4$ ) using five groups of testing datasets, and the ground truth phase images.



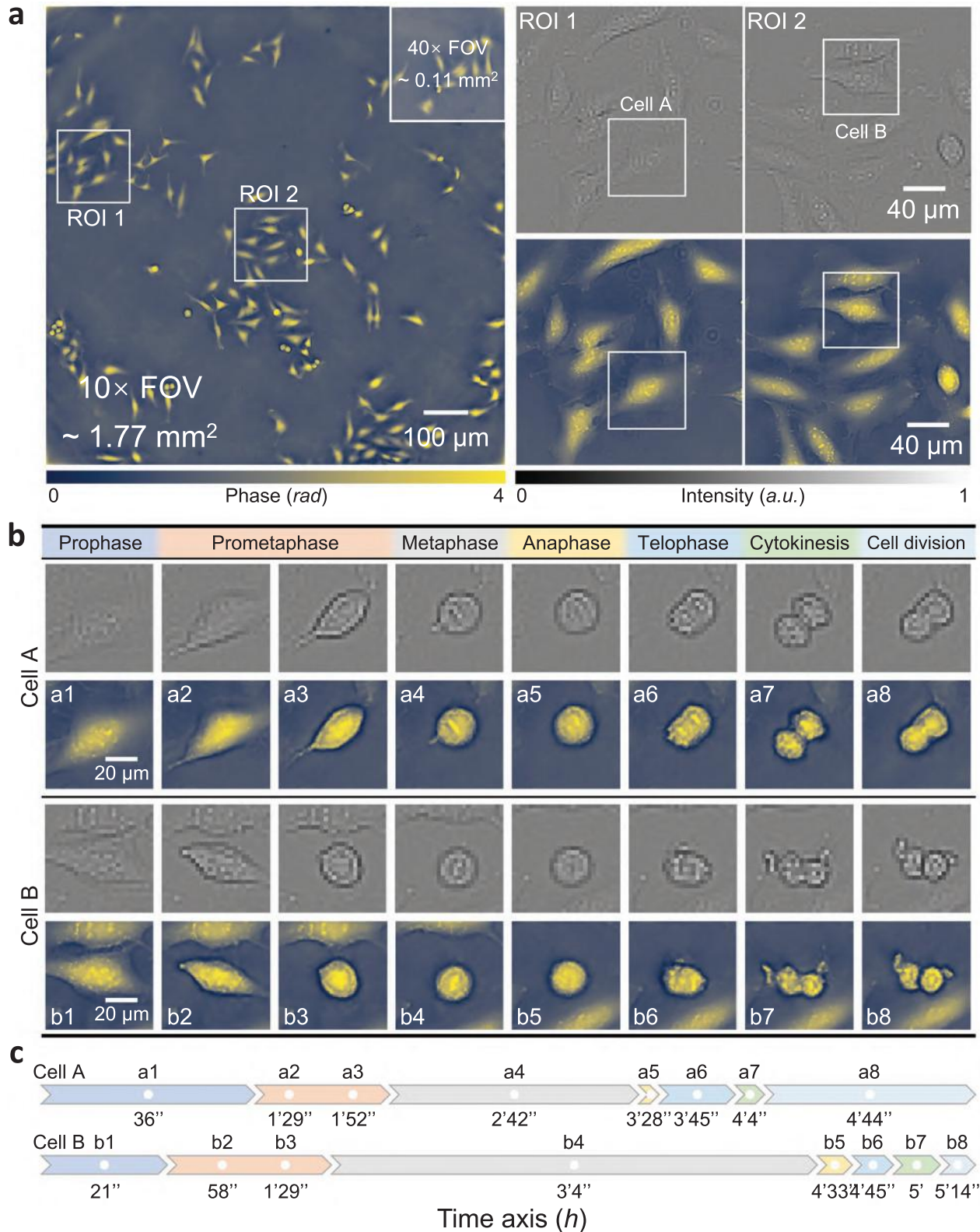
**Figure 6.** Validation of quantitative and generalization capabilities of experimental data based on a microlens array and a defocused HeLa cell. a1) The low-resolution phase image of the microlens array retrieved from low-resolution intensity images via TIE. a2) The low-resolution defocused intensity image of the microlens array. a3) The high-resolution phase reconstruction result output by DL-SRQPI. a4) The 3D topography of the microlens array. a5) Comparison of DL-SRQPI output, ground truth HR phase and LR phases. a6) The error of DL-SRQPI output, that is, the difference between ground truth phase and DL-SRQPI output phase. b1–b3) The input defocused LR intensity images of a HeLa cell ( $z = +3/+6/+9 \mu\text{m}$ ). b4–b6) DL-SRQPI reconstructed HR phase images from intensity images of (b1–b3). b7) The ground truth HR phase image of the HeLa cell.

practical application of DL-SRQPI in the fields of cytomorphology, cytokinetics, and cytogenetics.

### 3.5. Experimental Phase Imaging of HeLa Cells In Vitro Using DL-SRQPI

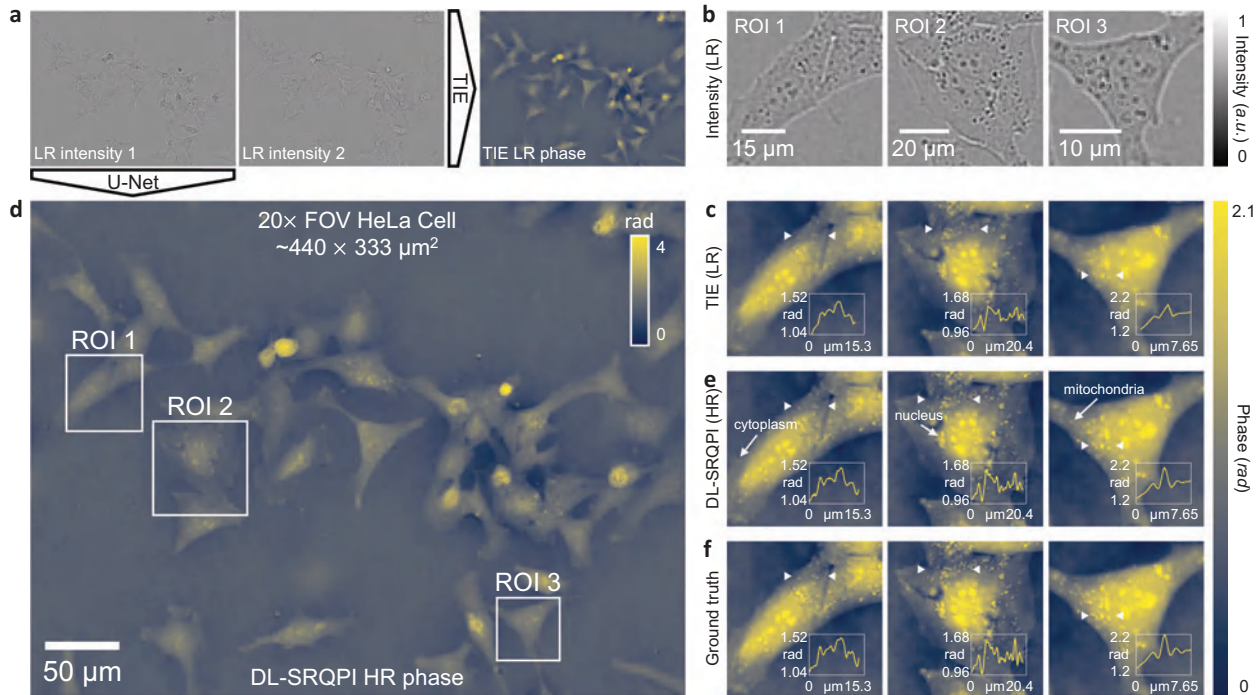
After validating the feasibility and the generalization capability of DL-SRQPI on simulated data, we further demonstrate the full-field high-resolution phase recovery from experimentally acquired intensity images of HeLa cells in vitro using DL-SRQPI. The HeLa cell image used to test the network was acquired at different culturing time point from the HeLa cell image used to construct the training dataset. As mentioned in Section 2.2, the ground truth phase images were retrieved based on TIE using the experimentally acquired high-resolution intensity images, and

the low-resolution intensity images were used as inputs of DL-SRQPI and TIE. As can be seen in Figure 8a,b, the low-resolution intensity images suffer from severe pixel aliasing problem, where the high-frequency components almost completely disappear in the extremely low contrast and oversized pixels. With the low-resolution intensity image stack as input, TIE could only provide phase image with the same low resolution (Figure 8c), creating a great obstacle to the observation of the internal structure of cells. In contrast, DL-SRQPI has a minimum requirement and possesses pixel super-resolution capability. With a single frame of low-resolution intensity as the only input, DL-SRQPI rapidly generates the high-resolution phase image with pixel resolution of  $2588 \times 1960$  pixels (Figure 8d) in 0.2 s, achieving fourfold pixel super-resolution. As shown in three enlarged ROIs in Figure 8e, DL-SRQPI enables the precise observation of plentiful subcellular features of HeLa cells, such as nucleus with large phase,



**Figure 7.** Time-lapse full-FOV high-resolution phase reconstruction of unstained HeLa cells undergoing division using DL-SRQPI. a) The full-FOV high-resolution phase reconstruction of the low-resolution intensity image by DL-SRQPI, and the comparison of low-resolution intensity images and high-resolution phase images of two ROIs. b) Sample frames of the DL-SRQPI reconstructed video (Video S2, Supporting Information) for cells A and B across 5 h, showing their different stages of cell division. c) The timeline of cell division of cells A and B within 5 h judging from DL-SRQPI reconstructed video. Different colors corresponding to (b) represent different cell division stages.





**Figure 8.** The experimental result of DL-SRQPI to reconstruct the full-FOV high-resolution phase image of unstained HeLa cells. a) The full-FOV low-resolution defocused intensity images, and the low-resolution phase image reconstructed by TIE. b) The low-resolution defocused intensity images of three ROIs. c) The low-resolution phase images of three ROIs reconstructed by TIE. d) The full-FOV high-resolution phase image reconstructed by DL-SRQPI from a single frame of low-resolution intensity image. e) The high-resolution phase images of the three ROIs reconstructed by DL-SRQPI. f) The ground truth phase images of the three ROIs.

mitochondria in transport, and cytoplasm in high contrast. The accuracy of the DL-SRQPI can be quantified by the SSIM index. The SSIM index of the full-field output phase reaches 0.9948, and the SSIM indexes of ROI 1, ROI 2, and ROI 3 are 0.9933, 0.9898, and 0.9920, which show extremely high similarity between the network outputs and the ground truth phases (Figure 8f). Compared with the TIE method that requires two intensity images at different defocus distances, the network reduces the demand for intensity images to only one, lightening the data burden as well as avoiding the phase error caused by the inaccurate estimation of axial intensity derivative. DL-SRQPI improves the resolution without sacrificing FOV, enhancing the throughput of the system effectively. Video S3, Supporting Information, shows the experimental time-lapse full-field phase reconstruction results of DL-SRQPI. These experimental results validate the efficacy and promptness of the DL-SRQPI utilization within a bright-field optical microscopy system.

#### 4. Conclusion

In this study, we have introduced DL-SRQPI, a novel deep learning-based technique for quantitative phase microscopy with pixel super-resolution capability. DL-SRQPI enables full-FOV high-resolution phase imaging of unlabeled specimens using only a single frame of low-resolution intensity image as input, as validated by experimental data. The robust generalization capability of DL-SRQPI has been demonstrated using datasets of simulated and experimental intensity images with varying defocus distances and coherent parameters, which highlights the versatil-

ity and adaptability of our approach in handling different imaging conditions. The quantitative property of DL-SRQPI has also been validated by a microlens array. Furthermore, our method exhibits notable advantages in terms of fast speed and high-throughput, as evidenced by successful phase reconstruction of unstained biological cells and dynamic phase observation of HeLa cells. These results demonstrate the feasibility of DL-SRQPI for video-rate living cell phase imaging, opening up new possibilities for real-time cellular dynamics analysis. Importantly, DL-SRQPI significantly reduces the need for intensity data redundancy compared to conventional QPI approaches, thereby mitigating the trade-off between FOV and resolution. This advancement enhances the SBP of fundamental bright-field system equipment, facilitating more efficient and cost-effective imaging capabilities. Overall, DL-SRQPI showcases its potential as a powerful QPI technique with wide-ranging applications in high-throughput microscopy. It holds promise for various fields such as drug discovery, cellular phenotype characterization, and the identification of disease mechanisms.<sup>[58]</sup> Future investigations will delve into understanding the dependencies of DL-SRQPI on specific cell types and culture conditions, as well as addressing the impact of various experimental configurations on phase retrieval.

#### 5. Experimental Section

**Sample Preparation:** To prepare biological material, the HeLa cells were cultivated in a glass bottom Petri dish (35 mm, MatTek) with L-glutamine Dulbecco's modified Eagles medium (Gibco, American)



supplemented with 10% Nu-serum (Corning, American), 10% fetal calf serum (Gibco, Australia), and 1% vitamin mix (100×) (Lonza, Cologne, Germany). The cells were cultured in a stage-mounted climate chamber (Tokai Hit INUF-IX3W, Japan) for stabilization of temperature at 37 °C and CO<sub>2</sub> gas at 5%. The medium was changed every other day and cells were passed with trypsin upon reaching 80% confluency. In preparation for cell division imaging, cells were washed once with phosphate-buffered saline and detached with either accutase or trypsin.

**Experimental Configuration:** For experimental measurements of HeLa cells, the intensity images used for TIE phase retrieval were obtained with an optical system equipped with an inverted microscope (IX71, Olympus, Japan), utilizing a halogen white-light source with a green interference filter (central wavelength  $\lambda = 550$  nm, 45-nm bandwidth) for illumination. The microscope was equipped with a 5-megapixel charge-coupled device (CCD) camera (UC50, Olympus, Germany) with a pixel resolution of 2588 × 1960 and a pixel pitch of 3.4  $\mu\text{m}$ . The microscope also included an electrically tunable lens (EL-C-10-30-VISLD, Optotune AG, Switzerland) module that was synchronized with the camera at different focal distances along the z-axis and controlled by software via a standard USB cable. The image stack was acquired via plan semiapochromat objective (LUCPlanFLN 40×, Olympus, Half magnification, NA 0.6) in an 8-bit grayscale range.

The microlens array (SUSS MicroOptics pitch 240ROC 297  $\mu\text{m}$ ) was imaged using an inverted bright-field microscope (Olympus IX71), and in-focus and out-of-focus intensity images were captured by axially translating the camera. The camera used here had a pixel size of 2.2  $\mu\text{m}$  (The Imaging Source DMK 72BUC02, 1280 × 960 resolution), and the illumination was set to a central wavelength of 550 nm.

**Implementation Details:** The deep neural network was implemented using Pytorch 1.10.2 based on Python 3.9.7. The network training and testing were performed on a workstation with Intel(R) Core (TM) i9-10900K CPU (3.70 GHz) and 32 GB of RAM, running a Windows 10 operating system (Microsoft) using NVIDIA GeForce RTX 3090 GPU. With a batch size of eight, the training process of each dataset took  $\approx 1$  h for 200 epochs. The network took  $\approx 0.3$  s to reconstruct a full-field 2048 × 2048 pixels phase image from the 512 × 512 pixels simulated intensity image, and took  $\approx 0.2$  s to reconstruct a full-field 2588 × 1960 pixels phase image from the 647 × 490 pixels experimentally acquired intensity image.

## Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (62105151, 62175109, U21B2033, 62227818), the Leading Technology of Jiangsu Basic Research Plan (BK20192003), the Youth Foundation of Jiangsu Province (BK20210338), the Biomedical Competition Foundation of Jiangsu Province (BE2022847), the Key National Industrial Technology Cooperation Foundation of Jiangsu Province (BZ2022039), Fundamental Research Funds for the Central Universities (30920032101), the Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense (JSGP202105, JSGP202201), and the Postgraduate Research Practice Innovation Program of Jiangsu Province (122104010479).

## Conflict of Interest

The authors declare no conflict of interest.

## Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## Keywords

deep learning, high-throughput microscopy, phase retrieval, quantitative phase imaging, super-resolution

Received: May 29, 2023

Revised: July 30, 2023

Published online: September 17, 2023

- [1] J. Mertz, *Introduction to Optical Microscopy*, Cambridge University Press, Cambridge 2019.
- [2] G. Popescu, *Quantitative Phase Imaging of Cells and Tissues*, McGraw-Hill Education, New York 2011.
- [3] M. G. Gustafsson, *J. Microsc.* **2000**, *198*, 82.
- [4] S. W. Hell, J. Wichmann, *Opt. Lett.* **1994**, *19*, 780.
- [5] P. Gao, B. Prunsche, L. Zhou, K. Nienhaus, G. U. Nienhaus, *Nat. Photonics* **2017**, *11*, 163.
- [6] E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacio, M. W. Davidson, J. Lippincott-Schwartz, H. F. Hess, *Science* **2006**, *313*, 1642.
- [7] R. H. Webb, *Rep. Prog. Phys.* **1996**, *59*, 427.
- [8] D. J. Stephens, V. J. Allan, *Science* **2003**, *300*, 82.
- [9] Y. Cotte, F. Toy, P. Jourdain, N. Pavillon, D. Boss, P. Magistretti, P. Marquet, C. Depeursinge, *Nat. Photonics* **2013**, *7*, 113.
- [10] P. Gao, C. Yuan, *Light Adv. Manuf.* **2022**, *3*, 105.
- [11] Z. Huang, P. Memmolo, P. Ferraro, L. Cao, *PhotonIX* **2022**, *3*, 3.
- [12] J. Li, A. Matlock, Y. Li, Q. Chen, C. Zuo, L. Tian, *Adv. Photonics* **2019**, *1*, 066004.
- [13] J. Li, N. Zhou, J. Sun, S. Zhou, Z. Bai, L. Lu, Q. Chen, C. Zuo, *Light Sci. Appl.* **2022**, *11*, 154.
- [14] S. Zhou, J. Li, J. Sun, N. Zhou, H. Ullah, Z. Bai, Q. Chen, C. Zuo, *Optica* **2022**, *9*, 1362.
- [15] V. Mico, Z. Zalevsky, J. Garcia, *Opt. Commun.* **2008**, *281*, 4273.
- [16] Y. Park, C. Depeursinge, G. Popescu, *Nat. Photonics* **2018**, *12*, 578.
- [17] P. Gao, B. Yao, I. Harder, N. Lindlein, F. J. Torcal-Milla, *Opt. Lett.* **2011**, *36*, 4305.
- [18] Y. Fan, J. Li, L. Lu, J. Sun, Y. Hu, J. Zhang, Z. Li, Q. Shen, B. Wang, R. Zhang, et al., *PhotonIX* **2021**, *2*, 19.
- [19] D. Dong, X. Huang, L. Li, H. Mao, Y. Mo, G. Zhang, Z. Zhang, J. Shen, W. Liu, Z. Wu, *Light Sci. Appl.* **2020**, *9*, 11.
- [20] W. Choi, C. Fang-Yen, K. Badizadegan, S. Oh, N. Lue, R. R. Dasari, M. S. Feld, *Nat. Methods* **2007**, *4*, 717.
- [21] Y. Sung, W. Choi, C. Fang-Yen, K. Badizadegan, R. R. Dasari, M. S. Feld, *Opt. Express* **2009**, *17*, 266.
- [22] M. R. Teague, *J. Opt. Soc. Am.* **1983**, *73*, 1434.
- [23] G. Zheng, R. Horstmeyer, C. Yang, *Nat. Photonics* **2013**, *7*, 739.
- [24] C. Zuo, J. Sun, J. Li, J. Zhang, A. Asundi, Q. Chen, *Sci. Rep.* **2017**, *7*, 7654.
- [25] J. Park, D. J. Brady, G. Zheng, L. Tian, L. Gao, *Adv. Photonics* **2021**, *3*, 044001.
- [26] G. Zheng, C. Shen, S. Jiang, P. Song, C. Yang, *Nat. Rev. Phys.* **2021**, *3*, 207.
- [27] J. Sun, Q. Chen, Y. Zhang, C. Zuo, *Opt. Express* **2016**, *24*, 15765.
- [28] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.
- [29] Y. Li, Y. Xue, L. Tian, *Optica* **2018**, *5*, 1181.
- [30] Y. Rivenson, Z. Göröcs, H. Günaydin, Y. Zhang, H. Wang, A. Ozcan, *Optica* **2017**, *4*, 1437.
- [31] H. Wang, Y. Rivenson, Y. Jin, Z. Wei, R. Gao, H. Günaydin, L. A. Bentolila, C. Kural, A. Ozcan, *Nat. Methods* **2019**, *16*, 103.
- [32] Y. Rivenson, H. Wang, Z. Wei, K. de Haan, Y. Zhang, Y. Wu, H. Günaydin, J. E. Zuckerman, T. Chong, A. E. Sisk, *Nat. Biomed. Eng.* **2019**, *3*, 466.

- [33] T. Nguyen, Y. Xue, Y. Li, L. Tian, G. Nehmetallah, *Opt. Express* **2018**, 26, 26470.
- [34] Y. Rivenson, Y. Zhang, H. Günaydin, D. Teng, A. Ozcan, *Light Sci. Appl.* **2018**, 7, 17141.
- [35] K. Wang, J. Di, Y. Li, Z. Ren, Q. Kema, J. Zhao, *Opt. Lasers Eng.* **2020**, 134, 106233.
- [36] C. Zheng, D. Jin, Y. He, H. Lin, J. Hu, Z. Yaqoob, P. T. So, R. Zhou, *Adv. Photonics* **2020**, 2, 065002.
- [37] X. Chang, L. Bian, J. Zhang, *eLight* **2021**, 1, 4.
- [38] J. Sun, C. Zuo, L. Zhang, Q. Chen, *Sci. Rep.* **2017**, 7, 7654.
- [39] P. Stępień, D. Korbuszewski, M. Kujawińska, *ETRI J.* **2019**, 41, 73.
- [40] Y. Fan, J. Sun, Y. Shu, Z. Zhang, G. Zheng, W. Chen, J. Zhang, K. Gui, K. Wang, Q. Chen, C. Zuo, *Laser Photonics Rev.* **2023**, 17, 2200201.
- [41] W. Bishara, T.-W. Su, A. F. Coskun, A. Ozcan, *Opt. Express* **2010**, 18, 11181.
- [42] S. Jiang, J. Zhu, P. Song, C. Guo, Z. Bian, R. Wang, Y. Huang, S. Wang, H. Zhang, G. Zheng, *Lab Chip* **2020**, 20, 1058.
- [43] X. Chang, S. Jiang, Y. Hu, G. Zheng, L. Bian, *ACS Photonics* **2023**.
- [44] S. Jiang, C. Guo, P. Song, N. Zhou, Z. Bian, J. Zhu, R. Wang, P. Dong, Z. Zhang, J. Liao, *ACS Photonics* **2021**, 8, 3261.
- [45] C. Zuo, J. Sun, Q. Chen, *Opt. Express* **2016**, 24, 20724.
- [46] Y. Shu, J. Sun, J. Lyu, Y. Fan, N. Zhou, R. Ye, G. Zheng, Q. Chen, C. Zuo, *PhotonIX* **2022**, 3, 24.
- [47] L. Lu, J. Li, Y. Shu, J. Sun, J. Zhou, E. Y. Lam, Q. Chen, C. Zuo, *Adv. Photonics* **2022**, 4, 056002.
- [48] M. Born, E. Wolf, H. Haubold, *Astron. Nachr.* **1980**, 301, 257.
- [49] C. Zuo, J. Li, J. Sun, Y. Fan, J. Zhang, L. Lu, R. Zhang, B. Wang, L. Huang, Q. Chen, *Opt. Lasers Eng.* **2020**, 135, 106187.
- [50] J. Sun, C. Zuo, J. Zhang, Y. Fan, Q. Chen, *Sci. Rep.* **2018**, 8, 7669.
- [51] J. Li, Q. Chen, J. Sun, J. Zhang, C. Zuo, *J. Biomed. Opt.* **2016**, 21, 126003.
- [52] L. Waller, L. Tian, G. Barbastathis, *Opt. Express* **2010**, 18, 12552.
- [53] D. Paganin, K. A. Nugent, *Phys. Rev. Lett.* **1998**, 80, 2586.
- [54] T. Falk, D. Mai, R. Bensch, Ö. Çiçek, A. Abdulkadir, Y. Marrakchi, A. Böhm, J. Deubner, Z. Jäckel, K. Seiwald, *Nat. Methods* **2019**, 16, 67.
- [55] X. Yang, L. Huang, Y. Luo, Y. Wu, H. Wang, Y. Rivenson, A. Ozcan, *ACS Photonics* **2021**, 8, 2174.
- [56] C. Zuo, Q. Chen, A. Asundi, *Opt. Express* **2014**, 22, 9220.
- [57] C. Dong, C. C. Loy, K. He, X. Tang, *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, 38, 295.
- [58] Z. Chen, M. Segev, *eLight* **2021**, 1, 2.

DOI: [10.29026/oes.2023.220023](https://doi.org/10.29026/oes.2023.220023)

# Deep learning assisted variational Hilbert quantitative phase imaging

Zhuoshi Li<sup>1,2,3</sup>, Jiasong Sun<sup>1,2,3</sup>, Yao Fan<sup>1,2,3</sup>, Yanbo Jin<sup>1,2,3</sup>, Qian Shen<sup>1,2,3</sup>, Maciej Trusiak<sup>4</sup>, Maria Cywińska<sup>4</sup>, Peng Gao<sup>5\*</sup>, Qian Chen<sup>3\*</sup> and Chao Zuo<sup>1,2,3\*</sup>

We propose a high-accuracy artifacts-free single-frame digital holographic phase demodulation scheme for relatively low-carrier frequency holograms—deep learning assisted variational Hilbert quantitative phase imaging (DL-VHQPI). The method, incorporating a conventional deep neural network into a complete physical model utilizing the idea of residual compensation, reliably and robustly recovers the quantitative phase information of the test objects. It can significantly alleviate spectrum-overlapping-caused phase artifacts under the slightly off-axis digital holographic system. Compared to the conventional end-to-end networks (without a physical model), the proposed method can reduce the dataset size dramatically while maintaining the imaging quality and model generalization. The DL-VHQPI is quantitatively studied by numerical simulation. The live-cell experiment is designed to demonstrate the method's practicality in biological research. The proposed idea of the deep learning-assisted physical model might be extended to diverse computational imaging techniques.

**Keywords:** quantitative phase imaging; digital holography; deep learning; high-throughput imaging

Li ZS, Sun JS, Fan Y, Jin YB, Shen Q et al. Deep learning assisted variational Hilbert quantitative phase imaging. *Opto-Electron Sci* 2, 220023 (2023).

## Introduction

Quantitative phase imaging (QPI), as a powerful label-free imaging technique, enables dynamic 2D and 3D non-destructive imaging of completely transparent structures<sup>1–3</sup>. It uses the refractive index as an endogenous contrast agent to generate subcellular-specific quantitative maps of analyzed live bio-structure<sup>4,5</sup>. QPI solutions based on digital holographic microscopy (DHM) encode a complex wavefront information into intensity

modulations by the interference of a scattered sample wave and a reference wave<sup>6–9</sup>. And it can robustly perform the quantitative analysis of wave-matter interactions by decoding phase delay from a hologram. DHM has emerged as a valuable means in the biomedical fields, such as measurements for stain-free biological cells<sup>3,10</sup>, optical metrology of nanostructures<sup>11–14</sup>, and drug release monitoring *in vitro*<sup>15</sup>.

Regarding the phase demodulation strategy employed,

<sup>1</sup>Smart Computational Imaging Laboratory (SCILab), School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China; <sup>2</sup>Smart Computational Imaging Research Institute (SCIRI) of Nanjing University of Science and Technology, Nanjing 210094, China; <sup>3</sup>Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense, Nanjing 210094, China; <sup>4</sup>Institute of Micromechanics and Photonics, Warsaw University of Technology, 8 Sw. A. Boboli St., Warsaw 02-525, Poland; <sup>5</sup>School of Physics, Xidian University, Xi'an 710126, China.

\*Correspondence: P Gao, E-mail: [peng.gao@xidian.edu.cn](mailto:peng.gao@xidian.edu.cn); Q Chen, E-mail: [chenqian@njust.edu.cn](mailto:chenqian@njust.edu.cn); C Zuo, E-mail: [zuochao@njust.edu.cn](mailto:zuochao@njust.edu.cn)

Received: 24 November 2022; Accepted: 3 March 2023; Published online: 18 May 2023



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License.

To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023. Published by Institute of Optics and Electronics, Chinese Academy of Sciences.

there are two main configurations for holographic wavefront acquisition in DHM, *i.e.*, in-line and off-axis digital holography (DH). In-line DH records complete wavefront information by the interference of the object light and the reference light on the same optical axis, which can realize full detector-bandwidth phase reconstruction. However, due to the superimposed twin image, the phase retrieval results of samples are severely impacted by imaging artifacts. It always needs to be processed via iterative phase retrieval<sup>16,17</sup> or noniterative phase-shifting methods<sup>18–20</sup>, which dramatically sacrifices the temporal resolution. Therefore, it is difficult for the in-line DH, which is vulnerable to external disturbance and vibration, to be applied to dynamic measurement. Alternatively, off-axis DH implements twin-image separation by introducing a slight angle between the object beam and reference beam and recovers the complex wavefront of the sample from the single-frame off-axis hologram. Whereas, for achieving the separation of autocorrelation and cross-correlation terms in the spatial frequency domain (SFD), the off-axis DH needs to provide a sufficiently high carrier frequency at the expense of the space-bandwidth product (SBP) of the imaging system<sup>21</sup>. The slightly off-axis DH regime, as a single-frame high-SBP DH imaging solutions, is therefore proposed<sup>22–24</sup>. It optimizes SBP through full spectral separation of conjugated object lobes while leaving the autocorrelation term partially overlapped with information-carrying cross-correlation terms. Under this configuration, the inevitable spectrum overlapping causes phase artifacts, which greatly degrades the imaging quality and impairs the practicality of the slightly off-axis DH configuration.

High-accuracy artifacts-free phase recovery from the low-carrier frequency holograms is the key to slightly off-axis DH application. This process is presently implemented by suppressing autocorrelation term iteratively<sup>25</sup>, utilizing dual-frame decoding scheme<sup>26,27</sup>, employing second wavelength assistance<sup>28</sup> and performing the 1D limited processing<sup>29,30</sup>. With inspiration from the theory of “*cepstrum*” and homomorphic filtering<sup>31</sup>, a slightly off-axis DH demodulation scheme based on the Kramers-Kronig (KK) relations is proposed, which utilizes the half-space bandwidth of the sensor to achieve high-SBP imaging<sup>32,33</sup>. Although it is able to increase the SBP of full complex field recovery significantly, it inevitably requires intensity restrictions on the object and reference beams and the separation of the cross-correlation terms of the interferogram in the extended SFD. Noteworthy,

an exquisite low-carrier frequency fringe demodulation approach has been presented recently, namely variational Hilbert quantitative phase imaging (VHQPI)<sup>34</sup>. The VHQPI, as an end-to-end pure numerical add-on module, deploys the merger of tailored variational image decomposition<sup>35</sup> and enhanced Hilbert spiral transform<sup>36</sup> to achieve quantitative phase recovery. It adaptively alleviates the overlapped-spectrum problem and robustly demodulates high-quality phase information, performing excellent practicality in biological applications.

Although VHQPI has demonstrated excellent low-carrier frequency fringe demodulation capability, the algorithm-inherent limitations (*e.g.*, parameter robustness and iterative stability) still cause non-sufficient image frequency component extraction, resulting in imaging artifacts in the phase reconstruction results. Deep learning (DL), as a subfield of machine learning, has currently gained extensive attention in the field of optical metrology and demonstrated great potential in solving optical metrology tasks<sup>37–46</sup>. When sufficient training data is collected in an environment that reproduces real experimental conditions, the trained model may have advantages over physics-model-based approaches on some issues (*e.g.*, computing speed, parameter adaptivity, algorithm complexity)<sup>37</sup>. Specifically, in terms of a series of ill-posed inverse phase retrieval problems, the traditional physical model tends to exhibit higher physics complexity and time consumption. Driven by a large dataset, the deep neural network (DNN) can directly and efficiently reconstruct the phase and amplitude images of the objects from the captured holograms<sup>47–49</sup>. Nevertheless, in DL-based phase recovery tasks, it is pretty tricky and laborious to capture massive datasets and generate the corresponding ground truth, especially when applied to bio-samples. Deep image prior (DIP) applies an untrained network to the solution of several inverse problems without a massive training dataset and ground truth, which can fit a randomly initialized DNN to a single corrupted image<sup>50</sup>. Inspired by the DIP, an untrained network model named “PhySenNet” is proposed, which incorporates a complete physical model into the conventional DNN to achieve phase retrieval from a single intensity image<sup>51</sup>.

Inspired by the successful application of the interplay between DNN and the physical model, in this work, we propose a DL-assisted variational Hilbert quantitative phase imaging approach (DL-VHQPI). Unlike the massive-data-driven DL training model, DL-VHQPI,



which utilizes DNN to compensate and optimize the possible solutions of the physics-driven model, can achieve high-precision artifacts-free phase recovery using only a small fraction datasets. Specifically, VHQPI, as the underlying physical model, can complete the preliminary extraction of the background components of the fringes to provide a physical prior for the deep learning model. The DNN compensates for the image frequencies that cannot be extracted by the physical model using the idea of residual compensation. Due to the physical model reducing the information entropy of the dataset, the DL-VHQPI performs higher reconstruction accuracy utilizing less than one-tenth of the dataset of the conventional end-to-end model (without the physical model). The simulation experiments quantitatively demonstrate that the proposed method can achieve high-accuracy artifacts-free quantitative phase imaging from single-frame low-carrier frequency holograms. And the results of live-cell experiments demonstrate the practicality of the method in biological research.

## Principle of VHQPI

The VHQPI, as the physical model of the DL-VHQPI, adaptively and effectively completes the low-carrier frequency fringe demodulation employing the unsupervised variational image decomposition (uVID) and enhanced Hilbert spiral transform (HST). This section will focus on describing the process details and physical limitations of this method. In the DH wavefront recording, the interferogram containing the required object information is constructed upon the coherent superimposition of the object and reference beams. The intensity distribution of the recorded hologram can be expressed as:

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos(\theta) + n = a + b \cos(\theta) + n. \quad (1)$$

It consists of a sum of three fundamental intensity components: background ( $a$ , incoherent sum of intensities  $I_1$  and  $I_2$  of interfering beams), high-frequency noise ( $n$ ), and coherent interference fringes term comprised by a cosine function modulated in phase ( $\theta$ ) and amplitude ( $b$ ,  $2\sqrt{I_1 I_2}$ ). Acquiring the accurate fringes term from the three components is the prerequisite of high-precision artifacts-free phase recovery. The uVID approach achieves image frequency components extraction, which is based on the notion of the classical variational image decomposition to separate the information components of the image with two steps in terms of methodology<sup>52,53</sup>: 1) A block-matching 3D (BM3D) algorithm is employed

to remove noise with remarkable efficiency<sup>54</sup>; 2) Background-fringes differentiation is performed using modified Chambolle projection algorithm with an automatic stopping criterion to set the number of projections, and there is no need to pre-set any parameter values<sup>55</sup>. The based-on uVID image frequency components extraction process is shown in Step 1 of Fig. 1. Although the uVID provides a robust and automatic one-stop-shop solution for single-frame fringe pattern analysis, there are physical limitations in the process of frequency component extraction, i.e., iterative instability and parameter robustness, which directly cause non-sufficient background term removal and then impair phase recovery accuracy and artifacts-suppression effect<sup>52</sup>.

To recover the phase information of the object, the uVID-filtered noise-free zero-mean-valued interferogram is then analyzed using the HST algorithm<sup>36</sup>, as shown in Step 2 of Fig. 1. The HST is the two-dimensional variant of the Hilbert transform (HT), in which the complex analytic signal can be constructed, whereas several requirements must be fulfilled. First, the processed interferogram must be of zero mean value, which is satisfied based on background term removal using the uVID approach. And the amplitude term ( $b$  in Eq. (1)) has to be a slowly varying function. This is the so-called Bedrosian theorem which can be applied to general pure-phase objects at relatively low carrier frequencies<sup>55</sup>. The complex analytic signal constructed by HST can be expressed as

$$AFP = 2\sqrt{I_1 I_2} \cos(\theta) - i \exp(-i\beta) \cdot \mathcal{F}^{-1}\{SPF * \mathcal{F}[2\sqrt{I_1 I_2} \cos(\theta)]\}, \quad (2)$$

where,  $AFP$  denotes the analytic fringe pattern and  $SPF$  is the spiral phase function;  $\mathcal{F}$  and  $\mathcal{F}^{-1}$  denote Fourier transform (FT) and inverse Fourier transform (IFT) operator respectively. It is important to emphasize that carrier-free single-shot interferogram analysis is a fully 2D phase demodulation problem, whereas carrier-based FT phase demodulation is a 1D simplification of the HT analytic relation. The HST, therefore, requires the local fringe direction map ( $\beta$ , modulo  $2\pi$ )<sup>56</sup>. The modulus value and angle of the 2D complex analytic signal constitute the intensity and phase in QPI, respectively.  $SPF$  is defined as

$$SPF(u, v) = \frac{u + i \cdot v}{\sqrt{u^2 + v^2}} = \exp[i \cdot \phi(x, y)], \quad (3)$$

where  $(u, v)$  is the coordinate of  $(x, y)$  corresponding to the SFD.  $\phi(x, y)$  is the polar coordinate phase expression.

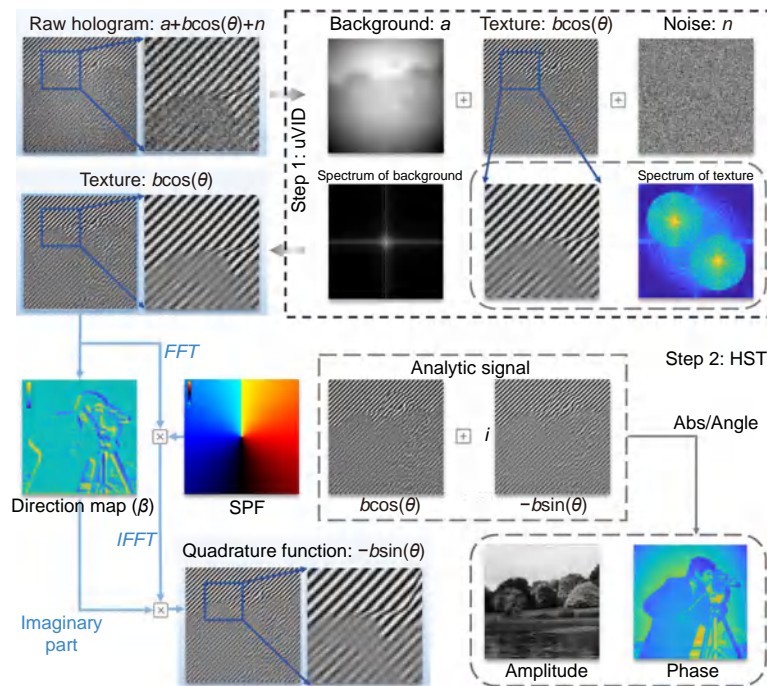


Fig. 1 | Flow chart of slightly off-axis interferometric fringe demodulation based on VHQPI.

Figure 1 specifically showcases the schematic diagram of the VHQPI-based low-carrier interferogram quantitative phase demodulation algorithm.

### Deep learning assisted VHQPI model

VHQPI has been proven to have excellent robustness and practicality in low-carrier frequency fringe demodulation issue though<sup>34</sup>. However, the algorithm-inherent iterative instability and parameter robustness restrict the image frequency component extraction capability, which will cause the non-perfect background term removal. DL methods driven by massive datasets provide a new route to address this problem by virtue of their high-powerful image feature extraction characteristics. Whereas, when encountering insufficient training data, which is very common, the DL method based on massive datasets may have a poor effect. A feasible scheme is to train the DNN on a stronger-constrained available standardized dataset<sup>57</sup>. Here, we employ Shannon entropy theory of the images in the dataset for that purpose: the lower the entropy of the datasets is, the more constrained prior information is, giving it a better same-domain generalization ability<sup>58,59</sup>. Therefore, in the proposed DL-assisted VHQPI model, the uVID is utilized to extract the image background term as the physical prior of the network to reduce the dataset's entropy. The first convolutional neural network (CNN1) is used to “learn” the residual terms and assists the physical model to complete the pre-

liminary estimation of the background components of the fringes. Furthermore, to further improve the imaging accuracy, the original hologram and the preliminary estimation background are re-fed into the model (CNN2) for advanced component extraction. Dual-channel input is used because the preliminarily estimated background terms have been very close to the ground truth after the first residual compensation by CNN1. Hence, the preliminarily estimated background can be used to provide the network with feature guidance and helps CNN2 achieve the advanced component extraction.

As depicted in Fig. 2, with the original hologram as input, CNN1 completes the preliminary background component extraction by compensating for the residual ( $\varepsilon_1$ ) of the background component acquired by uVID, as shown in Fig. 2(b). With the preliminarily estimated background term and the original hologram, the CNN2 (as shown in Fig. 2(c)) uses the two as dual-channel inputs to implement the more advanced background residual ( $\varepsilon_2$ ) compensation. After the high-accuracy fringes terms extraction, the complete complex analytic signal can be constructed by HST. And then the final phase results are recovered by calculating the angle of the 2D complex analytic signal. The whole method flow chart is shown in Fig. 2(a).

Moreover, both CNN1 and CNN2 networks are composed of a convolutional layer (Conv), a group of residual blocks (containing four residual blocks), and two

convolutional layers. Each residual block comprises two sets of Convs stacked one above the other. The network architecture uses Batch Normalization<sup>60</sup> and ReLU activation<sup>61</sup> to accelerate the model convergence. It establishes a shortcut between input and output, which can solve the problem of accuracy decline as the network deepens, thereby easing the training process. The output of the Convs is a 3-D tensor of shape  $(H, W, C)$ , where  $H$  and  $W$  are the height and width of pixels of the hologram respectively, and  $C$  is the number of channels. The hyperparameters of the two networks, *i.e.*, the weights, bias, and convolutional kernels, are trained using back-propagation on mean-squared errors between the results of the network output and the ground truth. The loss function is computed as

$$\text{Loss}(\omega) = \frac{1}{H \times W} \|Y_{\text{output}}^{\omega} - G\|^2, \quad (4)$$

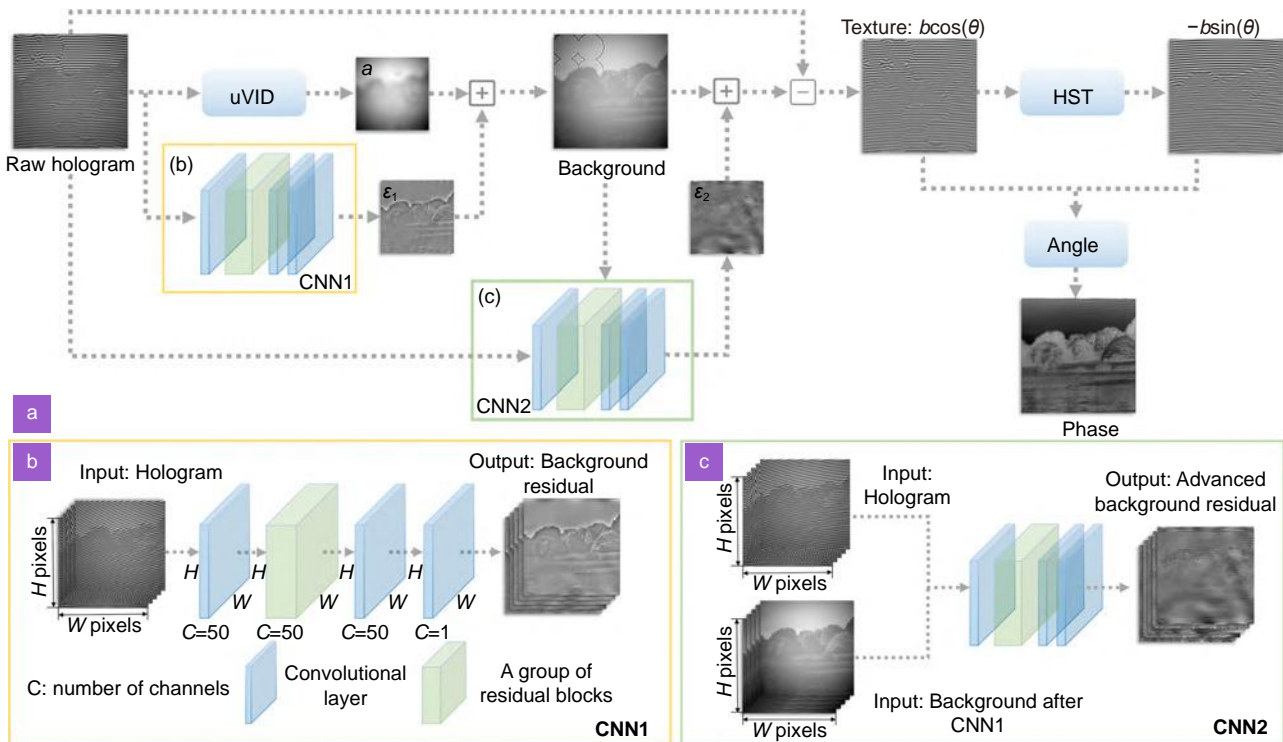
where  $\omega$  represents the parameter space of the model,  $Y_{\text{output}}$  is the results predicted by the model, and  $G$  is the ground truth.

## Experiments and results

In this section, we demonstrate the performance of the

proposed DL-VHQPI method over the conventional physics-driven low-carrier frequency fringe demodulation techniques and pure DL approach without a physical model (DL-noPhy) through numerical simulation and live-cell experiment. A rich set of paired training data is the prerequisite for network generalization during DL training. It is challenging to acquire a reliable ground truth in the real-world DH system due to environment-induced instability and system-inherent speckle noise. Consequently, we simulated low-carrier frequency holograms and the corresponding ground truth for training and quantitative analysis. We separately constructed the complex amplitude distributions of the object and reference light waves, and then the holograms can be constructed by solving the square of the modulus of the sum of the two. The sum of the squares of the modulus values of the two was calculated to obtain the background (ground truth) needed for training. The more specific process can be found in Supplementary information Section 1.

In the live-cell experiment, we used the Digital holographic smart computational light microscope (DH-SCLM) developed by SCILab, and turned it to a slightly off-axis state for hologram acquisition<sup>1</sup>. In the DH-



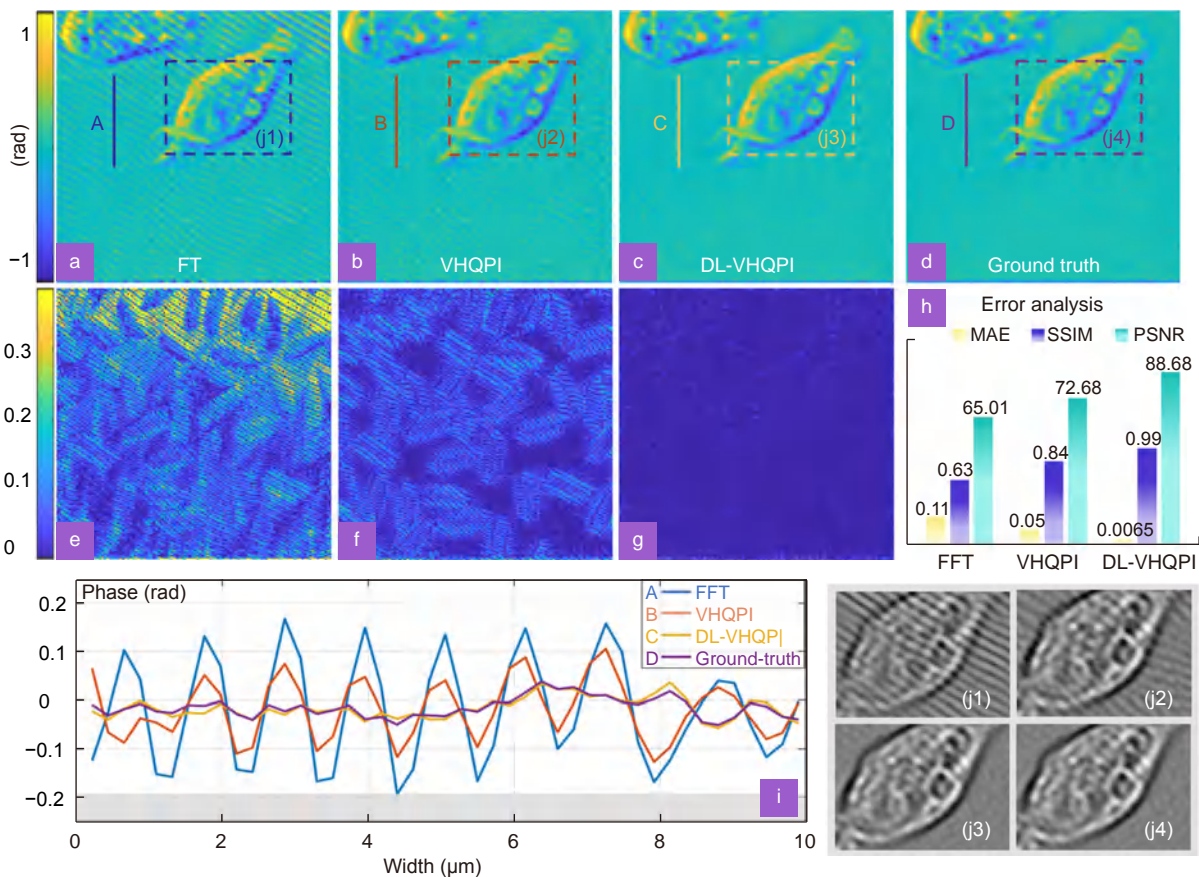
**Fig. 2 | Deep learning-assisted VHQPPI.** (a) Total network structure, combining uVID and HST with CNN respectively for phase reconstruction. (b) CNN1 takes a hologram as input and consists of three convolutional layers and a group of residual blocks to achieve compensation of background residuals by learning. (c) The CNN2 network structure is the same as CNN1, except that CNN2 combines the original hologram and the result of the first process into a two-channel input for advanced background compensation.



SCLM, the object wave transmitting the objective lens (UPLanSAPO  $\times 20/0.45\text{NA}$ , Olympus, Japan) interferes with the reference light and is recorded by the camera (The Imaging Source DMK 23U274,  $1600 \times 1200$ ,  $4.4 \mu\text{m}$ ). The central wavelength of the illumination is  $532 \text{ nm}$ . The used sample is Henrietta Lacks (HeLa) human cervical cancer cells cultured in DMEM medium with 10% fetal bovine serum under standard cell culture conditions ( $37.2 \text{ }^\circ\text{C}$  in 5%  $\text{CO}_2$  in a humidified incubator). To acquire the ground truth from the configuration, each intensity map of the object and reference light paths needs to be captured separately under the highly stable condition of the holographic system (Refer to Section 2 of the Supplementary Information for detailed processing). The complete training process was implemented using the TensorFlow framework (Google) and was computed on a GTX Titan graphics card (NVIDIA). A fixed learning rate of 0.0001 for the experiment is adopted for the Adam optimizer<sup>62</sup>.

## Simulation

Figure 3 presents the experimental results under the numerical simulation, demonstrating the quantitative analysis between DL-VHQPI and the conventional single-frame fringe demodulation techniques. Figure 3(a) shows the phase result recovered by the conventional Fourier transform (FT) method. It can be seen that the phase artifacts severely disturb imaging results due to the spectrum-overlapped problem in the SFD. Although reducing the filtering window size can attenuate the phase artifacts, this will sacrifice the SBP of the system while causing blurred imaging. More details about it can be found in Supplementary Section 3. The size of the filtering window used in the FT-based phase reconstruction results shown in Fig. 3(a) is calculated under the simulated numerical aperture (NA), as shown in the red filtering window in Fig. S2(a) of Supplementary Section 3. In VHQPI, the uVID can extract the fringes term from the hologram; however, the non-perfect background term



**Fig. 3 | The experiment results under the numerical simulation.** (a) The FT method phase recovery result. (b) The phase recovery result of VHQPI. (c) The phase result reconstructed by DL-VHQPI. (d) The ground truth. (e–g) The difference between the phase results of the three methods (*i.e.* FT, VHQPI, DL-VHQPI) and the ground truth. (h) Quantitative error analysis of three methods. (i) The cross-section of the phase results of FT, VHQPI, DL-VHQPI, and ground truth, and (j1–j4) are the DIC views of the partially enlarged views of their corresponding phase maps respectively.



removal still inevitably causes imaging artifacts, as shown in Fig. 3(b). In contrast, as presented in Fig. 3(c), the experimental results demonstrate DL-VHQPI's excellent performance, in terms of artifacts suppression, over two physics-driven methods: FT and VHQPI. Figure 3(d) is the ground truth recovered by phase recovery after theoretical background removal. The magnified views of the corresponding rectangular boxes in Fig. 3(a–d), as shown in Fig. 3(j1–j4), are the phase gradient images by digital differential interference contrast (DIC) for them. To discuss the performance of methods intuitively and quantitatively, we respectively calculated the Mean Absolute Error (MAE), Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR) between the FT, VHQPI, and DL-VHQPI phase recovery results and the ground truth. Compared with the FT and VHQPI methods, Fig. 3(h) quantitatively demonstrates that DL-VHQPI has an excellent phase recovery accuracy and artifacts-suppression effect (More than 10 times improvement in precision.). The background-part cross-section of the four phase results depicted in Fig. 3(i) shows the phase result reconstructed by DL-VHQPI has a higher similarity to the ground truth, which also demonstrates that it can be more effective in suppressing the fringe-like error of the background part.

In addition, we also designed a comparison experiment with DL-noPhy (The specific network is provided in the Section 4 of Supplementary information) to demonstrate the high-efficiency and high-accuracy characteristics exhibited by the proposed method. Table 1 quantitatively shows the comparison results of the DL-VHQPI and DL-noPhy; DL-VHQPI performs a higher phase reconstruction accuracy while only utilizing one-tenth of the datasets of DL-noPhy. The reason is that DL-VHQPI adopts a physical model (uVID) to the background-component extraction process of the fringe pattern and acquires the residual components for training, which is inherently a process of image entropy reduction. According to the Shannon entropy theory, lower image entropy implies more image constraints, which provides

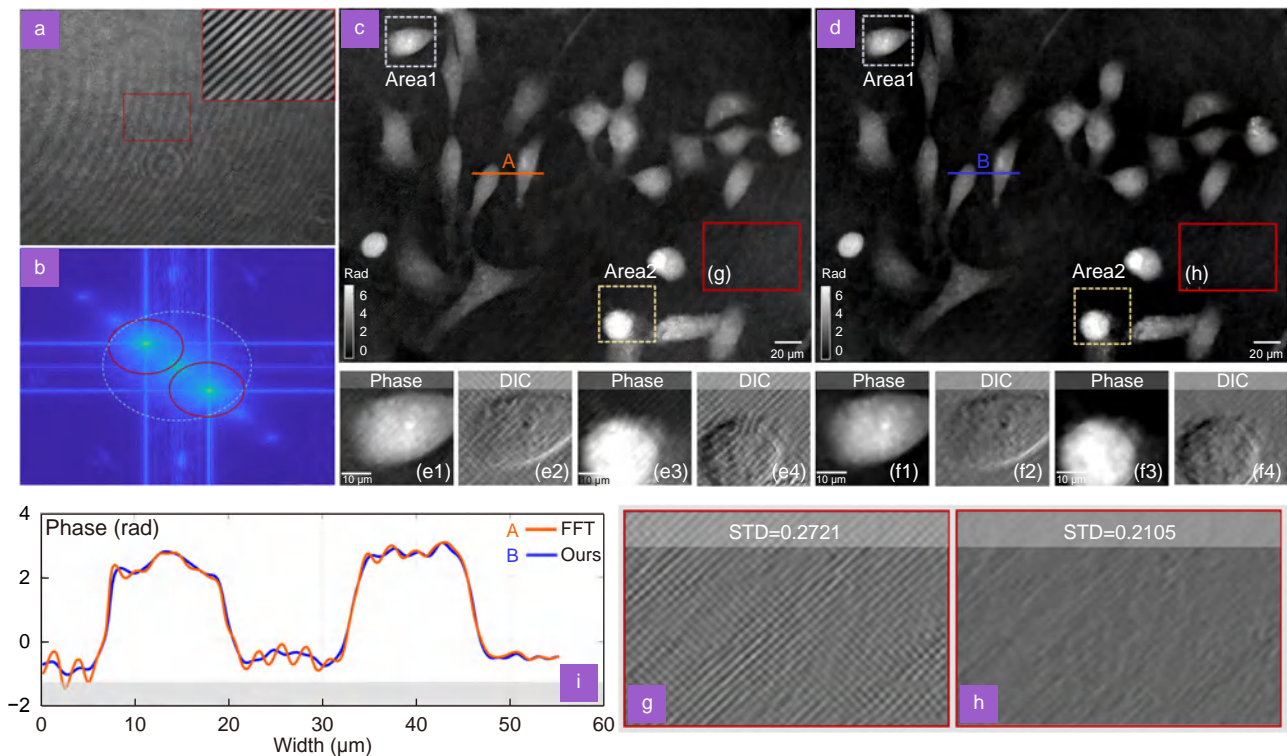
DNN with a more powerful same-domain generalization ability. The simulated holograms with the size of  $160 \times 160$  were fed to the network. During the training process, the CNN1 of DL-VHQPI over 150 epochs took 1 hour and 20 minutes, and CNN2 over 150 epochs took 1.5 hours; in contrast, DL-noPhy over 150 epochs took 7 hours and 50 minutes. Fewer training datasets for the same DNN model naturally mean shorter training time, so our method performs higher training efficiency than DL-noPhy while ensuring excellent imaging quality.

### Live-cell experiment on HeLa cells

We performed holographic biological experiments on HeLa cells under a  $\times 20/0.45\text{NA}$  lens to demonstrate the application of the method in biological research. The denoised interferogram presented in Fig. 4(a) is of overall low spatial carrier frequency, which results in a spectrum overlapping of cross-correlation and autocorrelation terms, as shown in Fig. 4(b). Figure 4(c) and 4(d) respectively show the phase reconstruction of captured low-carrier frequency holograms utilizing the FT and DL-VHQPI methods for HeLa cells. The field of view (FoV) of Fig. 4(c) and 4(d) is  $0.093 \text{ mm}^2$  (The Imaging Source DMK 23U274,  $1600 \times 1200$ ,  $4.4 \mu\text{m}$ ), and the SBP of the complex amplitude image is 210000 pixels [the area of the FoV, multiplied by the area of the spatial frequency band,  $\pi(NA/\lambda)^2$ ]. To compare the imaging results of the two methods in detail, we selected two regions of interest (ROI, Area1 and Area2) on the specimens, and their magnified views are shown in Fig. 4(e1, e3, f1, f3). Additionally, Fig. 4(e2, e4, f2, f4) vividly depict the reconstructed phase gradient images by digital DIC. It can be revealed that spectrum-overlapping-caused fringe-like error dramatically degrades the phase recovery quality. The selected regions in the red rectangle box of Fig. 4(c) and 4(d) highlight the artifacts-suppression capability on the phase background. And the enlarged views after DIC processing are shown in Fig. 4(g) and 4(h), respectively. The background part of the FT-based reconstructed phase result features many

**Table 1 | The quantitative comparison results of DL-VHQPI and DL-noPhy.**

Evaluation Index	Group1		Group2		Group3		Group4	
	DL-noPhy	DL-VHQPI	DL-noPhy	DL-VHQPI	DL-noPhy	DL-VHQPI	DL-noPhy	DL-VHQPI
MAE	0.0105	0.0065	0.0172	0.0142	0.0202	0.0171	0.0191	0.0170
SSIM	0.9914	0.9969	0.9712	0.9781	0.9637	0.9718	0.9646	0.9697
PSNR	84.8196	88.6846	79.7276	81.5757	78.1123	79.9083	79.4679	80.0348
Size of dataset	3600	324	3600	324	3600	324	3600	324



**Fig. 4 | Results of holographic experiments on HeLa cells.** (a) Low-carrier-frequency high-contrast hologram collected by slightly off-axis interferometry system. (b) Corresponding spatial frequency spectrum. (c) The result of phase recovery by slightly off-axis holography using FT method under  $\times 20$  lens. (d) The result of phase recovery using DL-VHQPI. (e1–e4) and (f1–f4) correspond to the local amplification results of “Area1” and “Area2” for the two samples under different phase recovery methods. Where (e2, e4, f2, f4) are the corresponding DIC views, respectively. (g) and (h) The DIC views after partial magnification of the phase map in the corresponding red box. (i) The numerical distribution of the cross-section and detail-preservation feature of the DL-VHQPI.

coarse diagonal-fringe distributions; in comparison, that of DL-VHQPI is much smoother. The calculated Standard Deviation (STD) quantitatively demonstrates that DL-VHQPI performs a better flatness distribution. As can be readily observed in the cross-section presented in Fig. 4(i), in the FT phase recovery, the reconstruction errors brought by the autocorrelation term will introduce noticeable artifacts to the correct phase result. The results demonstrate that DL-VHQPI can excellently suppress phase artifacts and own the effectiveness and applicability for a practical slightly off-axis DH system.

Indeed, reducing the size of the FT filter window may also be a good way to alleviate artifacts, but this will not fundamentally address the problem of the overlapped spectrum and will cause phase imaging blur. The reason is that reducing the filtering window is at the expense of the system’s SBP and the high-frequency information of the object cannot be enclosed in the limited filtering window. In the Section 3 of Supplementary information, we experimentally present the imaging effects under different FT filtering windows for living cells. To verify the

generalization of DL-VHQPI, we supplemented a new group of experimental results for living cells in Supplementary Section 5, in which we added a comparison and discussion with the VHQPI method and the traditional FT method. The results demonstrate that DL-VHQPI still performs the best artifact-suppression ability and generalizability under a new group of biological applications.

## Conclusions and discussions

In summary, we proposed a high-accuracy artifacts-free single-frame low-carrier frequency fringe demodulation approach for the slightly off-axis DH system, *i.e.*, a model using the DNN-assisted physical process. When the cross-correlation and autocorrelation are inevitably aliased in the SFD, the phase reconstruction based on the conventional FT method cannot eliminate the effect of phase artifacts caused by zero-order term<sup>6</sup>. Although reducing the size of the FT filter window may alleviate the problem of imaging artifacts, the high-frequency information loss of the object caused by the limited filtering

window will cause imaging blur. The method based on Kramers-Kronig relation is proposed on the basis of the concept of “cepstrum” and homomorphic filtering<sup>31</sup>, however, this method must depend on the limited condition of the object-reference ratio and need the separation of the high-order terms in the extended SFD<sup>32,33</sup>. Furthermore, the VHQPI implements the background component removal of single-frame hologram utilizing the principle of image frequency components extraction, while it inevitably suffers from the non-sufficient background term removal caused by the physical method<sup>34</sup>. In contrast, DL-VHQPI, a novel DL-assisted physical model method, can better suppress phase artifacts while improving imaging accuracy. The simulation result quantitatively demonstrates that the phase recovery accuracy obtained by DL-VHQPI is greatly superior to that by FT and VHQPI. Moreover, the live-cell experiment results demonstrate that our method is applicable in biological research.

In addition, it is noteworthy that in the classical end-to-end DNN model (without a physical model), massive data pairs are required to train the network model for a higher reconstruction precision. However, it may be prohibitively laborious and time-consuming for the real-world DH system to collect datasets and generate the corresponding ground truth. Conversely, the proposed DL-VHQPI can perform better same-domain generalization ability and image data-feature extraction capability without a large of datasets. Compared to the classical end-to-end DNN model (i.e., DL-noPhy), DL-VHQPI can achieve a higher reconstruction accuracy utilizing only a small fraction of the datasets due to the physical model reducing the information entropy of DL training objects. Meanwhile, fewer datasets mean shorter training time and higher training efficiency.

The significance of our work lies in the multiple possibilities of applying the proposed DL-assisted physical model idea to the QPI. This idea can be applied to many scenarios in which deep learning methods are applied to the QPI, e.g., addressing a series of ill-posed inverse phase retrieval problems and holography-based high-throughput optical diffraction tomography (ODT) problems<sup>63–65</sup>. Specifically, the artifacts-free low-carrier-frequency fringe demodulation capability of the proposed method has application possibilities for ODT imaging of wide-bandwidth objects. In addition, it has also implications for high-throughput studies of high-robust common-path off-axis interferometer systems<sup>66,67</sup>. We envi-

sion that the idea presented in this research can be applicable to a diverse range of future computational imaging techniques, not just limited to what we discussed here.

## References

1. Fan Y, Li JJ, Lu LP, Sun JS, Hu Y et al. Smart computational light microscopes (SCLMs) of smart computational imaging laboratory (SCILab). *PhotonIX* 2, 19 (2021).
2. Lee K, Kim K, Jung J, Heo J, Cho S et al. Quantitative phase imaging techniques for the study of cell pathophysiology: from principles to applications. *Sensors* 13, 4170–4191 (2013).
3. Park Y, Depeursinge C, Popescu G. Quantitative phase imaging in biomedicine. *Nat Photonics* 12, 578–589 (2018).
4. Vicar T, Balvan J, Jaros J, Jug F, Kolar R et al. Cell segmentation methods for label-free contrast microscopy: review and comprehensive comparison. *BMC Bioinformatics* 20, 360 (2019).
5. Gao P, Wirth R, Lackner J, Sunbul M, Jaeschke A et al. Super-resolution imaging of live cells with genetically encoded silicon rhodamine-binding RNA aptamers. *Biophys J* 118, 145A (2020).
6. Li ZS, Fan Y, Sun JS, Zuo C, Chen Q. A commercialized digital holographic microscope with complete software supporting. *Proc SPIE* 11571, 115711C (2020).
7. Kim MK. Principles and techniques of digital holographic microscopy. *SPIE Rev* 1, 018005 (2010).
8. Kemper B, von Bally G. Digital holographic microscopy for live cell applications and technical inspection. *Appl Opt* 47, A52–A61 (2008).
9. Gao P, Yuan CJ. Resolution enhancement of digital holographic microscopy via synthetic aperture: a review. *Light Adv Manuf* 3, 105–120 (2022).
10. Bettenworth D, Lenz P, Krausewitz P, Brückner M, Ketelhut S et al. Quantitative stain-free and continuous multimodal monitoring of wound healing *in vitro* with digital holographic microscopy. *PLoS One* 9, e107317 (2014).
11. Coppola G, Ferraro P, Iodice M, De Nicola S, Finizio A et al. A digital holographic microscope for complete characterization of microelectromechanical systems. *Meas Sci Technol* 15, 529–539 (2004).
12. Anand V, Han ML, Maksimovic J, Ng SH, Katkus T et al. Single-shot mid-infrared incoherent holography using Lucy-Richardson-Rosen algorithm. *Opto-Electron Sci* 1, 210006 (2022).
13. Xu K, Wang X E, Fan X H et al. Meta-holography: from concept to realization. *Opto-Electron Eng* 49, 220183 (2022).
14. Gao H, Fan XH, Xiong W, Hong MH. Recent advances in optical dynamic meta-holography. *Opto-Electron Adv* 4, 210030 (2021).
15. Gabai H, Baranes-Zeevi M, Zilberman M, Shaked NT. Continuous wide-field characterization of drug release from skin substitute using off-axis interferometry. *Opt Lett* 38, 3017–3020 (2013).
16. Huang ZZ, Memmolo P, Ferraro P, Cao LC. Dual-plane coupled phase retrieval for non-prior holographic imaging. *PhotonIX* 3, 3 (2022).
17. Wu XJ, Sun JS, Zhang JL, Lu LP, Chen R et al. Wavelength-scanning lensfree on-chip microscopy for wide-field pixel-super-resolved quantitative phase imaging. *Opt Lett* 46, 2023–2026

- (2021).
18. Wang HD, Göröcs Z, Luo W, Zhang YB, Rivenson Y et al. Computational out-of-focus imaging increases the space–bandwidth product in lens-based coherent microscopy. *Optica* 3, 1422–1429 (2016).
  19. Micó V, García J, Zalevsky Z, Javidi B. Phase-shifting Gabor holography. *Opt Lett* 34, 1492–1494 (2009).
  20. Poon TC. *Digital Holography and Three-Dimensional Display: Principles and Applications* (Springer, New York, 2006).
  21. Claus D, Iliescu D, Bryanston-Cross P. Quantitative space-bandwidth product analysis in digital holography. *Appl Opt* 50, H116–H127 (2011).
  22. Zhong Z, Bai HY, Shan MG, Zhang YB, Guo LL. Fast phase retrieval in slightly off-axis digital holography. *Opt Lasers Eng* 97, 9–18 (2017).
  23. Xue L, Lai JC, Wang SY, Li ZH. Single-shot slightly-off-axis interferometry based Hilbert phase microscopy of red blood cells. *Biomed Opt Express* 2, 987–995 (2011).
  24. Shaked NT, Zhu YZ, Rinehart MT, Wax A. Two-step-only phase-shifting interferometry with optimized detector bandwidth for microscopy of live cells. *Opt Express* 17, 15585–15591 (2009).
  25. Pavillon N, Arfire C, Bergoënd I, Depeursinge C. Iterative method for zero-order suppression in off-axis digital holography. *Opt Express* 18, 15318–15331 (2010).
  26. Trusiak M, Picazo-Bueno JA, Patorski K, Zdzankowski P, Mico V. Single-shot two-frame  $\pi$ -shifted spatially multiplexed interference phase microscopy. *J Biomed Opt* 24, 096004 (2019).
  27. León-Rodríguez M, Rayas JA, Cordero RR, Martínez-García A, Martínez-Gonzalez A et al. Dual-plane slightly off-axis digital holography based on a single cube beam splitter. *Appl Opt* 57, 2727–2735 (2018).
  28. Han JH, Gao P, Yao BL, Gu YZ, Huang MJ. Slightly off-axis interferometry for microscopy with second wavelength assistance. *Appl Opt* 50, 2793–2798 (2011).
  29. Ikeda T, Popescu G, Dasari RR, Feld MS. Hilbert phase microscopy for investigating fast dynamics in transparent systems. *Opt Lett* 30, 1165–1167 (2005).
  30. Guo CS, Wang BY, Sha B, Lu YJ, Xu MY. Phase derivative method for reconstruction of slightly off-axis digital holograms. *Opt Express* 22, 30553–30558 (2014).
  31. Pavillon N, Seelamantula CS, Kühn J, Unser M, Depeursinge C. Suppression of the zero-order term in off-axis digital holography through nonlinear filtering. *Appl Opt* 48, H186–H195 (2009).
  32. Baek Y, Lee K, Shin S, Park Y. Kramers–Kronig holographic imaging for high-space-bandwidth product. *Optica* 6, 45–51 (2019).
  33. Baek Y, Park Y. Intensity-based holographic imaging via space-domain Kramers–Kronig relations. *Nat Photonics* 15, 354–360 (2021).
  34. Trusiak M, Cywińska M, Micó V, Picazo-Bueno JA, Zuo C et al. Variational Hilbert quantitative phase imaging. *Sci Rep* 10, 13955 (2020).
  35. Cywińska M, Trusiak M, Patorski K. Automated fringe pattern preprocessing using unsupervised variational image decomposition. *Opt Express* 27, 22542–22562 (2019).
  36. Larkin KG, Bone DJ, Oldfield MA. Natural demodulation of two-dimensional fringe patterns. I. General background of the spiral phase quadrature transform. *J Opt Soc Am A* 18, 1862–1870 (2001).
  37. Zuo C, Qian JM, Feng SJ, Yin W, Li YX et al. Deep learning in optical metrology: a review. *Light Sci Appl* 11, 39 (2022).
  38. Feng SJ, Chen Q, Gu GH, Tao TY, Zhang L et al. Fringe pattern analysis using deep learning. *Adv Photonics* 1, 025001 (2019).
  39. Feng SJ, Zuo C, Hu Y, Li YX, Chen Q. Deep-learning-based fringe-pattern analysis with uncertainty estimation. *Optica* 8, 1507–1510 (2021).
  40. Cywińska M, Brzeski F, Krajnik W, Patorski K, Zuo C et al. DeepDensity: convolutional neural network based estimation of local fringe pattern density. *Opt Lasers Eng* 145, 106675 (2021).
  41. Pan B. Optical metrology embraces deep learning: keeping an open mind. *Light Sci Appl* 11, 139 (2022).
  42. Zuo C, Qian JM, Feng SJ, Yin W, Li YX et al. Correction: deep learning in optical metrology: a review. *Light Sci Appl* 11, 74 (2022).
  43. Feng SJ, Zuo C, Zhang L, Yin W, Chen Q. Generalized framework for non-sinusoidal fringe analysis using deep learning. *Photonics Res* 9, 1084–1098 (2021).
  44. Li YX, Qian JM, Feng SJ, Chen Q, Zuo C. Deep-learning-enabled dual-frequency composite fringe projection profilometry for single-shot absolute 3D shape measurement. *Opto-Electron Adv* 5, 210021 (2022).
  45. Zheng CH, Wang TS, Liu ZQ et al. Deep transfer learning method to identify orbital angular momentum beams. *Opto-Electron Eng* 49, 210409 (2022).
  46. Zheng ZH, Zhu SK, Chen Y, Chen HY, Chen JH. Towards integrated mode-division demultiplexing spectrometer by deep learning. *Opto-Electron Sci* 1, 220012 (2022).
  47. Rivenson Y, Zhang YB, Günaydin H, Teng D, Ozcan A. Phase recovery and holographic image reconstruction using deep learning in neural networks. *Light Sci Appl* 7, 17141 (2018).
  48. Rivenson Y, Wu YC, Ozcan A. Deep learning in holography and coherent imaging. *Light Sci Appl* 8, 85 (2019).
  49. Chen HL, Huang LZ, Liu TR, Ozcan A. Fourier Imager Network (FIN): a deep neural network for hologram reconstruction with superior external generalization. *Light Sci Appl* 11, 254 (2022).
  50. Lempitsky V, Vedaldi A, Ulyanov D. Deep image prior. In *Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9446–9454 (IEEE, 2018);<http://doi.org/10.1109/CVPR.2018.00984>.
  51. Wang F, Bian YM, Wang HC, Lyu M, Pedrini G et al. Phase imaging with an untrained neural network. *Light Sci Appl* 9, 77 (2020).
  52. Duran J, Coll B, Sbert C. Chambolle's projection algorithm for total variation denoising. *Image Process Line* 3, 311–331 (2013).
  53. Zhu XJ, Chen ZQ, Tang C. Variational image decomposition for automatic background and noise removal of fringe patterns. *Opt Lett* 38, 275–277 (2013).
  54. Bianco V, Memmolo P, Paturzo M, Finizio A, Javidi B et al. Quasi noise-free digital holography. *Light Sci Appl* 5, e16142 (2016).
  55. Klüber JW. Elimination of slip and instability effects in certain  $M$ -type electron beams. *Proc IEEE* 51, 868–868 (1963).
  56. Yang X, Yu QF, Fu SH. A combined method for obtaining fringe orientations of ESPI. *Opt Commun* 273, 60–66 (2007).
  57. Deng M, Li S, Zhang ZY, Kang I, Fang NX et al. On the interplay between physical and content priors in deep learning for computational imaging. *Opt Express* 28, 24152–24170 (2020).



58. Shannon CE. A mathematical theory of communication. *ACM SIGMOBILE Mob Comput Commun Rev* 5, 3–55 (2001).
59. Cover TM. *Elements of Information Theory*. John Wiley & Sons, 1999).
60. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning* (JMLR.org, 2015).
61. Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning* 807–814 (Omnipress, 2010).
62. Kingma DP, Ba J. Adam: a method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*. <https://arxiv.org/abs/1412.6980> (2015).
63. Choi W, Fang-Yen C, Oh S, Lue N, Dasari RR et al. Tomographic phase microscopy: quantitative 3D-mapping of refractive index in live cells. *Imaging Microsc* 10, 48–50 (2008).
64. Sung Y, Choi W, Fang-Yen C, Badizadegan K, Dasari RR et al. Optical diffraction tomography for high resolution live cell imaging. *Opt Express* 17, 266–277 (2009).
65. Li JJ, Matlock AC, Li YZ, Chen Q, Zuo C et al. High-speed *in vitro* intensity diffraction tomography. *Adv Photonics* 1, 066004 (2019).
66. Mico V, Zalevsky Z, García J. Superresolution optical system by common-path interferometry. *Opt Express* 14, 5168–5177 (2006).
67. Zhang JW, Dai SQ, Ma CJ, Xi TL, Di JL et al. A review of common-path off-axis digital holography: towards high stable optical instrument manufacturing. *Light Adv Manuf* 2, 333–349 (2021).

## Acknowledgements

We are grateful for financial supports from the National Natural Science Foundation of China (61905115, 62105151, 62175109, U21B2033, 62227818), Leading Technology of Jiangsu Basic Research Plan (BK20192003), Youth Foundation of Jiangsu Province (BK20190445, BK20210338), Biomedical Competition Foundation of Jiangsu Province (BE2022847), Key National Industrial Technology Cooperation Foundation of Jiangsu Province (BZ2022039), Fundamental Research Funds for the Central Universities (30920032101), and Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense (JSGP202105, JSGP202201), National Science Center, Poland (2020/37/B/ST7/03629). The authors thank F. Sun for her contribution to this paper in terms of language expression and grammatical correction.

## Competing interests

The authors declare no competing financial interests.

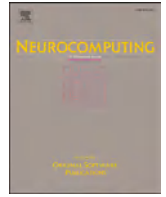
## Supplementary information

Supplementary information for this paper is available. <https://doi.org/10.29026/oes.2023.220023>



Contents lists available at ScienceDirect

## Neurocomputing

journal homepage: [www.elsevier.com/locate/neucom](http://www.elsevier.com/locate/neucom)

Survey paper

Deep learning-powered biomedical photoacoustic imaging<sup>☆</sup>Xiang Wei<sup>a,b,c</sup>, Ting Feng<sup>d</sup>, Qinghua Huang<sup>e</sup>, Qian Chen<sup>a,c</sup>, Chao Zuo<sup>a,b,c</sup>, Haigang Ma<sup>a,b,c,\*</sup><sup>a</sup> Smart Computational Imaging Laboratory (SCILab), School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China<sup>b</sup> Smart Computational Imaging Research Institute (SCIRI) of Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210019, China<sup>c</sup> Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing, Jiangsu Province 210094, China<sup>d</sup> Academy for Engineering & Technology, Fudan University, Shanghai 200433, China<sup>e</sup> School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China

## ARTICLE INFO

Communicated by Zidong Wang

## Keywords:

Biomedical imaging  
Photoacoustic imaging  
Deep learning  
Convolutional networks

## ABSTRACT

Photoacoustic Imaging (PAI) is an emerging hybrid imaging modality that combines optical imaging and ultrasound imaging, offering advantages such as high resolution, strong contrast, and safety. Despite demonstrating superior imaging capabilities, PAI still has certain limitations in its clinical application, such as the trade-off between imaging depth and spatial resolution, and the need for further improvement in imaging speed. Deep Learning, as a novel machine learning technique, has gained significant attention for its ability to improve medical image data and has been widely applied in PAI in recent years to overcome these limitations. In this review, we first introduce the principles of photoacoustic imaging, followed by the development and applications of popular deep neural network structures such as U-Net and GAN networks. Furthermore, we comprehensively discuss the recent advancements in the application of deep learning in photoacoustic imaging. Finally, a summary and discussion are provided.

## 1. Introduction

## 1.1. Photoacoustic imaging

Photoacoustic Imaging (PAI) is a novel non-invasive photon imaging technique used for disease detection, observing biological tissue structure, and assessing function. The physical basis of PAI is the photoacoustic effect in biological tissue. When a short-pulsed laser illuminates the imaged sample, the tissue or substance absorbs the light energy, resulting in thermal elastic expansion and causing instantaneous expansion and contraction of the surrounding medium, thereby generating ultrasound waves propagating towards the tissue surface and being received. By receiving the ultrasound signals and using acoustic inverse problems, the initial sound pressure signal map of the tissue surface can be reconstructed, enabling observation and diagnosis of biological tissue structure and function [1,2]. Due to the significant difference in scattering intensity between ultrasound waves and photons in biological tissue (approximately 2–3 orders of magnitude), ultrasound scattering is much lower than that of photons. Therefore, PAI can overcome the

diffraction limit of optical imaging depth (i.e., 1 mm). Moreover, PAI combines the high imaging depth of ultrasound imaging with the high contrast and high resolution of optical imaging, thereby achieving high-depth, high-contrast, and high-resolution imaging of biological tissue by leveraging the advantages of both technologies.

The most common forms of photoacoustic imaging are photoacoustic tomography (PAT), photoacoustic microscopy (PAM), and photoacoustic endoscopy (PAE) [3,4]. PAT uses a non-focused large-diameter pulsed laser beam to achieve full-field illumination of the tissue surface and employs an array transducer to collect signals, which are then reconstructed into an image using inversion algorithms. Existing inversion algorithms include filtered back-projection (FBP), delay-and-sum (DAS) beamforming algorithm, Fourier-based algorithms, and time reversal (TR) algorithm. PAM, on the other hand, uses a focused short-pulsed laser to illuminate the target point and employs a focused transducer to collect the PA signal point-by-point, allowing for image reconstruction without the need for additional inversion algorithms. PAE is a type of endoscope-based photoacoustic imaging technology. Due to its unique imaging principles and the advantages of

<sup>☆</sup> © 2012 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of Global Science and Technology Forum Pte Ltd

\* Corresponding author at: Smart Computational Imaging Laboratory (SCILab), School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China.

E-mail address: [mahaigang@njjust.edu.cn](mailto:mahaigang@njjust.edu.cn) (H. Ma).<https://doi.org/10.1016/j.neucom.2023.127207>

Received 12 September 2023; Received in revised form 17 November 2023; Accepted 26 December 2023

Available online 29 December 2023

0925-2312/© 2023 Elsevier B.V. All rights reserved.

optics and acoustics, photoacoustic imaging has broad application prospects and has gradually played a role in clinical medicine, biomedical research, drug development, material science, and other biomedical fields. Its application areas are also constantly expanding and deepening.

1.2. Deep learning

With the advent of the big data era, computer computational capabilities have significantly improved, and continuously emerging open-source and user-friendly software frameworks have led to unprecedented development of artificial intelligence (AI) technology. Classical AI technology machine learning has attracted great interest in both industry and academia, especially data-driven artificial neural network technology, that is, deep learning [3–6]. The deep learning method aims to discover complex mappings from training data to achieve optimization of the existing parameter space problem. Unlike the lack of computing power in the past, today’s graphic processing units allow neural networks to continuously improve their depth [7,8], width [9], computation speed and other aspects, gradually developing various basic network architectures. Deep learning has become an important method in computer vision, natural language processing, and AI fields. This article introduces various networks’ applications and effects in the optical-acoustic field based on supervised learning perspective, focusing on classic deep neural network structures.

1.3. Deep learning-powered photoacoustic imaging

In the two previous sections, we introduced the advantages of photoacoustic imaging and deep learning and found that both have very good prospects in their respective fields. Especially in the medical imaging field, photoacoustic imaging has many advantages, such as combining acoustic depth, optical resolution, and non-invasiveness. However, photoacoustic imaging still faces many challenges, including image quality limited by sound and light diffraction, and various problems in the data acquisition, processing, and inversion processes. For example, in PAT, it is difficult to achieve low-cost equipment and high signal-to-noise ratio image reconstruction at the same time, and the widely used sparse detectors currently have difficulty obtaining good reconstruction results through conventional inversion methods. In PAM, there are also deficiencies in imaging speed. Although scanning speed can be improved by changing the repetition rate of the excitation light pulse and the scanning mechanism, these methods often have an unavoidable impact on image quality. In short, there is a certain contradiction between image quality, economic benefits, and time efficiency in photoacoustic imaging. Although many methods have been proposed to solve these problems, and these methods have achieved some effectiveness, further exploration and improvement are still needed.

The intervention of deep learning has had a huge impact in the field of photoacoustic imaging. We have found that a large number of photoacoustic imaging works based on deep learning have achieved imaging quality and efficiency that previous methods have difficulty achieving. This is also the reason why we want to write this review and organize and analyze recent related work. We want to organize and analyze our work in recent years from four important directions of photoacoustic imaging: PAT image reconstruction, PAM image reconstruction, photoacoustic image processing, and photoacoustic signal processing. Not only that, we also introduced the development and current status of common network structures such as U-Net and GAN networks in image processing. Finally, we summarized and prospected the review.

The first chapter of this article introduces the principle of photoacoustic imaging, the principle of deep learning network, and analyzes the current problems of photoacoustic imaging. Chapter 2 details the development of current popular deep learning networks, including U-Net, Residual Network (ResNet), and Super-Resolution Generative Adversarial Network (SRGAN). Chapter 3 lists and analyzes the

application results of current deep learning technology in various fields of photoacoustic imaging. Chapter 4 summarizes the application results and problems of deep learning in the field of photoacoustics, and looks forward to future development directions. The following is the article flowchart and Chart of Recent works on Deep Learning-powered photoacoustic imaging,.

2. The neural network structures based on photoacoustic imaging

In recent years, the combination of photoacoustic imaging and deep learning has brought significant improvements to photoacoustic imaging. Considering the effectiveness, real-time performance, and economy of the method, U-Net has emerged in various aspects of photoacoustic imaging in recent years due to its simple and efficient network structure. It has been applied to PAT reconstruction, PAM reconstruction, denoising, and image processing of photoacoustic images. Its superiority in image recognition and segmentation tasks was first discovered, and then it was applied to image denoising. It is worth noting that the skip connections in U-Net ensure the validity of the image, which effectively improves the signal-to-noise ratio of the image and greatly suppresses the possible artifacts produced during the processing. Subsequently, the U-Net network has also been widely applied to other aspects of various photoacoustic imaging methods, which also proves that network optimization for image recognition tasks is applicable to optimizing network performance for other image tasks. In addition, U-shaped deep neural networks also have certain robustness and generalization capabilities, can process different types and qualities of data, and can further improve model performance through techniques such as data augmentation.

Overall, U-Net has great advantages in the field of photoacoustic imaging. It can effectively process high-dimensional data, learn features in the data, and achieve accurate image segmentation and localization. This makes it a very promising tool in the field of photoacoustic imaging, which can help doctors and researchers make more accurate diagnoses and treatments. This chapter mainly introduces the U-Net and SRGAN network structures, details the birth of U-Net and the development process of its network architecture, shares the photoacoustic microscopy method we are working on based on SRGAN, and finally introduces the classic residual block structure.

2.1. U-Net

2.1.1. The proposal of U-Net network

Ronneberger et al. first proposed the U-Net network in 2015 [10],

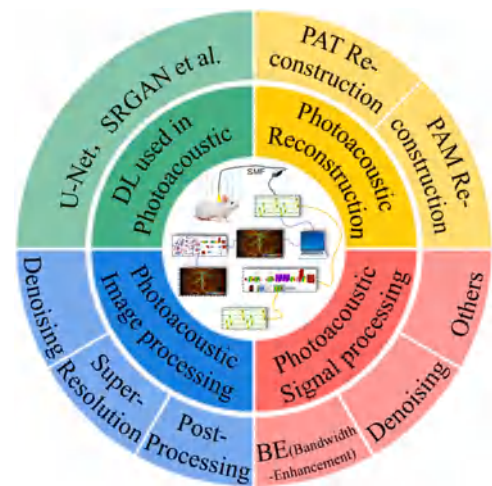


Fig. 1. The flowchart of this paper.

Recent works on Deep Learning-powered photoacoustic imaging				
Type	Name	Published time	Structure	Notes
PAT reconstruction	Deep learning for photoacoustic tomography from sparse data	2019	U-net with residual connection	It contains PAT filtered backprojection algorithm for the first layer.
	A New Deep Learning Network for Mitigating Limited-view and Under-sampling Artifacts in Ring-shaped Photoacoustic Tomography	2020	ring-array deep learning network (RADL-net)	It improves the quality of reconstructed images on a three-quarter ring transducer array.
	Limited-View and Sparse Photoacoustic Tomography for Neuroimaging with Deep Learning	2020	pixel-wise deep learning (Pixel-DL)	It contains pixel-wise interpolation governed by the physics of photoacoustic wave propagation.
	Deep-learning image reconstruction for real-time photoacoustic system	2020	a network using 3-D transformed arrays a multi-channel input 'upgUNET'	There is Reformatting raw channel data into a multi-channel array as a pre-processing step.
PAM reconstruction	Photoacoustic Microscopy with Sparse Data Enabled by Convolutional Neural Networks for Fast Imaging	2020(arxiv)	ResNet with Residual Block and SE Block	The CNN model utilized both squeeze and excitation blocks and residual blocks to achieve the enhancement.
	Deep Learning Enables Superior Photoacoustic Imaging at Ultralow Laser Dosages	2021	multitask residual dense network (MT-RDN)	It utilized an innovative strategy of integrating multi-supervised learning, dual-channel sample collection, and a reasonable weight distribution.
PA image processing	De-Noising of Photoacoustic Microscopy Images by Attentive Generative Adversarial Network	2022	an attention enhanced generative adversarial network	It is an attention enhanced GAN that uses an improved U-net generator to remove noise from PAM images.
	High-resolution photoacoustic microscopy with deep penetration through learning	2022	Wasserstein distancegenerative adversarial Network (WGAN)	It uses Wasserstein distance to replace the Jensen-Shannon divergence as the objective to be optimized.
	Photoacoustic Image Classification and Segmentation of Breast Cancer: A Feasibility Study	2018	AlexNet & GoogLeNet	It introduces the pre-trained AlexNet and GoogLeNet-based transfer learning for photoacoustic breast cancer classification.
PA signal processing	Deep Neural Network-Based Sinogram Super-Resolution and Bandwidth Enhancement for Limited-Data Photoacoustic Tomography	2020	U-Net (Hybrid) variant	It provides the generalization by effectively modeling the negative values in sinogram through the final layers while maintaining the advantages of ReLUs in U-Net architecture.
	Photoacoustic digital brain and deep-learning-assisted image reconstruction	2023	U-Net structure	It Uses a combination of simulation, MRA, and MRI to obtain prior ground truth images.
	Temporal and spectral unmixing of photoacoustic signals by deep learning	2021	conditional generative adversarial network (cGAN)	It automatically learns a loss that adapts to the data and distinguishes photoacoustic signals with phase differences

Chart 1. Chart of Recent works on Deep Learning-powered photoacoustic imaging.

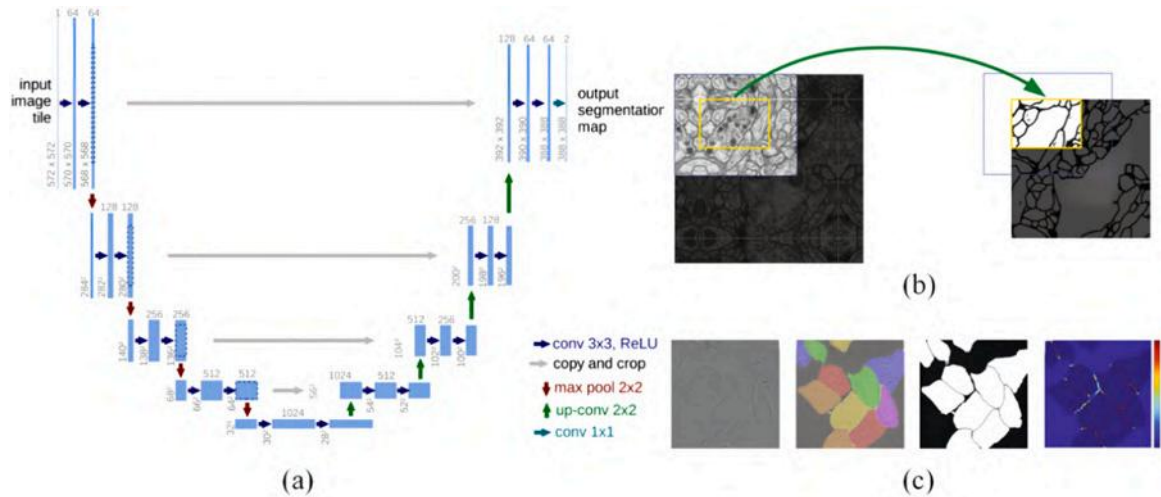
which was initially applied to image segmentation and won the championships of the ISBI 2015 Cell Tracking Challenge and Caries Detection Challenge. To this day, U-Net has inspired the development of many network structures, and more and more deep learning strategies continue to extend based on U-Net. The U-Net network structure consists of a contracting path for capturing context and a symmetric expanding path for precise localization. U-Net can also be combined with data augmentation techniques to achieve end-to-end training with a small amount of data input [11–13]. Due to its network structure resembling the letter "U", it is named U-Net. The initial U-Net network structure is shown in Fig. 2(a).

In U-Net network structure, the vertical arrows form the processes of the encoder and decoder, while the horizontal arrows represent skip connections that jump across multiple layers. Its multi-layer encoder and decoder structure together constitute an overall layout resembling the letter "U". The left part of U-Net is the encoder, and the right part is

the decoder. Let's discuss the encoder and decoder in detail. The encoder is responsible for extracting features from the input image. It gradually reduces the size of the feature map and increases the number of channels through multiple convolutional layers to extract more abstract features. Its structure consists of four blocks. Each block is composed of a  $3 \times 3$  convolution (using the ReLU activation function) and a pooling layer with a stride of  $2 \times 2$ . After processing through the four blocks, the feature map is gradually reduced. The output of the encoder is passed to the decoder, and at the same time, skip connections are made between the output of each stage of the encoder and the symmetric stage of the decoder to preserve the detailed information of the feature map.

U-Net was initially applied to image segmentation, as shown in Fig. 2 (b) and (c). Fig. 2(b) demonstrates the U-Net's overlapping-tile strategy for seamless segmentation of images of arbitrary sizes by predicting the segmentation results of small selected areas through inputting the large selected frame image. Fig. 2(c) shows the process of observing HeLa cells





**Fig. 2.** Network structure and image segmentation; (a) Basic structure of U-Net network; (b) Seamless segmentation effect; (c) Progressive treatment of HeLa cells, the four results are the original image, the image overlapped with the true value segmentation, the generated segmentation mask, and the result of using pixel loss weight mapping.

using differential interference contrast microscopy, where the four images represent the original image, the image overlaid with ground truth segmentation (different colors indicate different stages of HeLa cells), the generated segmentation mask (white represents foreground, black represents background), and the result using pixel loss weight mapping.

U-Net's overlapping-tile strategy has been widely used in medical image segmentation, effectively handling images of any size and achieving relatively accurate segmentation results. Meanwhile, U-Net can achieve end-to-end training with a small amount of data input by combining data augmentation and pixel loss weight mapping methods, making the network robust and capable of generalizing well.

### 2.1.2. The development of U-Net

U-Net is one of the currently popular network architectures, which was initially applied to image segmentation. With the continuous development and in-depth research of deep learning frameworks, the network structure of U-Net has also been continuously optimized and improved. More and more deep neural network structures have been discovered and combined with the U-Net architecture to further improve network efficiency. In addition, U-Net has been widely used in fields such as image reconstruction, image super-resolution, semantic segmentation, and signal processing, and has achieved good results.

In 2016, Cicek et al. proposed the 3D U-Net based on U-Net, which is used for 3D image segmentation [14]. Compared with U-Net, 3D U-Net only uses three downsampling operations and a normalization layer after each convolutional layer. It is worth noting that both 3D U-Net and U-Net do not use random dropout layers. In the 2018 MICCAI Brain Tumor Segmentation Challenge (BRATS), the team of the German Cancer Research Center used 3D U-Net and achieved the second place in the challenge with only a few modifications [15]. This indicates that compared to many new networks, 3D U-Net still has significant advantages.

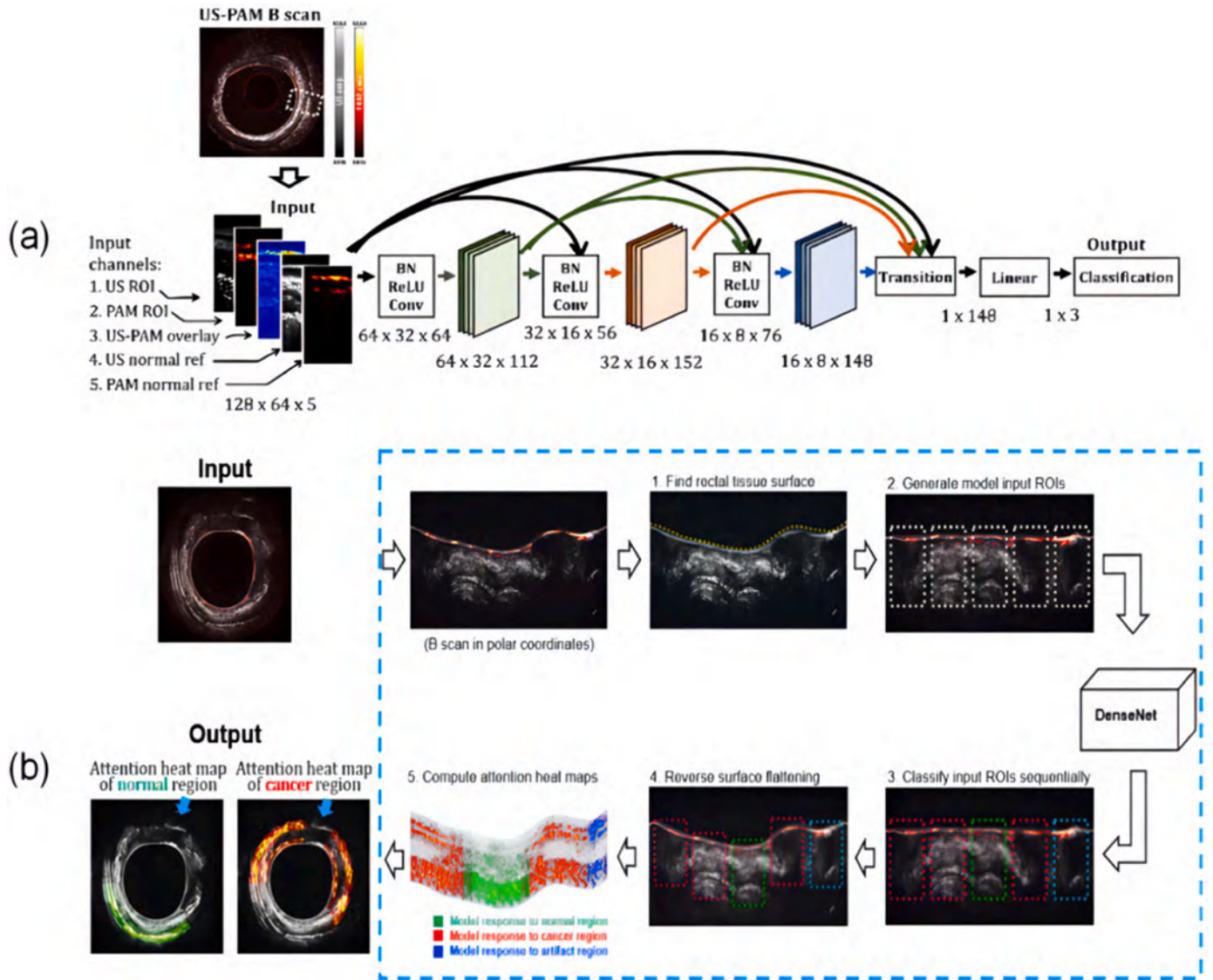
In 2018, residual U-shaped network (Res-UNet) and dense U-shaped network (Dense-UNet) were born based on the U-Net architecture. Res-UNet and Dense-UNet are inspired by residual connections and dense connections, respectively, replacing each sub-module of U-Net with a form of residual connection or dense connection. Among them, dense connection means that the output of a layer in the sub-module is used as part of the input of subsequent layers, while any layer's input comes from the combination of outputs of previous layers. Res-UNet has been applied to the segmentation of retinal images [16], while Dense-UNet has been used to remove artifacts in images [17], which is the first case of using the U-Net architecture for image processing. The authors

pointed out that U-Net is the most widely used CNN architecture for applying deep learning and post-processing methods to sparse tomographic image reconstruction [18]. It has many characteristics suitable for artifact removal, such as multi-level decomposition and multi-channel filtering. Moreover, on both synthetic data and experimental data [19], it shows better performance in removing sparse PAT image artifacts than iterative methods. The core idea of DenseNet is the Dense Block. In a Dense Block, the input of each layer is a concatenation of the outputs from all previous layers. Due to the direct connections between each layer and all preceding layers, DenseNet can effectively utilize parameters, resulting in a model with fewer parameters and reduced risk of overfitting. With the dense connectivity design, every layer in DenseNet has direct access to the feature maps from previous layers, facilitating feature propagation and reuse, which helps in learning richer feature representations.

Lin et al. proposed a robust deep learning network for ultrasonic photoacoustic microscopy with two modes dense network [20] (US-PAM DenseNet), aimed at improving the performance of the model in distinguishing malignant from non-cancerous tissues based on co-registration of dual-mode ultrasound (US) and PAM images, as well as individualized normal reference images, as training. In Fig. 3, the US-PAM DenseNet similarly classifies the entire US-PAM B scan by ROI grade and computes the ROI heat map, highlighting the rectal cancer region. In Fig. 3(a), Five channels are generated from the selected ROI as the model input, which has dimensions of  $128 \times 64 \times 5$ . Solid arrows indicate data flows and connections inside the model: different colors correspond to different data origins. Connections are made between every pair of layers in the DenseNet architecture. The model has three layers, with 64 initial kernels in the first layer, a kernel growth rate of 12 from one layer to the next, and block repetition numbers of 4, 8, and 6 respectively for the three layers. The size of each model layer is marked under the layer icons.

In the same year, U-Net began to be applied to direct PAT reconstruction of sparse data from raw sensors.

Guan et al. proposed a new deep learning method called Pixel-DL (Pixel-wise Deep Learning) [21]. It first utilizes pixel-wise interpolation controlled by the physical propagation of photoacoustic waves, and then employs convolutional neural networks (CNNs) to reconstruct images. Synthetic phantom data from mouse brain, lung, and retinal vascular system were used for training and testing. The results show that Pixel-DL achieves comparable or better performance compared to iterative methods, making it suitable for real-time photoacoustic tomography (PAT) rendering and improving image reconstruction quality in

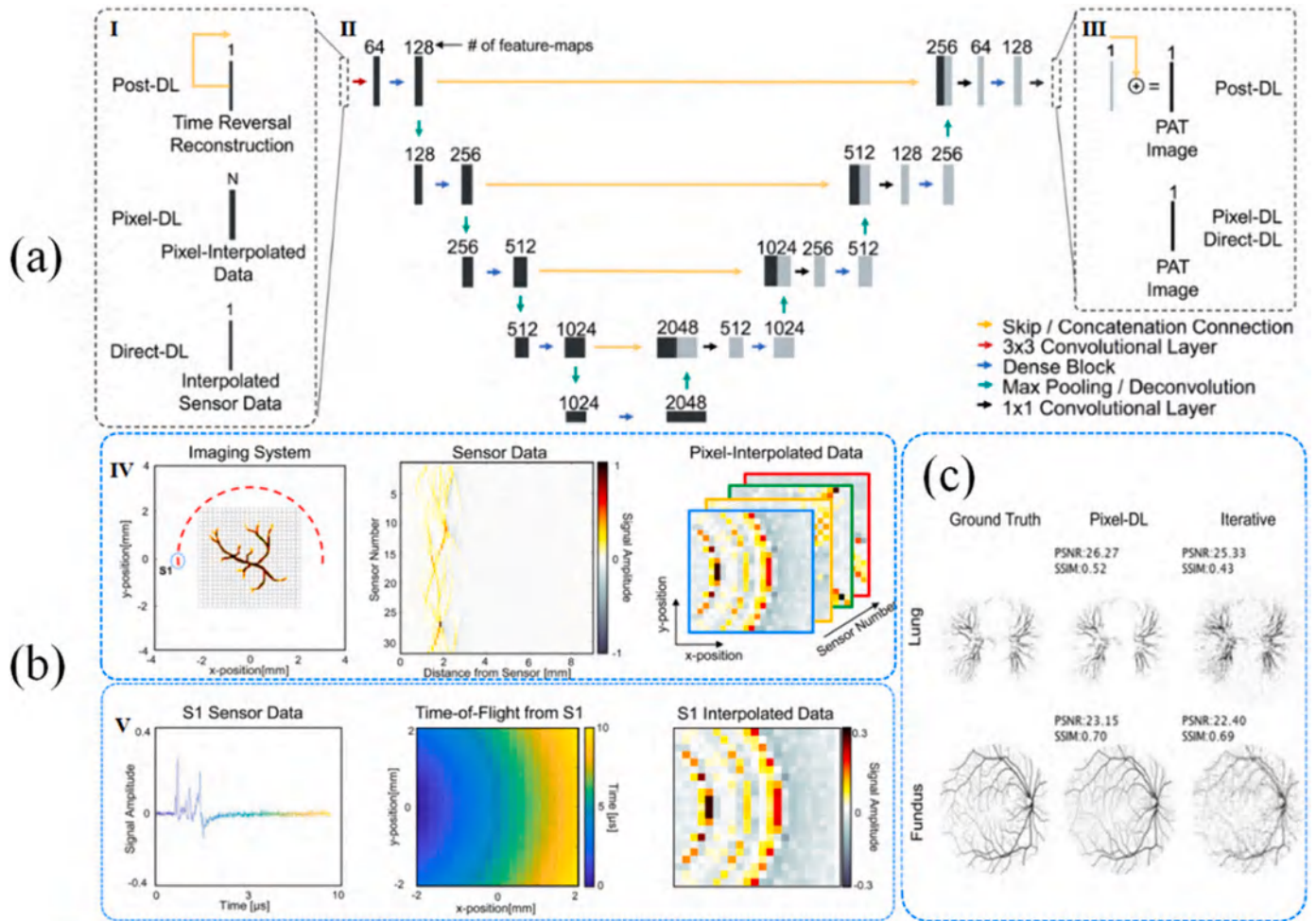


**Fig. 3.** The architecture of US-PAM DenseNet and the application of US-PAM DenseNet in generating thermal maps of suspicious tumor regions; (a) The white dotted box shows an example ROI selected from a co-registered US-PAM B scan. (b) The white dotted box shows an example ROI selected from a co-registered US-PAM B scan. Five channels are generated from the selected ROI as the model input, which has dimensions of  $128 \times 64 \times 5$ . Solid arrows indicate data flows and connections inside the model: different colors correspond to different data origins. Connections are made between every pair of layers in the DenseNet architecture. The model has three layers, with 64 initial kernels in the first layer, a kernel growth rate of 12 from one layer to the next, and block repetition numbers of 4, 8, and 6 respectively for the three layers. The size of each model layer is marked under the layer icons. Pipeline for applying US-PAM DenseNet to diagnose a whole US-PAM B scan and generate an attention heat map of suspicious cancer regions to facilitate surgeons decision making. The processing pipeline is illustrated in the blue box. In steps 3, 4, and 5, green dotted boxes show the ROIs that US-PAM DenseNet classifies as normal, red shows the cancer ROIs, and cyan shows artifacts. In step 5, guided backpropagation is computed for all three potential classification outcomes, i.e., normal, cancer and artifact, weighted with their respective prediction scores.

limited-view and sparse PAT scenarios. Fig. 4(c) shows the improvement of Pixel-DL. Comparable to iterative reconstruction, Pixel-DL had similar performance for the fundus vasculature and outperformed it for the lung vasculature dataset. In this work, three different CNN-based deep learning methods for limited-view and sparse PAT image reconstruction were used, as shown in Fig. 4(a). Fig. 4(a)(I) shows inputs into the CNN for each deep learning approach. The Post-DL CNN implementation used residual learning which included a skip connection between the input and final addition operation. The initial Pixel-DL input contains “N” feature-maps corresponding to the number of sensors in the imaging system; (II) The FD-UNet is comprised of a contracting and expanding path with concatenation connections; (III) The output of the CNN is the desired PAT image. In Post-DL, residual learning is used to acquire the final PAT image. In Post-DL, the sensor data is reconstructed into an image with artifacts using time reversal, and then CNN is utilized as a

post-processing step to remove the artifacts and enhance the image. In Pixel-DL, window-correlated information in the sensor data is interpolated on a pixel-by-pixel basis and mapped to the image space. In the improved Direct-DL implementation (mDirect-DL), a combination of linear interpolation and downsampling is used to ensure that the interpolated sensor data has the same dimensions as the final PAT image. In Fig. 5(b)(IV) The red semi-circle represents the sensor array, and the gray grid represents the defined reconstruction grid. In Fig. 5(b)(V) Color represents the time at which a pressure measurement was taken and is included to highlight the use of time-of-flight to map the sensor data to the reconstruction grid.

In 2019, Lan et al. proposed a Y-Net network based on the U-Net idea [22]. Unlike the general U-Net, Y-Net has two inputs and one output, i.e., two encoders and one decoder. By using the measured raw data and the beamformed image as inputs, Y-Net solves the PAT image



**Fig. 4.** The proposed FD-UNet network architecture, the introduced pixel interpolation process and PAT sensor data acquired with 32 sensors and a semi-circle view; (a)(I) Inputs into the CNN for each deep learning approach. The Post-DL CNN implementation used residual learning which included a skip connection between the input and final addition operation. The initial Pixel-DL input contains “N” feature-maps corresponding to the number of sensors in the imaging system; (II) The FD-UNet is comprised of a contracting and expanding path with concatenation connections; (III) The output of the CNN is the desired PAT image. In Post-DL, residual learning is used to acquire the final PAT image; (b)(IV) There are Schematic of the PAT system for imaging the vasculature phantom. The first sensor (S1) is circled and used as an example for applying pixel-wise interpolation to a single sensor; The PAT time series pressure sensor data measured by the sensor array; Resulting pixel-interpolated data after applying pixel-wise interpolation to each sensor based on the reconstruction grid; (V) There are Sensor data for S1; Calculated time-of-flight for a signal originating at each pixel position and traveling to S1; Pressure measurements are mapped from the S1 sensor data to the reconstruction grid based on the calculate time-of-flight for each pixel.(c)Data were acquired respectively on images of lung and fundus vasculature.

reconstruction problem, which can also be called hybrid processing. Inspired by the Y-Net network idea, in 2022, Guo et al. proposed an attention-guided network based on multi-feature fusion (AS-Net) for PA reconstruction, aiming to solve the PA reconstruction problem under sparse conditions of ultrasonic transducers in photoacoustic tomography [23].

In Fig. 5, Firstly, 2-D PA raw data is transformed into a 3-D square matrix by Folded Transformation (FT). Then AS-Net produces the multi-feature fusion base on the attention mechanism for PA reconstruction. ASKF-Net architecture consists of a basic PA reconstruction (BPR) module, semantic feature extraction (SFE) module, and feature fusion (FF) module. BPR module is a modified Auto-Encoder architecture used to reconstruct images from the PA signal, while the SFE module aims to extract semantic features from the DAS image. FF module is used to fuse the semantic feature into the output of the BPR module and generate the final reconstructed image.

In 2022, MENG et al. proposed a deep tissue acoustic-resolution photoacoustic microscopy technique based on a two-stage deep learning network [24]. This technique can adaptively restore high-resolution photoacoustic images at different defocusing depths,

thereby partially solving the problem of poor imaging quality of off-focus plane targets. Specifically, the network structure consists of two stages. The first stage of the deep learning network is used to reconstruct the region far away from the focus, and the second stage reconstructs the region near the focus. In order to achieve image reconstruction, a residual U-shaped network with attention gates (Res-UNet\_AG) is also designed in this study.

## 2.2. Generative adversarial network

Super-Resolution Generative Adversarial Network (SRGAN) is a network proposed by Christian Ledig et al. in 2017 in their paper [25]. This paper presents a super-resolution method based on generative adversarial networks, which can convert low-resolution images into high-resolution and realistic images. The appearance of SRGAN has attracted wide attention in the field of image processing and has achieved good results in practical applications.

The main body of the SRGAN network consists of two independent and combinable training network structures, namely the generator and discriminator. The network loss function consists of a perceptual loss



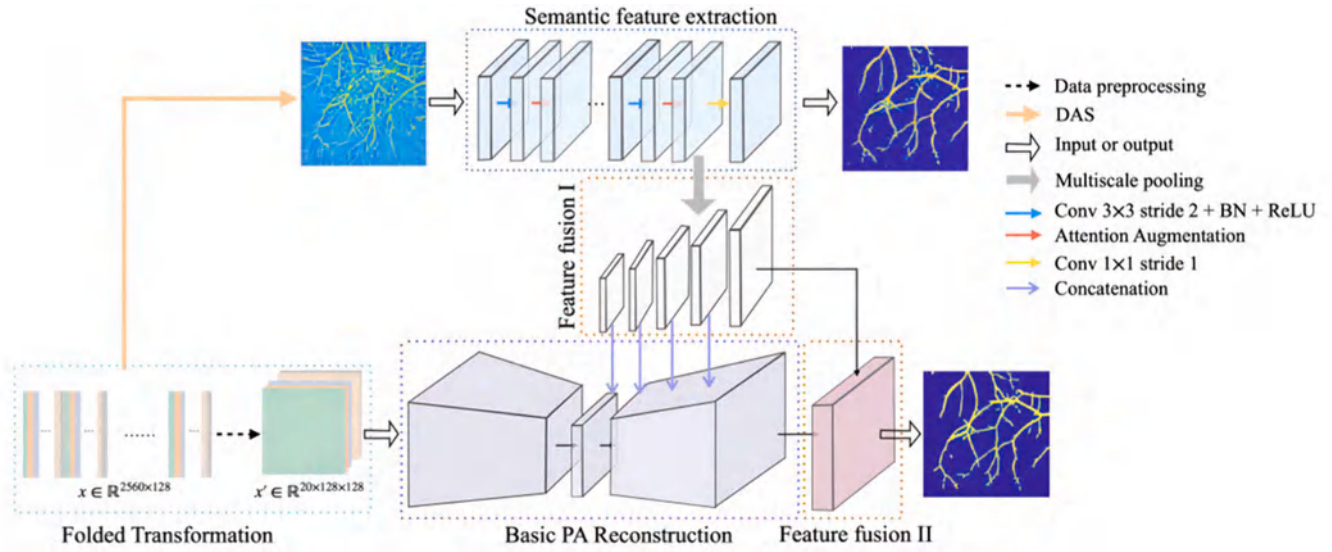


Fig. 5. Illustration of the reconstruction framework from Guo et al., which includes PA raw data preprocessing and AS-Net reconstruction network.

function, which corresponds to the content loss of the generator, and an adversarial loss function of the discriminator weighted by the two. The training idea of the network is to input low-resolution images into the generator network to obtain super-resolution images, and then input the super-resolution images into the discriminator network. The discriminator network will output a binary result representing the "truthfulness" of the image. If the discriminator outputs that the image is fake, the loss value will be returned to the generator network for further training; if it is true, the discriminator will continue to train. In summary, the generator and discriminator mutually constrain each other and train with a related loss function. The goal of the generator can be understood as "deceiving" the discriminator, while the goal of the discriminator is to optimize the authenticity of the generator. The network iteration stops when the minimum error is reached, and the generator at this time is taken as the final result of the network training. The content loss function used in this paper is different from the spatial loss and is based on the feature space Mean Squared Error (MSE) loss of a certain layer weight of the VGG19 model [26,27]. This loss function can improve the semantic recognition and readability of the image. The paper compares

the reconstruction effects of three methods, interpolation, ResNet, and SRGAN (the generator is ResNet). The results show that SRGAN has a better effect in extracting image features. Fig. 6 shows the SRGAN network structure.

In 2018, Wang et al. proposed an enhanced super-resolution generative adversarial network to solve the artifacts generated by SRGAN in image super-resolution [28]. The super-resolution generative adversarial network (SRGAN) can generate realistic textures during the single image super-resolution process. However, the details of the reconstructed image are often accompanied by artifacts. In order to further improve the visual quality of SRGAN, the authors conducted in-depth research on SRGAN and improved three key components: network structure, adversarial loss, and perceptual loss, to obtain an enhanced SRGAN (ESRGAN). The residual-in-residual dense block (RRDB) without batch normalization is introduced as the basic network building unit. In addition, the authors used the idea of relative error to let the discriminator predict the relative realism instead of the absolute value. Finally, the activation before the feature is used to improve the perceptual loss and provide stronger supervision for brightness consistency and texture

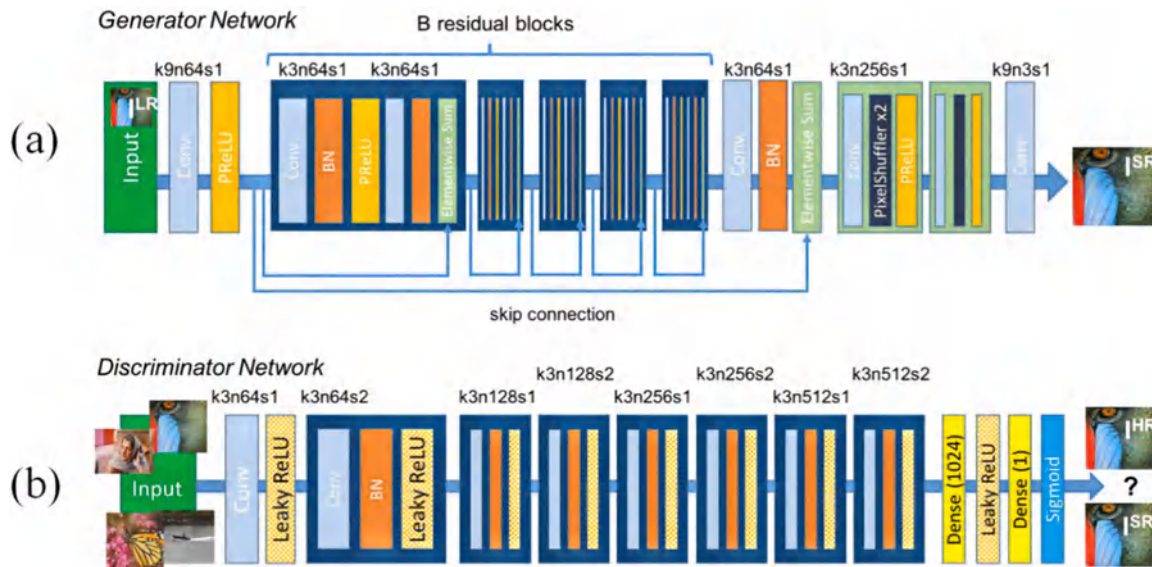


Fig. 6. The SRGAN structure. (a) The Generator Network of SRGAN structure. The main function is to generate parameters. (b) The Discriminator Network of SRGAN structure. The main function is to cooperate and build generate corresponding parameters which could be used to train Generator Network for improvement.



restoration. The network improvements are shown in Fig. 7.

Fig. 7(a) shows the network flow of ESRGAN, where Residual Blocks in the SRGAN generator are treated as a type of feature extraction layer called Basic Block. Fig. 7(b) shows that ESRGAN removes the normalization layer (BN, Batch Normalization) in the Residual Blocks and adds Dense Blocks as feature extraction layers after Residual Blocks. In each Residual Block, the BN layer appears twice to ensure that the network gradient does not explode, but it significantly slows down the network training speed and reduces the amount of feature information, which leads to a significant impact on the reconstruction effect. Dense Blocks can also solve the problem of gradient diffusion and explosion, and will not reduce the image reconstruction effect. The authors also proposed optimizing the loss function of the discriminator, which is different from the standard discriminator in SRGAN. It estimates the real probability of the input image, that is, trying to predict the probability that the real image is relatively more realistic than the fake image. This modification helps to learn sharper edges and more detailed textures.

### 2.3. Residual blocks and residual networks

U-Net is constantly developing in terms of width and depth in the field of photoacoustic imaging, especially in the areas of reconstruction and image processing in PAM. With the continuous improvement of imaging speed, image effectiveness, and the requirements for biaxial resolution, the ordinary U-Net network structure is difficult to meet the situation of deepening the width and depth. The introduction of Residual Blocks into U-Net has enabled the network to reach unprecedented depths, and Residual Blocks have been fully applied not only in U-Net but also in other network structures such as the generator network of SRGAN. Szegedy et al. summarized the impact of network structures including Residual Blocks on image recognition tasks [29]. The authors analyzed the inherent importance of residual connections for training very deep neural networks. High-performance networks are often very deep, and deep neural networks are difficult to train compared to shallow neural networks because of the problems of gradient vanishing and exploding, as well as the increased computational complexity that increases the hardware requirements for network training. Skip connections are an important component structure of Residual Blocks, which can obtain weights from a certain layer of the network layer and quickly feedback to another layer, usually skipping connections to deeper layers. This structure can reflect the weights of the lower layers of the network in the next layers of the network, thereby avoiding gradient vanishing and exploding problems, and improving the efficiency and stability of network training. Currently, it is common to use residual connections to replace filter cascading stages.

The authors also pointed out that optimizing convolutional neural networks with recognition performance as the goal can also be

transformed into performance improvements in other tasks. Using Residual Blocks to construct residual networks (ResNet) that can train deep networks not only has good results in image segmentation but can also be further extended to other fields such as medical imaging. He et al. first proposed ResNet to solve the problems of gradient vanishing and exploding in deep networks while successfully increasing the number of network layers to the order of  $10^3$  while ensuring the constraint of the loss function. ResNet is composed of Residual Blocks [30,31]. The proposal of ResNet and Residual Blocks ensures the effectiveness of training deep neural networks. Even if the network depth reaches the level of  $10^3$ , the loss function can be optimized to ensure a reduction in training error.

## 3. Application of deep learning in photoacoustic imaging

### 3.1. Photoacoustic image reconstruction

Deep learning methods, as a new information mining method, have a wide range of applications in multidimensional information processing, such as reconstruction, denoising, super-resolution, etc., and have achieved many good results. Currently, there are also some non-iterative reconstruction schemes proposed, such as direct estimation, PA signal model reconstruction, and PA signal or image enhancement through deep learning.

The direct reconstruction method solves the PA wave equation, which captures the mapping from signal to image with the PA signal as input. Waibel et al. established a direct estimation from light and sound signal detector data to PA imaging [32], input the synthetic data of the 128-element linear detector into an improved U-Net, and reconstructed the final initial PA pressure signal. Schwab et al. used deep learning to learn the weights of reflected data on different channels and trained neural networks for vessel phantom. Meanwhile, the model used Shepp-Logan phantom to verify. They also proposed a data-driven regularization method [33], which significantly suppresses noise by applying truncated singular value decomposition (SVD) [34] and then restoring truncated SVD coordinate coefficients. Lan proposed using three different sensor data (2.25 MHz, 5 MHz, 7.5 MHz) as input and using U-net for direct reconstruction. Feng et al. improved Res-UNet for direct reconstruction of simple phantoms and compared it with some U-net models [35]. Tong Tong [36] trained a feature pyramid network (FPnet) as post-processing using in vivo data. Mohammad Abu Anas et al. proposed a deep CNN network structure for beamforming PA data [37], which consists of five dense blocks consisting of convolutional layers with different sizes. The article discusses the influence of variable sound speed on this method and verifies its robustness under variable sound speed.

In particular, in the PAI system, due to the existence of optical

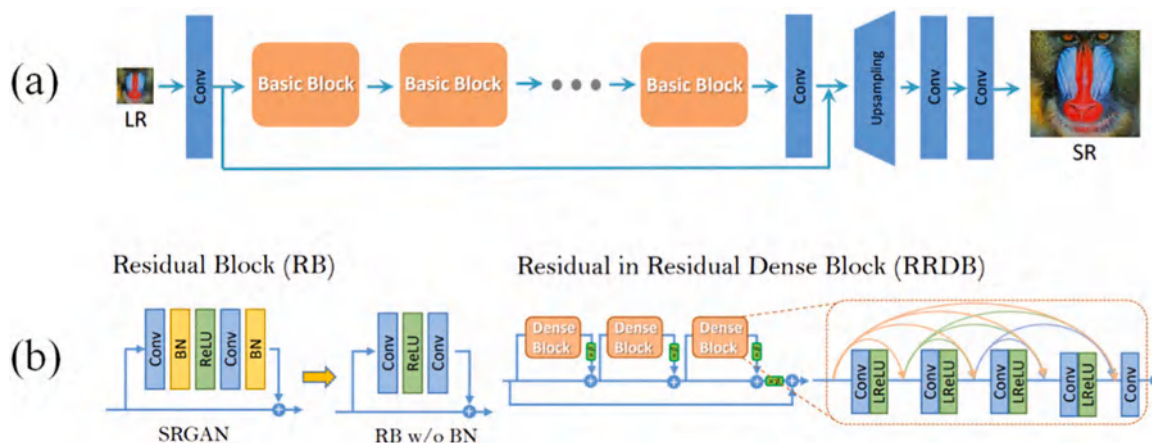


Fig. 7. ESRGAN Network; (a) ESRGAN flow; (b) Modified Residual Block and RRDB.

scattering, the effective excitation degree of deep targets is affected to a certain extent, which is a problem that cannot be ignored. To solve this problem, Johnstonbaugh et al. designed an encoder-decoder network for predicting objects in deep tissue [38]. This study introduced acoustic and optical attenuation in simulation and compared it with actual situations. Allman et al. used VGG16 beam to detect and eliminate reflections from point sources [39], and all experiments used simple (point) phantoms. This method uses neural networks to eliminate artifacts caused by reflections, which greatly improves imaging speed.

In summary, there is currently widespread research on non-iterative reconstruction methods, indicating the urgent need for real-time photoacoustic imaging. Not only in PAT, but also in PAM and PAE, there are various problems with real-time imaging. Factors affecting imaging speed in traditional PAM algorithms include the repetition rate of the excitation light pulse, scanning mechanism, signal preprocessing, and image post-processing. Common solutions such as increasing pulse repetition rate, sampling scanning method, and pixel stacking have been verified, but at the same time, they also face the lack of imaging quality. Considering the current research status of PAE, there are not many studies on PAE reconstruction using deep learning, so it will not be further described here.

### 3.1.1. Photoacoustic tomography image reconstruction

Image reconstruction is an important part of photoacoustic computed tomography (PAT), which is responsible for converting the raw signals received by the ultrasound transducer into an initial pressure distribution image. Due to the ill-posed nature of photoacoustic imaging and the lack of an accurate inverse model in practical situations (limited field of view and sparse sampling), photoacoustic tomography reconstruction is still challenging. In PAT, the purpose of image reconstruction is to reconstruct the initial PA pressure distribution, which is positively correlated with the optical absorption intensity of biological tissues. The sensor array receives PA signals  $P(P(r, t) | r, t)$  represents 3D position and time) excited by short-pulsed laser at different ionization levels, and based on these PA signals, the acoustic-thermal information  $H(r, t)$  is reconstructed through some inverse reconstruction methods, and then  $A(r)$  is further reconstructed, i.e., the distribution of tissue optical absorption intensity. Currently, the most commonly used inverse reconstruction methods include model-based methods such as back-projection (BP) and time reversal (TR); sparse data-based reconstruction methods such as compressed sensing (CS), wavelet transform (WT), and discrete cosine transform (DCT); data mining methods such as deep learning; and model-based iterative methods. Among them, back-projection method is the most widely used, while BP and its derived algorithms such as filtered backprojection (FBP) are considered the most famous PAT reconstruction algorithm due to their simple implementation [40,41].

If experimental conditions are sufficient, i.e., a sufficiently large and dense ultrasound transducer array is distributed on the inner radius of a circular or elliptical detector, the photoacoustic inverse problem of the backprojection method can be expressed as follows [40]:

$$A(r) \propto \int d\theta \frac{1}{t} \frac{\partial p(r_0, t)}{\partial t} \Big|_{t=|r_0-r|/c} \quad (1)$$

Here,  $r$  is the position of the acoustic pressure;  $c$  is the speed of sound;  $\theta$  is the angle between the ultrasound transducer and the acoustic pressure signal;  $r_0$  is any position of the ultrasound transducer on the inner radius of the circular or elliptical detector;  $p(r_0, t)$  is the known condition for the inverse operation, i.e., the acoustic pressure signal received by the ultrasound array at that position;  $A(r)$  is the spatial distribution of tissue optical absorption intensity.

Kim et al. proposed to modify 2D raw data (with time and detector dimensions) into a 3D array (with two spatial dimensions and one channel dimension), where the channel data packages correspond to the propagation delay distribution at a spatial point and serves as the input to the neural network [42]. Traditional popular machine learning

methods train on incomplete images obtained under ill-posed conditions through standard reconstruction methods [43–46]. Due to the loss of previously captured weak information that is difficult to reconstruct, the fine structure of the reconstructed image is often unsatisfactory. Kim's method trains on the basis of the first step of most traditional reconstruction methods, greatly simplifying the learning process. The expansion of the channel dimension preserves more information and improves learning accuracy.

Fig. 8(a) shows the input data of the neural network. Using simple acoustic propagation physics rules in  $r(x, z)$  and the linear array transducer system, 3D transformed data is obtained by the propagation delay distribution of specific image points at different depths, which are used as inputs to the network.

Fig. 8(b) shows the CNN network architecture used in the study. Prior to data input into the network, pre-processing was performed by looking up a priori LUT tables on the original signal ( $2048 \times 128$  obtained by adding noise to real images), converting it into a  $512 \times 128 \times 128$  data array containing delay information. Reformatting the original channel data into a multi-channel array as a pre-processing step improves learning efficiency for highly complex network structures. This neural network uses U-net as a basis and decomposes the signal through multi-scale feature mapping. By combining trainable networks with transformation methods, the structure of vascular networks was simulated in simulations and experiments. Overall, this method significantly improves image quality compared to traditional methods for reconstructing PA data, but loses a little complex absorption body geometry and may produce small artifacts.

Antholzer et al. proposed a direct and efficient reconstruction algorithm based on deep learning for the sparse data problem in reconstructions [47]. The first step uses the PAT filtered backprojection algorithm, followed by optimizing the reconstruction results using the U-net architecture. It not only solves the time-consuming forward and adjoint problems, but also has better imaging effects than direct filtered backprojection algorithms, and performs similarly to existing iterative methods for sparse data PAT. Because iterative algorithms have their own limitations. For example, the reconstruction quality strongly depends on the used a-priori model about the objects to be recovered. For example, TV minimization assumes sparsity of the gradient of the image to be reconstructed. Such assumptions are often not strictly satisfied in real world scenarios which again limits the theoretically achievable reconstruction quality. On the other hand, iterative reconstruction algorithms tend to be slower as they require repeated application of the PAT forward operator and its adjoint. Antholzer further proposed another three-layered S-net network for direct reconstruction for the sparse data problem, where the input is an image with artifacts and a real ground image obtained through a priori method. In simulation experiments, S-net can effectively eliminate artifacts caused by sparse data and greatly improve reconstruction efficiency compared to traditional image reconstruction methods [48–51]. The author also summarizes a deep network generally used for image enhancement after PAT image reconstruction. In the first step, the FBP algorithm (or another standard linear reconstruction method, using FBP as an example here) is applied to sparse data. In the second step, a deep CNN is applied to intermediate reconstruction, which outputs an image with almost no artificial artifacts. This can be explained as a deep network with FBP in the first layer and CNN in the remaining layers.

Image reconstruction is also an important part of functional imaging, including blood oxygen detection and various molecular detections. Due to the fact that hemoglobin is the main substance absorbed by human cells below 1000 nm, PAT can quantitatively detect hemoglobin ( $HbO_2$ ) and deoxyhemoglobin ( $HbR$ ). Since the oxygen saturation ( $sO_2$ ) of hemoglobin in normal tissue is higher than that in malignant tissue,  $sO_2$  is an important physiological index of the body [52–54].  $sO_2$  is defined as the fraction of  $HbO_2$  relative to the total hemoglobin concentration in the blood:

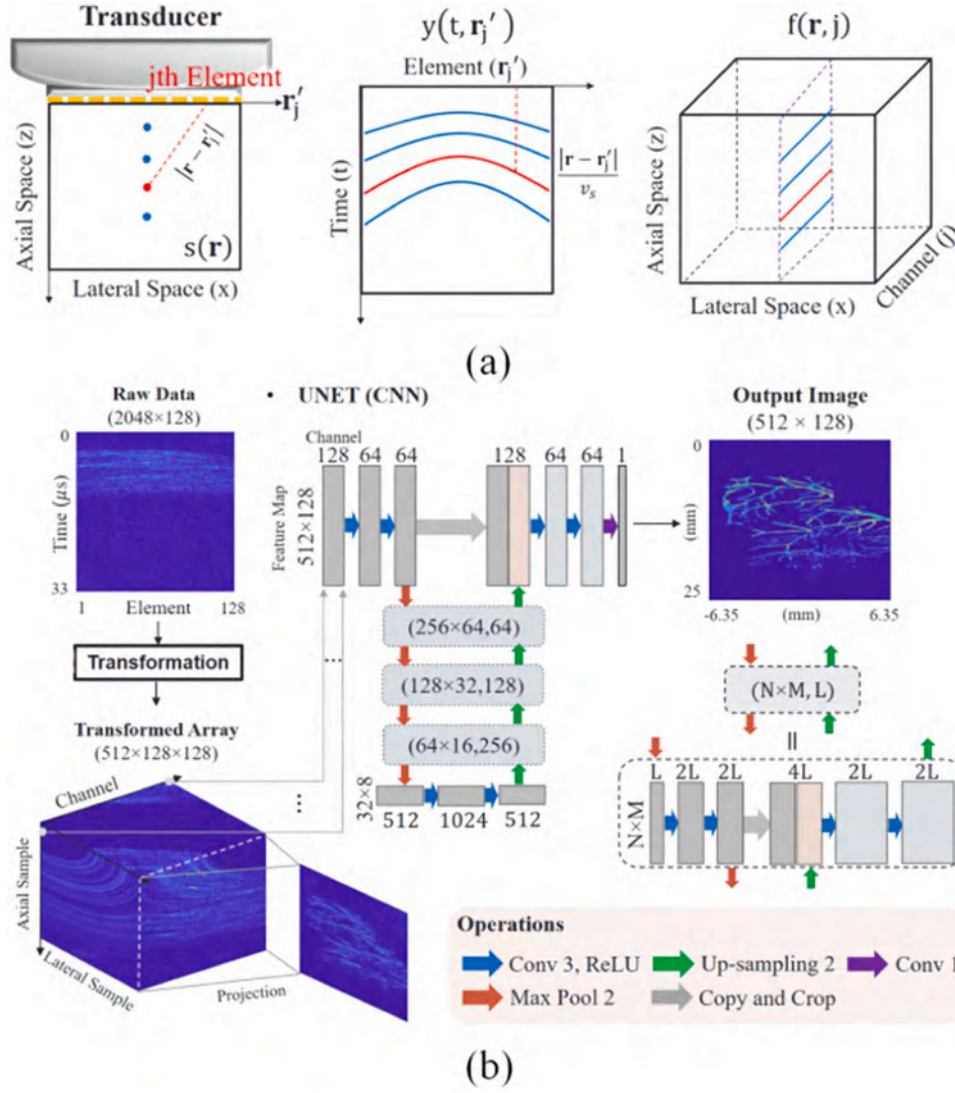


Fig. 8. System diagram. (a) Schematic of photoacoustic data acquisition; (b) CNN-Net.

$$sO_2(x, y) = \frac{C_{HbO_2}(x, y)}{C_{HbO_2}(x, y) + C_{HbR}(x, y)} \times 100\% \quad (2)$$

Here  $C_{HbO_2}$  and  $C_{HbR}$  represent the concentration of oxygenated and deoxygenated hemoglobin, respectively, while  $x$  and  $y$  denote the spatial position. According to the above formula, the basic principle of using photoacoustic tomography (PAT) for quantitative blood oxygen synthesis imaging is that  $HbO_2$  and  $HbR$  have significant absorption differences at different wavelengths of light. Similarly, quantitative spectroscopic photoacoustic imaging (QS-PAI) is an imaging technique that measures at multiple wavelengths of light to provide information related to molecular composition [55]. The aim is to convert multi-wavelength PA images into a final image that mainly highlights the quantitative and accurate estimation of chromophore spatial concentration changes in scattering media. The main problem with QS-PAI is essentially an inversion problem of light propagation operators. The current common two-stage inversion strategy can be summarized as follows: 1. determining the absorption coefficient; 2. determining the chromophore concentration. Due to the characteristics of scattering media, accurate nonlinear inversion of spatially structured light flux is difficult to achieve, and it is unrealistic to rely on strict conditions, such as known scattering coefficients and homogeneous background optical properties. In the inversion process, linear substitution instead of nonlinear inversion is used to determine the absorption coefficient by

using multi-wavelength PA images and light flux related to the absorption coefficient and scattering coefficient, which may result in large errors.

Cai et al. proposed the first deep learning framework Res-UNet for quantitative PA imaging [56]. Res-UNet takes the entire initial pressure image distributed at different wavelengths as input, so that reconstruction can best utilize all measurement signals. To prevent the degradation of deep networks, residual learning mechanism is adopted. In Res-UNet, comprehensive contextual information is extracted from multispectral initial pressure images to quantitatively estimate chromophore concentration or  $sO_2$ . The CNN architecture implemented using U-net is used to measure object contours, perform optical inversion, estimate the main absorbing chromophores and their absorption spectra, and perform linear decomposition.

Yang et al. proposed a deep residual and recursive neural network (DR2U-net) for quantitative estimation of hemoglobin oxygenation in photoacoustic imaging [57]. The proposed DR2U-net can extract flux distribution information from the optical absorption image using only two wavelengths of light in Monte Carlo simulations, and then generate quantitative  $sO_2$  images. Through testing on simulated biological tissues, the measured  $sO_2$  results have high accuracy, with an error as low as 1.27 %, compared to traditional linear mixing methods (48.76 %). In the network structure, deep networks can enrich feature information, so



the article uses residual connections mentioned above to solve possible gradient explosion and improve training accuracy [58]. Batch normalization is also used to accelerate convergence speed and reduce covariate shift. This approach effectively reduces the nonlinear effect of scattered light flux while increasing system robustness and reducing noise interference.

Rajendran and Pramanik proposed a novel deep learning architecture for tangential resolution in circular-scan photoacoustic tomography (PAT) imaging system [59]. The article uses a U-Net-based convolutional neural network combined with 9 residual blocks to improve the tangential resolution of PAT images. This is the first study to use a U-Net structure neural network for tangential resolution of PAT images. In general, in photoacoustic tomography, axial resolution does not change and is influenced by the detection bandwidth. However, tangential resolution will change with the size of the detector aperture. Especially when the aperture size is smaller, the tangential resolution is higher. However, if a small-aperture detector is used, the sensitivity of the sensor will decrease. Therefore, a large-aperture detector is the main choice for circular-scan PAT imaging systems. The proposed TARES network was implemented using Python 3.7 and TensorFlow v2.3 deep learning library [60]. The model was trained using simulated PA data and validated using experimental model data and human PA images [61–64]. The training model can detect data well and simulate body images of humans and animals.

Gao et al. proposed a U-Net-based convolutional neural network to extract effective photoacoustic information hidden in speckle patterns in a vascular network image dataset under porous media [65]. As shown in Fig. 9, human skull belongs to a typical multi-scattering medium, and traditional ultrasound imaging has many challenges in imaging deep and fine structures due to significant scattering of sound signals during excitation and reception. The article uses photoacoustic imaging principles and deep neural networks to solve the issues of frequency-domain wideband scattering in transcranial photoacoustic microvascular imaging and superposition of spatial domain main lobe and side lobe signals [66,67]. In short, the neural network can effectively extract valid information from highly blurred speckle patterns for rapid reconstruction of target images, providing broad application prospects in transcranial

photoacoustic imaging [68,69].

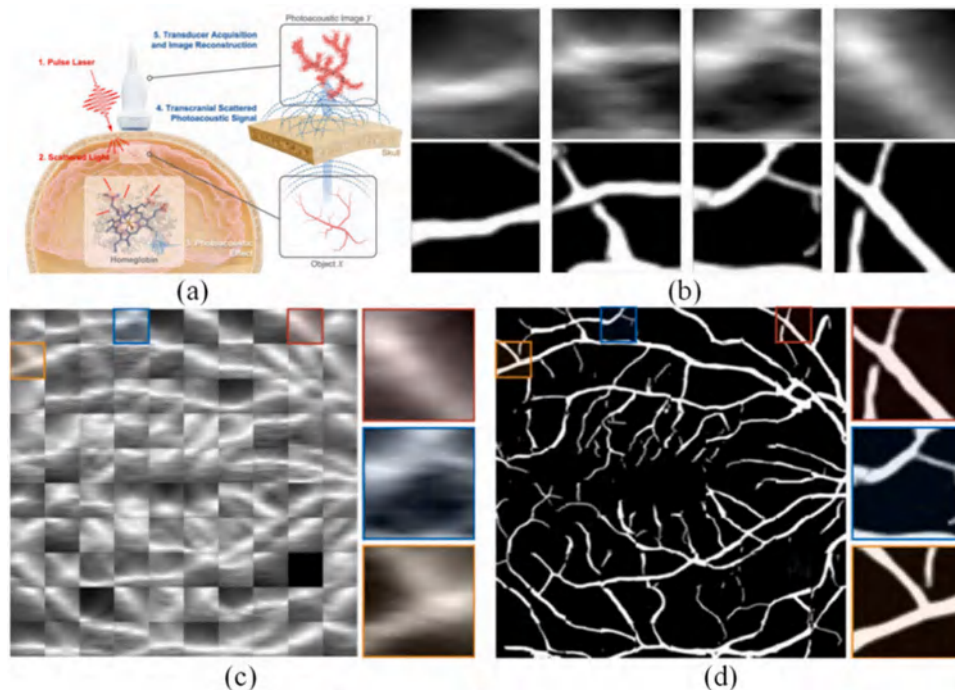
### 3.1.2. Photoacoustic microscopic imaging reconstruction

Zhou et al. proposed a method using ResNet to improve the quality of sparse PAM images [70], which can simultaneously maintain good image quality and accelerate image acquisition speed [71–73]. In this work, the dataset used was PAM images of oak and magnolia leaf veins. Immerse the leaves in a container with black ink for more than 7 h, then place them on a glass slide and seal them with silicone gel (GE sealant). For each PAM image, use an OR-PAM probe with a resolution of  $2 \mu\text{m}$ , consisting of a beam profiler and  $10 \times$  Beam expander measurement at  $256 \times$  Scan leaf samples at 256 scanning points with a scanning step of  $8 \mu\text{m}$ . Finally, a real image dataset of 268 original fully sampled PAM images was obtained. Corresponding low pixel images pass through  $2 \times$  And  $4 \times$  Downsampling acquisition.

The proposed ResNet structure is shown in Fig. 10(a). The authors used 16 residual blocks and 8 squeeze and excitation (SE) blocks as the key part of feature extraction. Inspired by SRGAN [74], the residual block shown in Fig. 10(b) can extract features well in the SR task. The SE block with channel attention mechanism (as shown in Fig. 10(c)) helps network convergence and performance. The "Upconv" block consists of  $2 \times$  upsampling layers and standard convolution layers (kernel size 3, filter number 256, stride 1). The Tanh activation function is used after the final output layer.

Zhao et al. proposed a multi-task residual dense network (MT-RDN) deep learning system and method [75]. The MT-RDN network adopts an innovative strategy combining multi-supervised learning, dual-channel sample collection, and reasonable weight allocation. The proposed deep learning method is combined with an improved OR-PAM system for application. This study obtained good images for the first time under ultra-low laser dose (reduced by 32 times). The network method aims to solve the challenges of image quality deterioration caused by low single-pulse laser energy and undersampling during high-speed imaging.

In the proposed system method, the original images (i.e., under-sampled images obtained under low excitation laser energy) are collected at 532 and 560 nanometer wavelengths and assigned to two different network input channels input1 and input2 respectively. The



**Fig. 9.** Schematic diagram and method comparison diagram. (a) the schematic diagram of transcranial photoacoustic imaging; (b) the reconstruction effect of DAS and this network on plaque respectively; (c) and (d) the reconstruction effect of DAS and this network on whole image respectively.



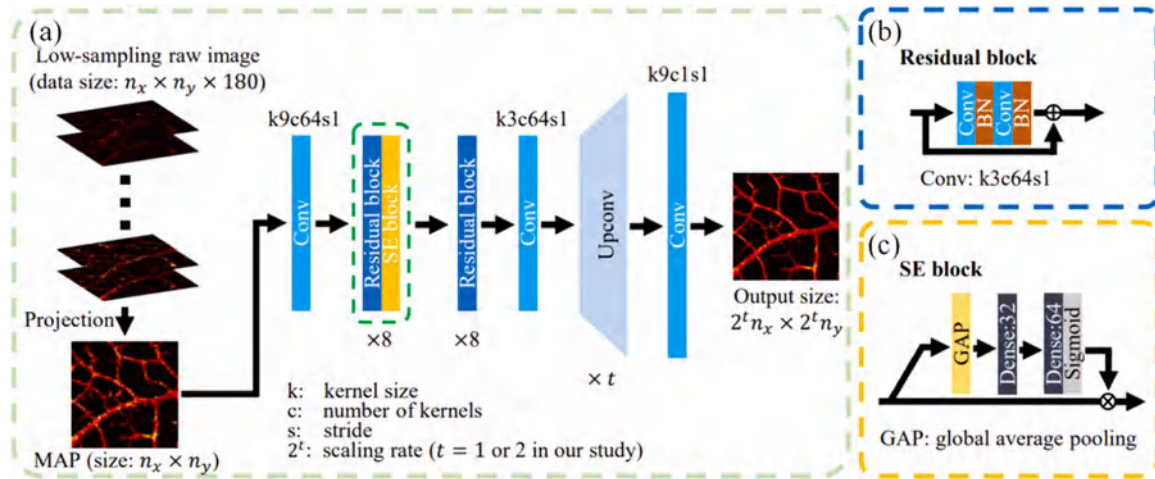


Fig. 10. Resnet Network diagram; (a) ResNet structure; (b) Residual Block structure; (c) Squeeze and Excitation (SE) Block.

low pixel input of the network is obtained by down-sampling the original image using 2x and 4x undersampling images at half of the single-pulse laser energy (i.e., ANSI limit of single-pulse laser energy), and then segmenting the original image as input to the MT-RDN network. The 2x undersampled image is cut into slices of  $100 \times 100$  pixels, and the 4x undersampled image is cut into slices of  $50 \times 50$  pixels. MT-RDN has three sub-networks. The first sub-network is used to process data input 1 (i.e., 532 nm data) to obtain output 1, and the second sub-network is used to process data input 2 (i.e., 560 nm data) to obtain output 2. Outputs 1 and 2 are further combined and processed by Sub-network 3 to obtain output 3. Ground truth images 1–3 are obtained from full-sampled images obtained at 532 nm and 560 nm ANSI limits of single-pulse laser energy, and ground truth images obtained using the Paivef method [76].

### 3.2. Photoacoustic image processing

The goal of image processing is to improve the quality and clarity of important details or targets in an image for specific applications by manipulating the image. Often, image enhancement is closely associated with the subsequent steps of photoacoustic image reconstruction. Image processing techniques such as noise reduction, smoothing, contrast stretching, sharpening, edge enhancement, and super resolution are commonly used to increase imaging readability and efficiency. These operations all belong to image processing, with the aim of improving the interpretability and effectiveness of the resulting image. Image processing often follows image reconstruction algorithms.

#### 3.2.1. Improvement of signal-to-noise ratio of photoacoustic images

In photoacoustic (PA) signals, the initially acquired PA signal and image often suffer from low signal-to-noise ratio (SNR) due to the weak amplitude of the PA signal and strong random noise from external instruments and the environment. In practice, the PA waves generated by low-cost, low-energy laser diodes are very weak and almost buried by noise. Additionally, deep tissue imaging is accompanied by severe attenuation, such as scattering, leading to the problem of low SNR in PA signals [77]. Consequently, the reconstructed PA images have poor quality with noise. Therefore, effective denoising techniques are required for reconstructing artifact-free PA images from measurements containing noise signals [78–80]. Although traditional Kalman filters (KF) [81,82] can remove Gaussian noise in the time domain [83–86], they lack adaptability under real-time estimation conditions due to their fixed model. The effectiveness of the traditional KF relies on the proper definition of two key parameters: the system noise matrix (Q) and the measurement noise matrix (R). However, it is often challenging to

obtain accurate statistical data for these parameters in practical situations. To overcome this challenge, there are existing methods for eliminating white noise. The most common one is data averaging, which has been used in PAI. However, it requires additional storage space for data and imposes high requirements on time [87]. In addition to white noise, electrical noise generated by the photoacoustic imaging system [88,89] and interference in the acquired photoacoustic signals can significantly degrade image contrast in multispectral photoacoustic tomography (MSOT).

He et al. proposed an attention enhanced GAN that uses an improved U-net generator to remove noise from PAM images [90]. The network does not need to manually select settings for different noisy images, but instead uses an attention enhanced generative adversarial network to extract image features and adaptively remove varying degrees of Gaussian, Poisson, and Rayleigh noise. The proposed method has been validated on both synthetic and real datasets, including phantom (leaf vein) and in vivo (mouse ear blood vessels and zebrafish pigment) experiments. The network structure diagram and denoising effect are shown in Fig. 12. To effectively capture features and distinguishing information with varying importance, an attention mechanism is applied in their network. Different from regular CNNs which may treat all information equally, attention blocks additionally introduce attention weights for different feature channels or spatial positions. Specifically, this method utilizes the attention block, i.e., the GC block, to enhance the attention to long range dependencies and that better handle unexpected noise instance of focusing on signal pixels. The detailed structure of the GC attention block includes  $1 \times 1$  convolutions and layer normalization. GC blocks are placed after each standard unit block of the encoder in the generator. Fig. 11(a) shows the GAN network structure diagram. The network structure includes a generator and a discriminator. Fig. 11(b) displays a comparison of the results of neural network imaging and other methods in the mouse ear vascular region. (Scale bar: 250  $\mu\text{m}$ . All images, excluding zoom images, share the same scale bar. The values in the colorbar indicate relative PA intensity). On the left side of Fig. 11(c) is the sample image before denoising, which includes mouse ear blood vessels, zebrafish pigment, and enlarged color box areas in the above samples. On the right is the denoised image, which includes mouse ear blood vessels, zebrafish pigment, and enlarged color box areas in the above image. (Scale: 500  $\mu\text{m}$ ).

#### 3.2.2. Improvement of photoacoustic image resolution

Deep learning methods can also be applied to improve the resolution of photoacoustic (PA) images. Traditional acoustic-resolution PA imaging systems are often limited to imaging resolutions on the order of 100 micrometers due to the optical diffraction limit and the acoustic

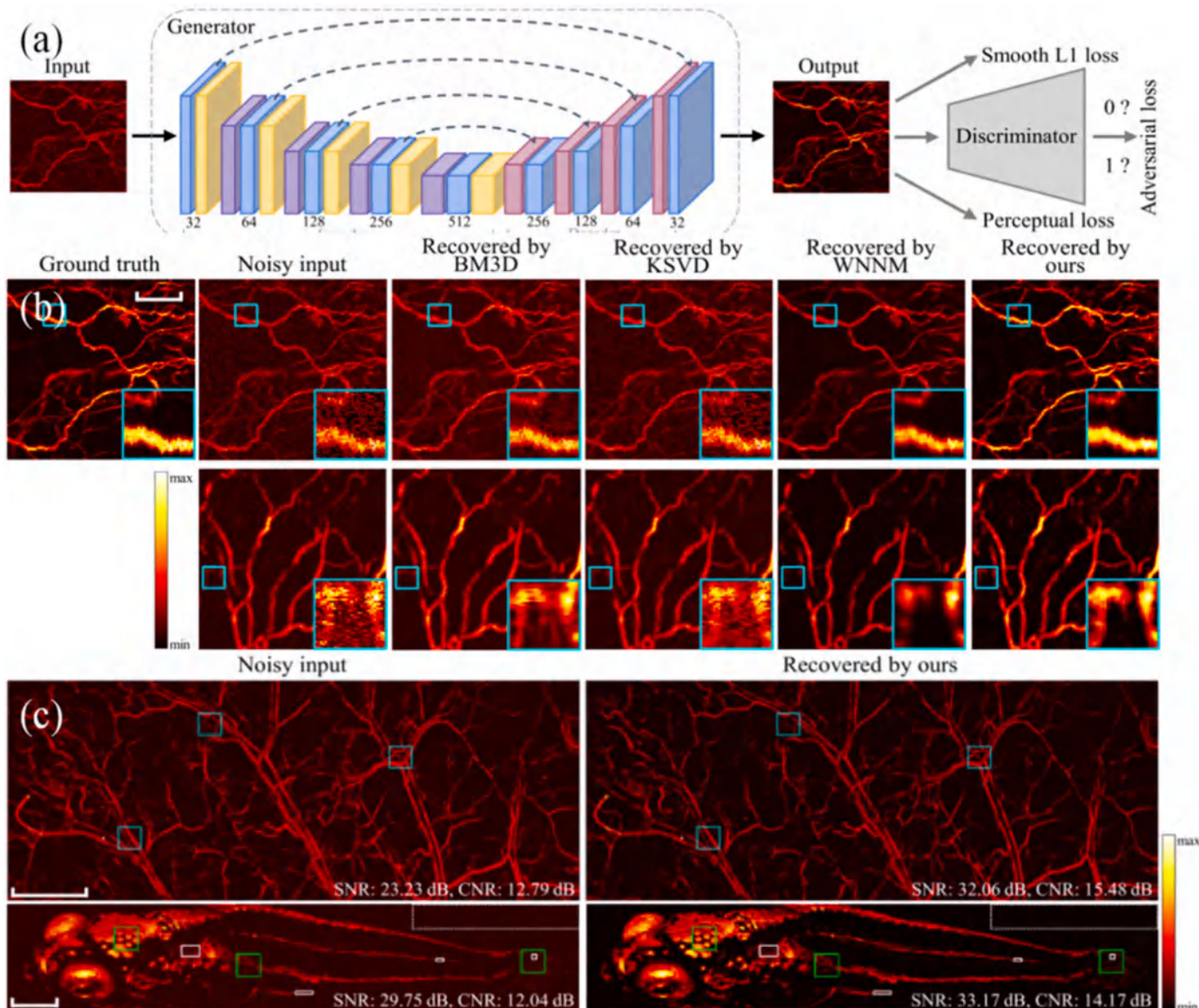


Fig. 11. Network structure diagram and denoising effect illustration. (a) Network structure diagram; (b) Representative results of the mouse ear blood vessel dataset acquired by in vivo experiment. Top row: a representative sample from the synthetic noisy dataset; bottom row: a representative sample from the real noisy dataset; (c) Demonstration of denoising effects on mouse ear vasculature and zebrafish pigment.

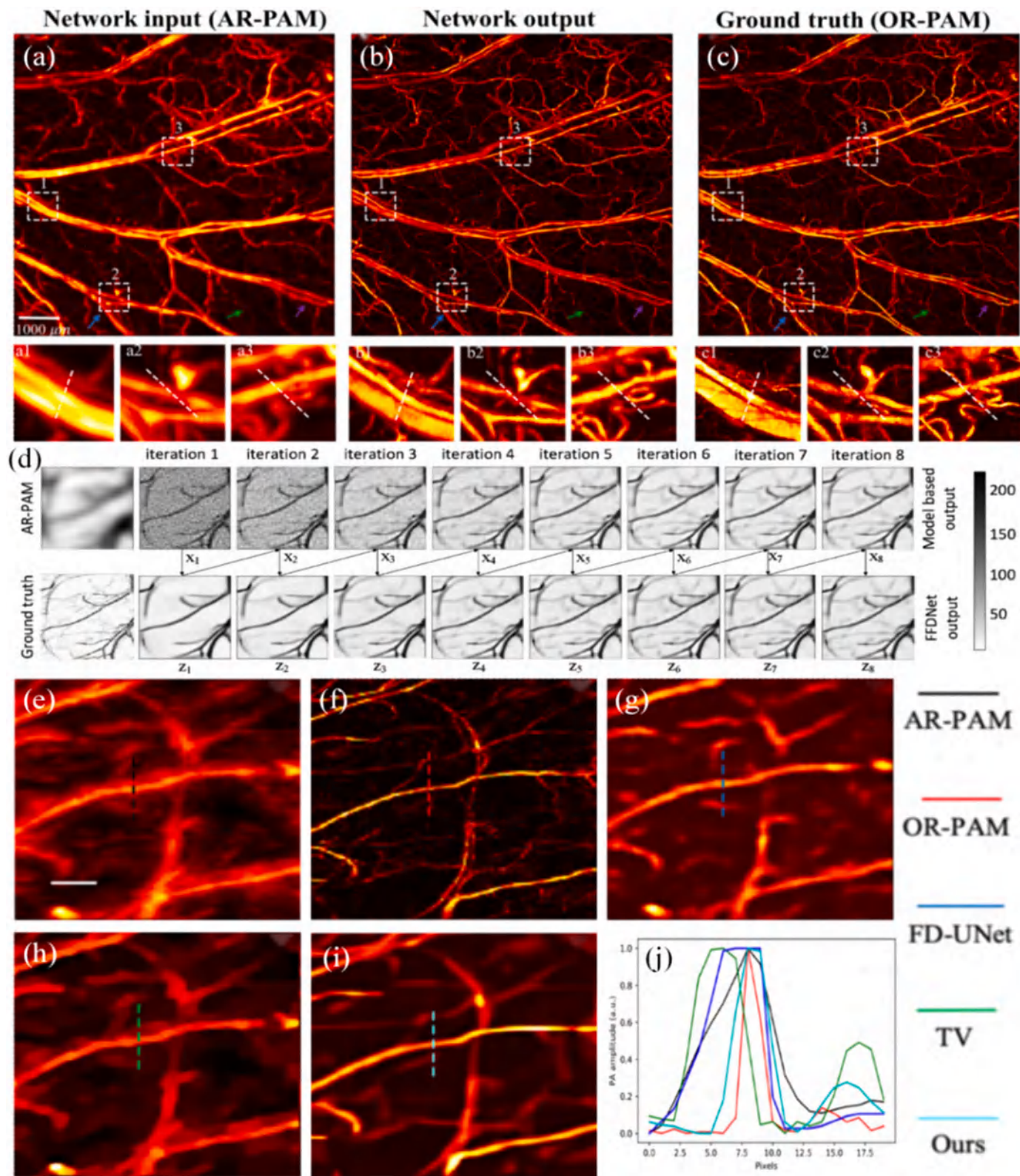
diffraction limit. On the other hand, optical-resolution imaging systems can achieve spatial depths of around 1 mm due to the optical diffraction limit but have limited applicability in clinical medicine. Similar to the post-processing methods for sparse data or array-angle-limited problems in PA tomography (PAT), deep learning has been widely used in super-resolution reconstruction of photoacoustic images by implementing end-to-end image optimization.

Cheng et al. proposed a deep-penetration high-resolution photoacoustic microscopy technique based on deep learning generative adversarial network (GAN) architecture [91]. This method employed Wasserstein GAN (WGAN) as the training network to learn from low-resolution absorption-reconstruction photoacoustic microscopy (AR-PAM) images towards high-resolution optical-resolution photoacoustic microscopy (OR-PAM) images at the same depth. In this WGAN network, the generator takes AR images as input and generates high-resolution images, which are then passed to the discriminator to determine their similarity to ground truth and high-resolution images. As mentioned earlier, this generative adversarial network involves an adversarial cooperative training between the generator (G) and the

discriminator (D): G generates an image that closely resembles the target image or its label to deceive D, while D provides feedback by discerning between real and generated images. In general, the network aims to minimize the mutual information difference (also known as Jensen-Shannon divergence) between the produced data and the real data. The article employed Wasserstein distance as the selected objective instead of Jensen-Shannon divergence to address the issues of vanishing gradients and model collapse in the generator [92–94]. The imaging results are shown in Fig. 12(a)–(c).

The degradation model of AR-PAM imaging is influenced by the imaging depth and the center frequency of the ultrasonic transducer, which may vary under different imaging conditions and cannot be processed using a single neural network model. To address this limitation, Zhang et al. proposed a supplementary framework that combines the advantages of model based and learning based methods and avoids their limitations, which can be used to enhance the image quality of AR-PAM images [95]. Firstly, a deep convolutional neural network is used to implicitly capture the image statistical and structural information of the target vascular image, thereby obtaining a Plug and Play (PnP) prior,





**Fig. 12.** WGAN network results for mouse ear vasculature and results of an adaptive enhancement method with a deep CNN prior. (a-c) WGAN network for mouse ear vasculature: (a) Network input AR-PAM image, (a1, a2, a3) enlarged regions selected by white dashed boxes; (b) Network output image, (b1, b2, b3) enlarged regions selected by white dashed boxes; (c) Ground truth OR-PAM image, (c1, c2, c3) enlarged regions selected by white dashed boxes; (d-i) Adaptive enhancement method with a deep CNN prior: (d) Example AR-PAM image enhancement in different iterations by model based equation (upper row) and FFDNet (bottom row); (e) AR-PAM imaging result; (f) OR-PAM imaging result; (g) Result enhanced using the FDU-Net on (e); (h) Enhancement result using the total variation algorithm on (e); (i) Result enhanced using the proposed algorithm on (e); (j) Signal intensity distribution along the vertical dashed line. (Scale bar: 1 millimeter).

while avoiding the process of designing complex manual regularization terms. Subsequently, this PnP prior is further inserted into the model based framework so that it can adaptively handle the variational degradation process. The proposed framework was first applied to simulation research and demonstrated its excellent performance and adaptability (with a huge dynamic range). In addition, some in vivo experiments were conducted to test the framework's ability to handle various real imaging scenes. The results show that this method can adaptively enhance AR-PAM images obtained across different imaging systems and depths, expanding the application scenarios of this method. This work adopted a combination of network optimization and degradation models, iterating eight times each other. The advantage of this approach is to use neural networks to correct artifacts generated by degradation methods, and model methods to correct images when the optimization effect of network images decreases, achieving the effect of improving image effectiveness and resolution. The iteration model and imaging results are shown in Fig. 12(d)–(j). In the three simulation scenarios created, the proposed algorithm achieved optimal performance in terms of PSNR and SSIM values; In vivo testing results using this algorithm showed significant increases in SNR and CNR values from 6.34 and 5.79, respectively, to 35.37 and 29.66, as shown in Fig. 12(j).

### 3.2.3. Photoacoustic image segmentation and recognition processing

Image detection and recognition involve the task of identifying specific elements in medical images [96–98]. In many cases, the images are three-dimensional, making efficient analysis crucial. The ability to differentiate and classify different elements is fundamental in medical image analysis, and image segmentation is a necessary method for processing medical images. Image segmentation has greatly benefited from the latest developments in deep learning. In image segmentation, the goal is to accurately delineate the contours of organs or anatomical structures, and methods based on convolutional neural networks (CNNs) have gradually become dominant in this field. Deep learning not only helps in selecting and extracting features but also aids in constructing new features [99–101]. Moreover, it can provide predictive models that not only diagnose diseases but also measure and predict targets, offering actionable insights to improve efficiency for medical professionals. There have been numerous successful examples of deep learning-assisted image processing in photoacoustic imaging, and the segmentation and recognition methods used in photoacoustic imaging can be applied to other medical imaging modalities as well.

Zhang et al. proposed an emerging deep learning-based method for breast cancer diagnosis in photoacoustic tomography (PAT) [96]. This method employed a preprocessing algorithm to enhance the quality and uniformity of input breast cancer images. Additionally, a transfer learning algorithm was utilized to address the issue of insufficient training data, resulting in improved classification performance. The network categorized existing breast cancer datasets into six classes based on the BI-RADS level, helping doctors better diagnose and treat cancer based on breast imaging reports and data system levels.

In magnetic resonance imaging (MRI) field, Wu et al. proposed an oriented novel attention-based glioma grading network (AGGN) [102]. By applying the dual-domain attention mechanism, both channel and spatial information can be considered to assign weights, which benefits highlighting the key modalities and locations in the feature maps. Multi-branch convolution and pooling operations are applied in a multi-scale feature extraction module to separately obtain shallow and deep features on each modality, and a multi-modal information fusion module is adopted to sufficiently merge low-level detailed and high-level semantic features, which promotes the synergistic interaction among different modality information. The results have demonstrated the effectiveness and superiority of the proposed AGGN in comparison to other advanced models, which also presents high generalization ability and strong robustness.

Li et al. proposed a feature learning enhanced convolutional neural network (FLE-CNN) for cancer detection from histopathology images

[103]. They built a highly generalized computer-aided diagnosis (CAD) system. The FLE-CNN included an information refinement unit employing depth- and point-wise convolutions is meticulously designed, where a dual-domain attention mechanism is adopted to focus primarily on the important areas. Experimental results demonstrate the merits of the proposed FLE-CNN in terms of feature extraction, which has achieved average sensitivity, specificity, precision, accuracy and F1 score of 0.9992, 0.9998, 0.9992, 0.9997 and 0.9992 in a five-class cancer detection task, and in comparison to some other advanced deep learning models, above indicators have been improved by 1.23 %, 0.31 %, 1.24 %, 0.5 % and 1.26 %, respectively.

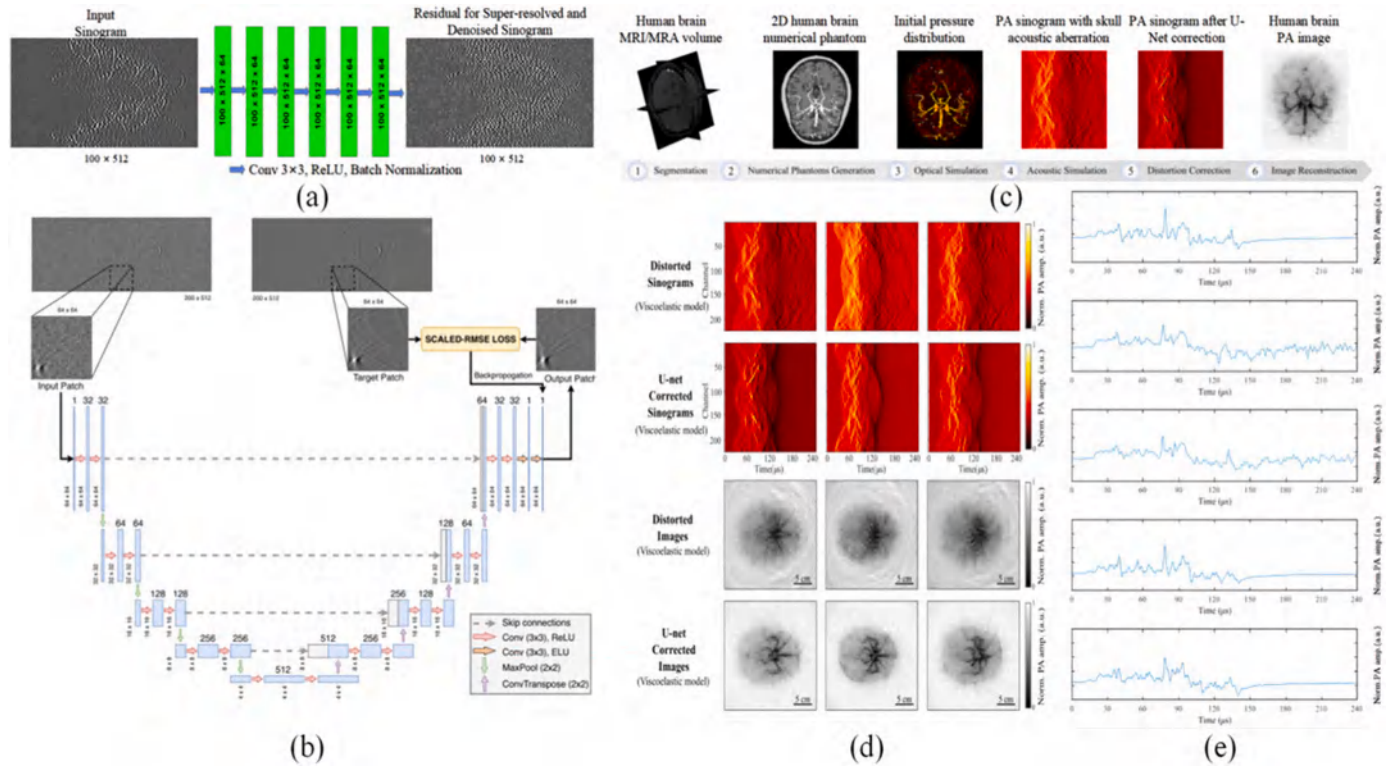
### 3.3. Photoacoustic signal processing

Awasthi et al. proposed a deep learning-based photoacoustic (PA) sinogram super-resolution denoising model [104]. The loss function of this model is scaled root mean square error, which is used for super-resolution, denoising, and bandwidth (BW) enhancement of PA signals acquired at region boundaries [105]. The network and method presented in the paper have the following characteristics: It is the first single network that performs super-resolution, denoising, and BW enhancement of PA data in the sinogram domain. Most deep learning networks are proposed in the image space to improve reconstructed images. This network exhibits inherent robustness and generalization abilities. It also demonstrates robustness when trained on numerical models. The improved structure can be used to enhance raw data (sinogram) acquired experimentally, improving the results of inverse problems and inherently reducing biases introduced by image reconstruction methods. The introduction of scaled root mean square loss function to train the network on sinogram data containing extremely low values can be extended to other applications with similar properties as PA data. Fig. 13(a) and (b) show the network flowchart and structure diagram.

Similarly, Zhang et al. also proposed using sinogram data as input to remove artifacts produced by photoacoustic tomography imaging [106]. In their work, a two-dimensional brain PA numerical phantom dataset was generated based on magnetic resonance angiography (MRA) and T1-weighted images from the *ixi* dataset. The dataset was then used as input to a U-net network for training. The simulated artifact images were corrected against prior high-resolution images, resulting in a trained network that effectively corrects the acoustic aberration caused by the skull. Fig. 13(c) illustrates the experimental workflow, (d) presents comparisons of three sets of simulated brain imaging sinogram maps before and after artifact removal, and (e) shows a comparison of normalized signals from one of the models.

In photoacoustic imaging, previous signal processing techniques have been found insufficient to eliminate the influence of electrical noise because they often rely on simplified models and fail to capture the complex characteristics of both the signal and the noise. Dehner et al. proposed a discriminative deep learning approach to separate electrical noise from the photoacoustic signal prior to image reconstruction as shown in Fig. 14 [107]. In Fig. 14(a), Data layout of a measured multispectral stack of sinograms. The depicted sinogram shows the recorded signals during a representative scan of a human breast lesion at 960 nm. Magnification of the marked signals, which were recorded prior to responses from tissue and thus are predominately comprised of electrical noise. Histogram and fitted Gaussian distribution ( $R^2 = 99.5\%$ ) for parts of the electrical noise with visually low amounts of parasitic noise (signals marked with the dashed rectangle) illustrating the characterization of the thermal noise of the system. In Fig. 14(b), there are Noisy sinogram from a representative scan of a human breast lesion. Electrical noise component inferred by the neural network on the left side. Denoised sinogram obtained by subtracting the above two. On the right side are Magnifications of the marked areas in the left charts. Quantitative evaluation of the denoising performance below. There are Comparison of the SNR distributions in simulated photoacoustic sinograms





**Fig. 13.** Application of Sinogram Graph as Network Input in PAT. (a) Network diagram proposed by Awasthi et al.; (b) Network structure. (c) The experimental flow chart proposed by Zhang et al., whose flow includes prior image segmentation, making skull simulation structure, optical simulation, acoustic simulation, sinogram from image training and finally obtaining the artifact free image; (d) Normalized PA sinograms and normalized DAS reconstructed human brain PAT images from viscoelastic media acoustic model. (e) Normalized PA signal is taken as the first channel of one of the skull simulation models. Signals are reference PA signal, PA signal with skull aberration obtained from fluid media acoustic model, PA signal with skull aberration obtained from viscoelastic media acoustic model, PA signal with skull aberration obtained from fluid media acoustic model after U-net Correction, PA signal with skull aberration obtained from viscoelastic media acoustic model after U-net correction.

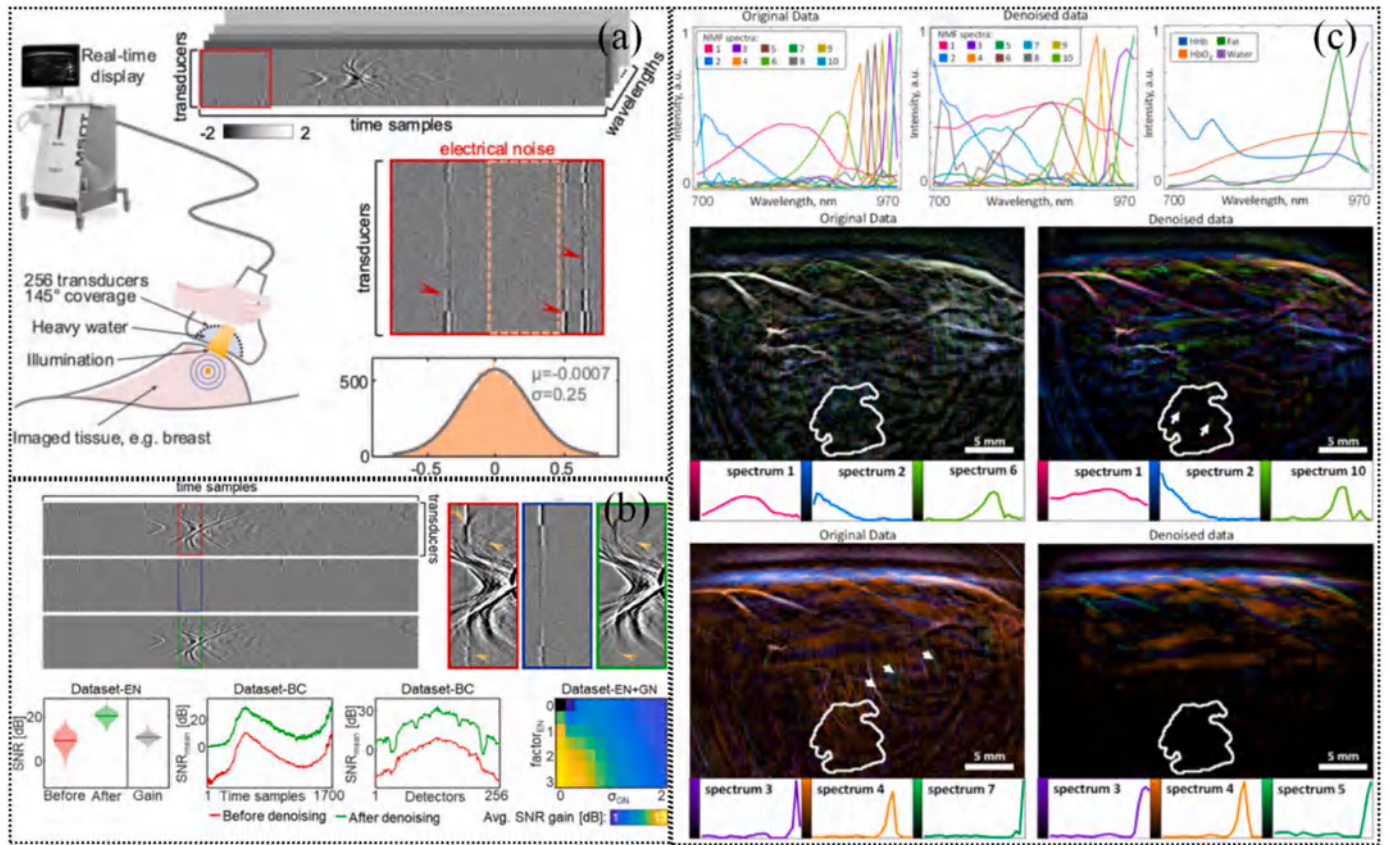
that are distorted by electrical noise before and after denoising. The mean gain is 10.9 dB. Evaluation of in vivo scans of human breast lesions. Mean SNR (SNR<sub>mean</sub>) of individual time samples. The average increase is 20.8 dB. Individual SNR<sub>mean</sub> of all detectors. The average increase is 22.4 dB. Average SNR gains (“SNR after denoising - SNR before denoising”) of the trained model for photoacoustic signals that were corrupted by a combination of measured electrical noise sinograms scaled with factor  $EN \in \{0, 0.5, 1, \dots, 3\}$ , and white Gaussian noise with standard deviations  $\sigma_{GN} \in \{0, 0.2, 0.4, \dots, 2\}$ . In Fig. 14(c), The first row shows the NMF spectra obtained from the original and denoised human breast lesion MSOT images from Dataset-BC, as well as the reference absorption spectra of the most prominent chromophores in breast tissue. The second and third rows show the before and after denoising comparison images, with the left column representing the pre-denoising image and the right column representing the post-denoising image. Visualizations of the NMF decomposition of a typical MSOT image are shown for pre-denoising and post-denoising at a depth of approximately 2 cm in malignant breast tumor. The contribution of the three spectra to the image is color-coded, with these spectra corresponding to the absorption spectra of hemoglobin (second row), fat, and water (third row). The tumor location determined from the ultrasound image is delineated by white contours. The proposed deep learning algorithm is based on two key features. Firstly, it learns the spatiotemporal correlation between the noise and the signal by using the entire photoacoustic sinogram as input. Secondly, it is trained on a large dataset consisting of experimentally acquired pure noise and synthetic photoacoustic signals.

The network utilizes a U-Net neural network architecture with 5 depths and 64 channel widths [108,109]. The basic expressive power of the network is reduced by estimating the interference signal. The L1 norm (L1 loss) is used as the loss function, and the ADAM optimizer

[110] is employed with a learning rate of 0.0001, decayed linearly to 0. The ADAM optimizer has a batch size of 1, and the momentum parameters are set to  $\beta_1 = 0.5$  and  $\beta_2 = 0.999$ . To speed up the learning process, Dehner et al. used a neural network input value of a constant 0.004, which brings the signal range to  $[-1; 1]$ . After passing through the artificial neural network, all signals were rescaled back to their original range. During training, the decomposition with the minimum loss over the data was validated, and the final model was selected.

Gutta et al. proposed a deep learning-based method for bandwidth enhancement of photoacoustic (PA) data [111]. During the process of photoacoustic tomography (PAT), the acquired PA signals from the surface of the tissue are always limited to a certain frequency band, while finer details of the image often reside in the high-frequency region of the PA signal. By utilizing a deep learning network, it is possible to effectively enhance the bandwidth of the PA signal without increasing computational complexity, thereby improving the contrast restoration and reconstruction quality of PA images. The network is trained with limited-bandwidth signals as input and full-bandwidth signals as output. The enhanced acoustic (PA) signal is then used as input to analysis reconstruction algorithms such as backprojection. This approach enables simple and efficient restoration of frequency band information but is limited by the constraints of prior algorithms and cannot achieve breakthroughs without real ground truth data.

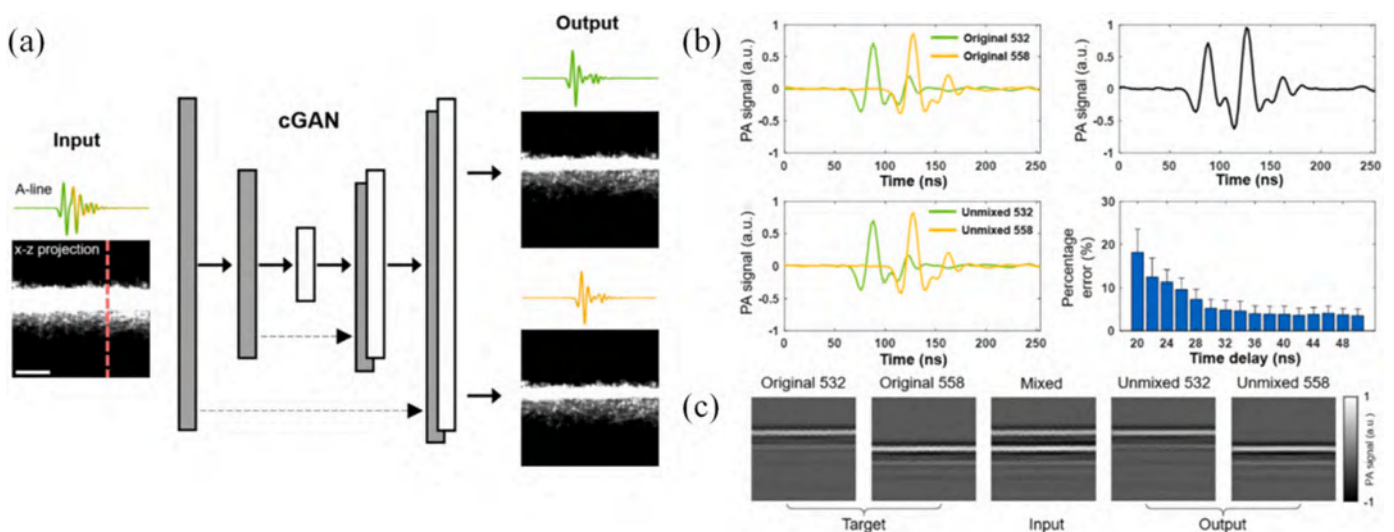
Zhou et al. proposed a conditional generative adversarial network (cGAN) for distinguishing the photoacoustic (PA) signals generated by fiber-separated dual-wavelength excitation lasers [112]. The time delay between the signals is approximately 38 nm. Improving the imaging speed of multi-parameter photoacoustic microscopy (PAM) is a key focus in this direction. To avoid temporal overlap, the A-line rate is limited to within 3 MHz due to the speed of sound in biological tissues. In



**Fig. 14.** The experiment is based on the schematic diagram of a handheld MSOT system, the evaluation diagram of signal signal-to-noise ratio, and the impact of denoising on the spectral content of photoacoustic images. (a) The experiment is based on the schematic diagram of the hand-held MSOT system Illustration of the scanning procedure using the handheld imaging probe of the test system. (b) Evaluation of the proposed denoising approach in the signal domain. (c) The impact of denoising on the spectral content of photoacoustic images.

order to achieve high-speed photoacoustic imaging of hemoglobin oxygen saturation, stimulated Raman scattering in optical fibers is widely used as a conventional method for generating dual-wavelength excitation at 558 nm from a commercially available 532 nm laser. However,

the length of the fiber used for efficient wavelength conversion is typically short, resulting in only a small time delay being obtained, leading to significant overlap in the acquired A-line signals at the two wavelengths. The proposed cGAN network allows for PAM excitation using



**Fig. 15.** Schematic of the cGAN and Dual wavelength A-scan signal graph. (a) The grayscale images are x – z projections of the three-dimensional dataset. A representative A-line, along the red dashed line, is shown above each of the x – z projection images. Scale bar: 300  $\mu$ m. (b) 532-nm excited A-line (green) and digitally delayed 558-nm excited A-line (yellow). a.u., arbitrary units. Digital sum of the two A-lines. Non-overlapping A-lines (green: 532nm and yellow: 558 nm) generated by the cGAN. Percentage error of sO<sub>2</sub> values as a function of time delay. The error bars represent standard deviations. (c) Representative B-scans as the target, input, and output of the cGAN, consisting of 256 original, mixed, and unmixed A-lines, respectively.



multi-spectral laser pulses, addressing the issue of insufficient energy in single-color laser pulses as shown in Fig. 15. This technology presents an innovative approach towards achieving ultra-high-speed multi-parameter PAM.

## 4. Summary and outlook

### 4.1. Summary

Deep learning, as a cutting-edge data acquisition technology, has been widely used in various fields of photoacoustic imaging such as image reconstruction, image processing, and signal processing. It can adjust parameters according to different network requirements to achieve a balance between strong robustness, high imaging speed, and artifact removal.

Compared with the iterative reconstruction method, the overall error of the denoising process after back projection reconstruction is higher than that of the iterative reconstruction method, especially in the case of limited angle scanning, which will produce mechanical artifacts, and the error at the imaging boundary is also more obvious. However, it is known that both deep learning-based image processing algorithms and traditional iterative reconstruction algorithms can overcome these mechanical artifacts by using prior mapping relationships. Compared with the post-processing methods of sparse data reconstruction in PAM, the deep learning-based reconstruction method not only has faster reconstruction speed but also greatly improves reconstruction efficiency. In addition, in terms of photoacoustic image processing, deep learning also has great advantages, including: higher flexibility and accuracy in handling complex and variable photoacoustic signals; outstanding performance in many image processing tasks such as image segmentation, classification, and reconstruction; ability to handle large amounts of training data, thereby improving model generalization ability and prediction performance; good scalability, allowing the model complexity to be adjusted based on task requirements and computing resources. Furthermore, deep learning models can utilize knowledge learned from other domains for transfer learning, thereby accelerating model training and improving performance.

The application prospects of deep learning in the field of photoacoustic imaging are vast, with continuously emerging network architectures for reconstruction algorithms in photoacoustic tomography, as well as for subsequent processing and forward sinogram processing. There is also ample room for improvement in various aspects of photoacoustic microscopy (PAM), such as scanning mechanism enhancements, excitation mechanism improvements, and post-processing techniques.

### 4.2. Data acquisition

Big data is the core of deep learning, but there is currently no open dataset for photoacoustic image reconstruction. In current experiments, the test sets used to train and validate the CNN are generally obtained through three methods: real human photoacoustic imaging results, imaging results of phantoms, and computer-simulated images. Since photoacoustic imaging has not been widely used for clinical diagnosis and treatment of diseases, the available clinical case data is severely lacking. The flexibility of phantom images is low, and the cost of making phantoms is high. Furthermore, it takes a long time to construct the data set required for deep learning. Computer simulation involves forward numerical simulation of the optical forward problem (the propagation of pulsed laser in tissue) and the acoustic forward problem (the process of tissue absorbing light energy, expanding due to heating, then emitting ultrasonic waves and propagating towards the tissue surface), obtaining the simulated initial sound pressure distribution map as the expected output image. The low-quality images reconstructed from limited-angle photoacoustic measurement data using standard reconstruction algorithms are used as input images to form the training set of the CNN. The

authenticity and effectiveness of the sample still need to be further discussed. In summary, there is currently a lack of large-scale open source training samples for photoacoustic imaging.

Gao et al. proposed a computing method of four-dimensional (4D) spectral-spatial imaging for PAD [113]. This method takes the optical and acoustic properties of heterogeneous skin tissues into account, which can be used to correct the optical field of excitation light, detectable ultrasonic field, and provide accurate single-spectrum analysis or multi-spectral imaging solutions of PAD for multilayered skin tissues. Simulation datasets obtained from the computational model were used to train neural networks to further improve the imaging quality of the PAD system. Most deep learning-based photoacoustic imaging needs thousands pairs of labeled input-output data to train the neural network, especially those applications in clinical skin imaging, which requires even larger amounts of data. However, in many cases the ground truth corresponding to the experimental data is inaccessible. This work as an efficient "learning from computational model" implemented an efficient method for obtaining simulation data. Considering human skin tissues are multilayered physiopathological structures with variability in optical absorption and acoustic impedance, this work further verify the simulation method from angles such as beam type, ultrasonic transducer performance, laser focusing position, and multi-spectral analysis. The article also proposes two neural networks trained on the dataset obtained through this method, namely the spread spectrum network and the enhanced imaging depth network. The feasibility of simulated datasets generated by computational modeling for neural network training was also demonstrated, helping to solve the major challenge of deep learning techniques in photoacoustic skin imaging that cannot obtain ground truth in many cases, with the potential to further improve the imaging quality of the PAD system through image reconstruction, information processing, and artificial intelligence methods as shown in Fig. 16(b–c). Fig. 16(a) verifies the effectiveness of the simulation model under different luminous flux conditions of 6 mJ/cm<sup>2</sup>, 12 mJ/cm<sup>2</sup>, and 18 mJ/cm<sup>2</sup>. Fig. 16(b) shows the results of palm skin imaging optimized by two networks.

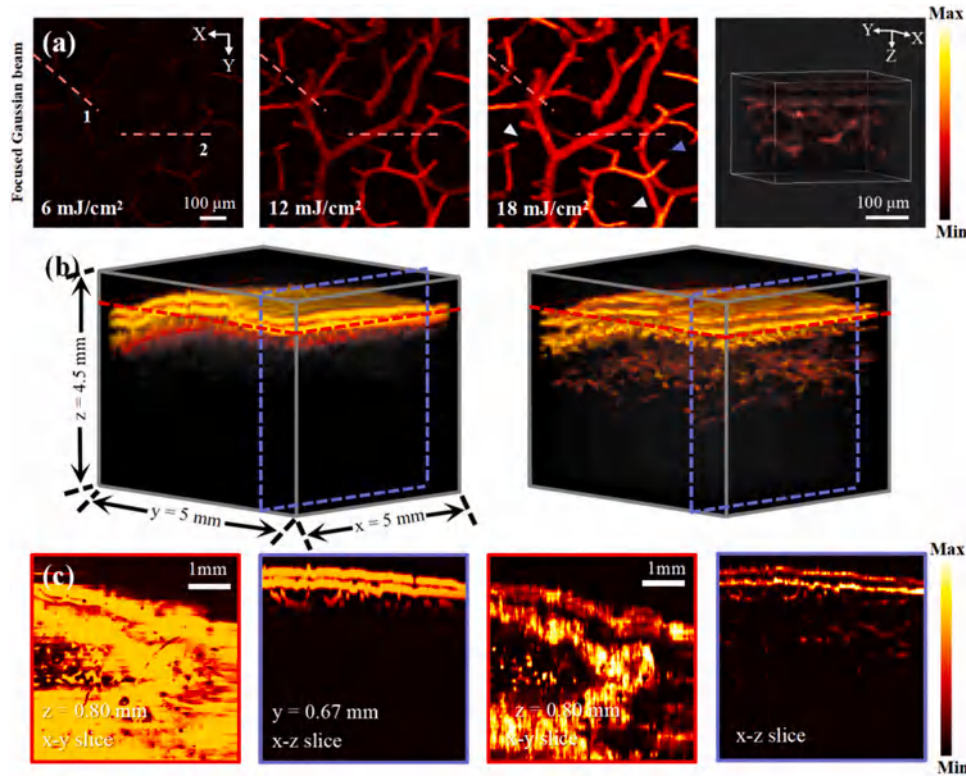
### 4.3. Interpretability

In addition to the lack of widely available and specialized data, the reliability and interpretability of deep learning methods are also receiving increasing attention. Lisboa et al. published a review of interpretability discussions in machine learning in 2020 [114], in which they classified interpretability discussions corresponding to the development of machine learning.

The article further proposes that there is currently no complete consensus on how to evaluate the quality of interpretable or interpretable methods. The evaluation methods that can explain ML include the "Real Humans in Real Tasks" proposed by Doshi Velez and Kim and the "AI rationalization" proposed by Ehsan et al. The quality of a given explanation needs to be evaluated in the context of its task, measuring the extent to which the explanation promotes and improves decision-making.

Salahuddin et al. published a review of interpretable methods for deep neural networks in medical image analysis in 2022 [115]. The article pointed out that interpretable artificial intelligence (XAI) refers to an AI solution that can provide some details about its functionality in a way that end users can understand. At present, the interpretability of deep neural networks is widely defined as attempting to explain the decision-making process of a model in a way that can be understood by end users.

Dai et al. [116] used a conceptual alignment deep autoencoder to analyze tongue images that represent different body constituent types based on traditional Chinese medicine principles. Koh et al. introduced Concept Bottleneck Models for osteoarthritis grading and used 10 clinical concepts such as joint space narrowing, bone spurs, calcification, etc. In their study, Dai et al. employed a novel deep autoencoder with



**Fig. 16.** The model adopts focused Gaussian beam images with different power densities of 532 nm wavelength incident beams and network generalization test results. (a) When the power densities are 6 mJ/cm<sup>2</sup>, 12 mJ/cm<sup>2</sup>, and 18 mJ/cm<sup>2</sup> and he three dimensional imaging result; (b) 3D PA image of palm skin and 3D PA image obtained after the spread spectrum network and the depth enhanced network processing (c) The corresponding color slices in the figure.

conceptual alignment to investigate tongue images, which are representative of diverse body constituent types according to the principles of traditional Chinese medicine. The utilization of this approach enabled a comprehensive analysis and interpretation of the underlying features associated with each body type.

Furthermore, Koh et al. [117] presented an innovative framework known as Concept Bottleneck Models for the purpose of osteoarthritis grading. This model incorporated ten important clinical concepts including joint space narrowing, bone spurs, calcification, among others. By leveraging these concepts, the researchers were able to establish a robust and informative grading system for the evaluation of osteoarthritis severity.

The above two are both based on the perspective of conceptual interpretability, and there are also more interpretable classifications, including: Case based models, Counter actual interpretation, Language description et al.

#### 4.4. Conclusion and outlook

In the switchable optical and acoustic resolution photoacoustic endomicroscope proposed by Ma et al. in 2020 [118], high-resolution imaging of the surface and deep layers is achieved by switching between optical and acoustic resolution systems at different depths in the skin. We can contemplate by acquiring a dataset from this system, the high-resolution surface images and deep-layer images are combined and fed into a deep neural network, enabling the high-resolution images to learn depth information from the deep-layer images, and the deep-layer images to learn resolution from the high-resolution images. This approach may ultimately lead to a single system that combines the advantages of both types of photoacoustic microscopy. There is also the possibility of mutual learning between systems with different numerical apertures (NA) and corresponding scanning mechanisms, or between different excitation wavelengths. Can we achieve complementary effects

between penetration depth and imaging resolution? These are all worth considering. Furthermore, in the field of photoacoustic endoscopy, the lack of corresponding datasets has limited the widespread use of deep learning methods. Therefore, exploring deep neural network-based approaches for photoacoustic endoscopy is also an important area of research.

In the previous article [94], the example of learning from low resolution AR-PAM images to high-resolution OR-PAM images can improve a certain imaging performance and expand the applicability of the system through learning between different imaging systems. In the field of photoacoustic imaging microscopy, AR-PAM and OR-PAM are complementary in imaging depth and resolution, OR-PAM can currently achieve an imaging depth of around 1.5 mm, with resolution at the micron or submicron level, while AR-PAM has an imaging depth of over 10 mm, but the corresponding resolution also has an order of magnitude attenuation. If a prior method can be used to obtain a prior of OR-PAM images at the same depth, the image features of this prior can be retained through a neural network method and applied in the corresponding AR-PAM system, that is, the results of the AR-PAM imaging system can be obtained through a neural network, and a high depth AR-PAM image with corresponding OR-PAM resolution can be obtained. Compared with general prior methods, the biggest advantage of OR-PAM prior is that it preserves the basic features of photoacoustic images, and its feedback signal composition is also ultrasound. This brings great convenience to the preservation of image features of tissue signal strength and phase. It is obvious that the combined imaging system can effectively improve imaging quality and obtain high-quality images at corresponding depths that were previously difficult to obtain. The corresponding potential mutual learning work can be envisioned. For example, OR-PAM learns imaging depth from AR-PAM images. Although the resolution of AR-PAM images is not as good as that of OR-PAM, it is possible to learn the intensity of AR-PAM signals from the perspective of photoacoustic signals by preserving the features of OR-PAM images, aiming to discover



weak photoacoustic signals in deep tissues. Due to the complementarity between AR-PAM and OR-PAM in the field of photoacoustic microscopy, it can be imagined that their mutual learning will become a reality in the near future. Photoacoustic tomography, on the other hand, has a higher imaging depth, and its imaging speed and imaging range are significantly different from microscopic systems. Its image features and system application scenarios are also inconsistent with microscopic systems. Therefore, the mutual learning between fault systems and microscopic systems still requires further development and integration of photoacoustic imaging. Not only is there mutual learning within photoacoustic imaging, but this approach can also be applied in bimodal imaging systems and in conditions of different system parameters. Bimodal imaging refers to the combination of two or more imaging techniques to obtain different types of information simultaneously or sequentially for image reconstruction and analysis. Mutual learning can be extended to, for example, mutual learning between ultrasound imaging and photoacoustic imaging, mutual learning between 1064 nm and 532 nm wavelength systems, and mutual learning between different NA systems. In summary, the ultimate goal of potential mutual learning currently lies in improving the system's penetration ability or imaging quality, provided that the two imaging systems are close or the imaging results can learn from each other.

The achievements of deep learning in photoacoustic imaging are undeniable, such as its applications in image reconstruction, signal-to-noise ratio improvement, and super-resolution. These achievements provide new ideas and methods for the development of photoacoustic imaging technology. However, there are still some challenges and limitations for deep learning models in photoacoustic imaging. For example, deep learning models require high training data demands, requiring a large amount of labeled data and computing resources. Although many network methods, such as U-Net structure and unsupervised learning, attempt to solve the data problem, there is still considerable room for improvement. Additionally, interpretability of deep learning models is also an issue that needs to be addressed.

In future research, we can try to further optimize the performance and interpretability of deep learning models to better meet the application requirements of photoacoustic imaging technology in clinical medicine and life sciences. At the same time, we can also explore combining deep learning with other technologies to discover more potential applications. As mentioned earlier, the work of Zhang et al. [95] is a good application and extension of neural network methods. For the interpretability of neural network methods, the author proposed a new approach that combines network and model methods. The model method corrected the image content forward, while the neural network corrects image artifacts. This alternating iteration method greatly improved the interpretability of neural networks. However, the drawback of this method is that the training difficulty and reconstruction time of the network have increased. Perhaps a more efficient network structure can be used as an alternative to iterative methods to improve imaging speed. In short, the idea of combining the principle of preserving models with neural networks is worth learning and continuing. Can deep learning also be better applied and explained from the perspective of photoacoustic signals? In photoacoustic tomography, the quality of the sine wave of the signal determines the quality of the reconstructed image. Unlike general image learning neural networks such as Awasthi [104] and Zhang [106], training photoacoustic signals to achieve signal amplification and denoising is also a way of applying deep learning. In future work, a collaborative learning approach can be envisioned. The signal learning network provides deep imaging signals and amplifies them, while the image learning network provides deep image features to achieve joint learning, discover deep structures, and efficiently image.

This review focuses on deep learning enabled photoacoustic imaging, and analyzes recent deep learning work from four perspectives: photoacoustic imaging PAT reconstruction, PAM reconstruction, image processing, and signal processing. The article also starts from neural network structures such as U-Net, GAN network, and Dense Block,

organizing their early work in the field of biomedical imaging, and introducing readers to common neural network structures and their origins in the biomedical field. Finally, the article summarizes the analysis and summary of deep learning in improving imaging capabilities from four perspectives, proposes the current problems and difficulties of neural networks, and further provides ideas for solving the problems. In summary, the rapid development of neural networks has continuously empowered photoacoustic imaging and even biomedical imaging in recent years. They have made epoch-making contributions to the depth and quality of imaging results, as well as to the improvement of imaging system efficiency and imaging speed.

#### CRediT authorship contribution statement

**Chen Qian:** Supervision. **Zuo Chao:** Funding acquisition. **Ma Hai-gang:** Conceptualization, Data curation, Investigation, Supervision, Writing – review & editing. **Wei Xiang:** Investigation, Methodology, Writing – original draft, Writing – review & editing. **Huang Qinghua:** Resources. **Feng Ting:** Conceptualization.

#### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Data Availability

No data was used for the research described in the article.

#### Acknowledgements

This work was supported by National Natural Science Foundation of China (62275121, 12204239, 12326609, 62071382, 62227818), Youth Foundation of Jiangsu Province (BK20220946), Fundamental Research Funds for the Central Universities (30923011024), Jiangsu Provincial Basic Research Program Frontier Leading Special Project (BK20192003).

#### References

- [1] Yang C., Lan H., Gao F. Accelerated photoacoustic tomography reconstruction via recurrent inference machines[C]//2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2019: 6371–6374.
- [2] P. Beard, Biomedical photoacoustic imaging, *Interface Focus* 1 (4) (2011) 602–631.
- [3] C. Yang, H. Lan, F. Gao, et al., Review of deep learning for photoacoustic imaging, *Photoacoustics* 21 (2021) 100215.
- [4] I. Goodfellow, Y. Bengio, A. Courville, *Deep learning*, MIT press., 2016.
- [5] Q. Huang, H. Tian, L. Jia, et al., A review of deep learning segmentation methods for carotid artery ultrasound images, *Neurocomputing* (2023) 126298.
- [6] Y. Guo, Y. Liu, A. Oerlemans, et al., Deep learning for visual understanding: a review, *Neurocomputing* 187 (2016) 27–48.
- [7] A. Salehi, M. Balasubramanian, DDCNet: deep dilated convolutional neural network for dense prediction, *Neurocomputing* 523 (2023) 116–129.
- [8] J. Schmidhuber, Deep learning in neural networks: an overview, *Neural Netw.* 61 (2015) 85–117.
- [9] K. Kawaguchi, J. Huang, L.P. Kaelbling, Effect of depth and width on local minima in deep learning, *Neural Comput.* 31 (7) (2019) 1462–1498.
- [10] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation[C]//medical image computing and computer-assisted intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015. Proceedings, Part III 18, Springer International Publishing, 2015, pp. 234–241. October 5–9, 2015.
- [11] Du G., Cao X., Liang J., et al. Medical image segmentation based on u-net: A review[J]. *Journal of Imaging Science and Technology*, 2020.
- [12] N. Man, S. Guo, K.F.C. Yiu, et al., Multi-layer segmentation of retina OCT images via advanced U-net architecture, *Neurocomputing* 515 (2023) 185–200.
- [13] Z. Huang, J. Miao, H. Song, et al., A novel tongue segmentation method based on improved U-Net, *Neurocomputing* 500 (2022) 73–89.
- [14] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, et al., 3D U-Net: learning dense volumetric segmentation from sparse annotation[C]//Medical Image Computing

- and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016. Proceedings, Part II 19, Springer International Publishing, 2016, pp. 424–432. October 17–21, 2016.
- [15] F. Isensee, P. Kickingereder, W. Wick, et al., No new-net[C]/Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018. Revised Selected Papers, Part II 4, Springer International Publishing, 2019, pp. 234–244. September 16, 2018.
- [16] X. Xiao, S. Lian, Z. Luo, et al., Weighted res-unet for high-quality retina vessel segmentation[C]/2018 9th international conference on information technology in medicine and education (ITME), IEEE (2018) 327–331.
- [17] S. Guan, A.A. Khan, S. Sikdar, et al., Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal, IEEE J. Biomed. Health Inform. 24 (2) (2019) 568–576.
- [18] Han Y.S., Yoo J., Ye J.C. Deep residual learning for compressed sensing CT reconstruction via persistent homology analysis[J]. arXiv preprint arXiv: 1611.06391, 2016.
- [19] S. Hakakzadeh, Z. Kavehvasht, M. Pramanik, Artifact removal factor for circular-view photoacoustic tomography[C]/2022 IEEE International Ultrasonics Symposium (IUS), IEEE (2022) 1–4.
- [20] Y. Lin, S. Kou, H. Nie, et al., Deep learning based on co-registered ultrasound and photoacoustic imaging improves the assessment of rectal cancer treatment response, Biomed. Opt. Express 14 (5) (2023) 2015–2027.
- [21] S. Guan, A.A. Khan, S. Sikdar, et al., Limited-view and sparse photoacoustic tomography for neuroimaging with deep learning, Sci. Rep. 10 (1) (2020) 8510.
- [22] H. Lan, K. Zhou, C. Yang, et al., Hybrid neural network for photoacoustic imaging reconstruction[C]/2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE (2019) 6367–6370.
- [23] M. Guo, H. Lan, C. Yang, J. Liu, et al., As-net: fast photoacoustic reconstruction with multi-feature fusion from sparse data, IEEE Trans. Comput. Imaging 8 (2022) 215–223.
- [24] J. Meng, X. Zhang, L. Liu, et al., Depth-extended acoustic-resolution photoacoustic microscopy based on a two-stage deep learning network, Biomed. Opt. Express 13 (8) (2022) 4386–4397.
- [25] Ledig C., Theis L., Huszár F., et al. Photo-realistic single image super-resolution using a generative adversarial network[C]/Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4681–4690.
- [26] T. Carvalho, E.R.S. De Rezende, M.T.P. Alves, et al., Exposing computer generated images by eye's region classification via transfer learning of VGG19 CNN[C]/2017 16th IEEE international conference on machine learning and applications (ICMLA), IEEE (2017) 866–870.
- [27] V. Rajnikanth, A.N. Joseph Raj, K.P. Thanaraj, et al., A customized VGG19 network with concatenation of deep and handcrafted features for brain tumor detection, Appl. Sci. 10 (10) (2020) 3429.
- [28] Wang X., Yu K., Wu S., et al. Esrgan: Enhanced super-resolution generative adversarial networks[C]/Proceedings of the European conference on computer vision (ECCV) workshops. 2018: 0–0.
- [29] Szegedy C., Ioffe S., Vanhoucke V., et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]/Proceedings of the AAAI conference on artificial intelligence. 2017, 31(1).
- [30] F. He, T. Liu, D. Tao, Why resnet works? residuals generalize, IEEE Trans. Neural Netw. Learn. Syst. 31 (12) (2020) 5349–5362.
- [31] H. Lin, S. Jegelka, Resnet with one-neuron hidden layers is a universal approximator, Adv. Neural Inf. Process. Syst. (2018) 31.
- [32] D. Waibel, J. Gröhl, F. Isensee, et al., Reconstruction of initial pressure from limited view photoacoustic images using deep learning[C]/Photons Plus Ultrasound: Imaging and Sensing 2018, SPIE 10494 (2018) 196–203.
- [33] J. Schwab, S. Antholzer, R. Nuster, et al., Deep learning of truncated singular values for limited view photoacoustic tomography[C]/Photons Plus Ultrasound: Imaging and Sensing 2019, SPIE 10878 (2019) 254–262.
- [34] H. Abdi, Singular value decomposition (SVD) and generalized singular value decomposition, Encycl. Meas. Stat. 907 (2007) 912.
- [35] J. Feng, J. Deng, Z. Li, et al., End-to-end Res-Unet based reconstruction algorithm for photoacoustic imaging, Biomed. Opt. Express 11 (9) (2020) 5321–5340.
- [36] T. Tong, W. Huang, K. Wang, et al., Domain transform network for photoacoustic tomography from limited-view and sparsely sampled data, Photoacoustics 19 (2020) 100190.
- [37] E.M.A. Anas, H.K. Zhang, C. Audigier, et al., Robust photoacoustic beamforming using dense convolutional neural networks[C]/Simulation, Image Processing, and Ultrasound Systems for Assisted Diagnosis and Navigation: International Workshops, POCUS 2018, BIVPCS 2018, CuRIOUS 2018, and CPM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16–20, 2018. Proceedings, Springer International Publishing, 2018, pp. 3–11.
- [38] K. Johnstonbaugh, S. Agrawal, D. Abhishek, et al., Novel deep learning architecture for optical fluence dependent photoacoustic target localization[C]/Photons Plus Ultrasound: Imaging and Sensing 2019, SPIE 10878 (2019) 95–102.
- [39] D. Allman, A. Reiter, M.A.L. Bell, Photoacoustic source detection and reflection artifact removal enabled by deep learning, IEEE Trans. Med. Imaging 37 (6) (2018) 1464–1477.
- [40] M. Xu, L.V. Wang, Universal back-projection algorithm for photoacoustic computed tomography, Phys. Rev. E 71 (1) (2005) 016706.
- [41] J. Provost, F. Lesage, The application of compressed sensing for photo-acoustic tomography, IEEE Trans. Med. Imaging 28 (4) (2008) 585–594.
- [42] M.W. Kim, G.S. Jeng, I. Pelivanov, et al., Deep-learning image reconstruction for real-time photoacoustic system, IEEE Trans. Med. Imaging 39 (11) (2020) 3379–3390.
- [43] H. Lan, C. Yang, D. Jiang, et al., Reconstruct the photoacoustic image based on deep learning with multi-frequency ring-shape transducer array[C]/2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE (2019) 7115–7118.
- [44] M.T. McCann, K.H. Jin, M. Unser, Convolutional neural networks for inverse problems in imaging: a review, IEEE Signal Process. Mag. 34 (6) (2017) 85–95.
- [45] C. Cai, K. Deng, C. Ma, et al., Bell, Photo-acoustic neural network for optical inversion in quantitative photoacoustic imaging, Opt. Lett. 43 (12) (2018) 2752–2755.
- [46] J. Bell, What is machine learning? Mach. Learn. City.: Appl. Archit. Urban Des. (2022) 207–216.
- [47] S. Antholzer, M. Haltmeier, J. Schwab, Deep learning for photoacoustic tomography from sparse data, Inverse Probl. Sci. Eng. 27 (7) (2019) 987–1005.
- [48] L.V. Wang, Tutorial on photoacoustic microscopy and computed tomography, IEEE J. Sel. Top. Quantum Electron. 14 (1) (2008) 171–179.
- [49] D. Allman, A. Reiter, M.A.L. Bell, Photoacoustic source detection and reflection artifact removal enabled by deep learning, IEEE Trans. Med. Imaging 37 (6) (2018) 1464–1477.
- [50] H. Shan, G. Wang, Y. Yang, Accelerated correction of reflection artifacts by deep neural networks in photo-acoustic tomography, Appl. Sci. 9 (13) (2019) 2615.
- [51] P.C.M. Van Zijl, S.M. Eleff, J.A. Ulatowski, et al., Quantitative assessment of blood flow, blood volume and blood oxygenation effects in functional magnetic resonance imaging, Nat. Med. 4 (1998) 159–167.
- [52] B. Cox, J.G. Laufer, S.R. Arridge, et al., Quantitative spectroscopic photoacoustic imaging: a review, J. Biomed. Opt. 17 (6) (2012), 061202-061202.
- [53] M. Li, Y. Tang, J. Yao, Photoacoustic tomography of blood oxygenation: a mini review, Photoacoustics 10 (2018) 65–73.
- [54] L.J. Rich, M. Seshadri, Photoacoustic imaging of vascular hemodynamics: validation with blood oxygenation level-dependent MR imaging, Radiology 275 (1) (2015) 110.
- [55] B.T. Cox, J.G. Laufer, P.C. Beard, et al., Quantitative spectroscopic photoacoustic imaging: a review, J. Biomed. Opt. 17 (6) (2012) 061202.
- [56] C. Cai, K. Deng, C. Ma, et al., End-to-end deep neural network for optical inversion in quantitative photoacoustic imaging, Opt. Lett. 43 (12) (2018) 2752–2755.
- [57] C. Yang, H. Lan, H. Zhong, et al., Quantitative photoacoustic blood oxygenation imaging using deep residual and recurrent neural network[C]/2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE (2019) 741–744.
- [58] N. Sobahi, A. Sengur, R.S. Tan, et al., Attention-based 3D CNN with residual connections for efficient ECG-based COVID-19 detection, Comput. Biol. Med. 143 (2022) 105335.
- [59] P. Rajendran, M. Pramanik, Deep learning approach to improve tangential resolution in photoacoustic tomography, Biomed. Opt. Express 11 (12) (2020) 7311–7323.
- [60] Abadi M., Agarwal A., Barham P., et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems[J]. arXiv preprint arXiv:1603.04467, 2016.
- [61] B.E. Treeby, B.T. Cox, k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields, J. Biomed. Opt. 15 (2) (2010), 021314-021314-12.
- [62] M. Schellenberg, K.K. Dreher, N. Holzwarth, et al., Semantic segmentation of multispectral photoacoustic images using deep learning, Photoacoustics 26 (2022) 100341.
- [63] B.E. Treeby, J. Jaros, B.T. Cox, Advanced photoacoustic image reconstruction using the k-Wave toolbox[C]/Photons Plus Ultrasound: Imaging and Sensing 2016, SPIE 9708 (2016) 517–530.
- [64] X. Song, X. Zhou, Photoacoustic microscopy simulation platform based on K-Wave simulation toolbox[C]/Photonics for Quantum 2021, SPIE 11844 (2021) 54–57.
- [65] Y. Gao, W. Xu, Y. Chen, et al., Deep learning-based photoacoustic imaging of vascular network through thick porous media, IEEE Trans. Med. Imaging 41 (8) (2022) 2191–2204.
- [66] M. Schmitt, C.M. Poffo, J.C. de Lima, et al., Application of photoacoustic spectroscopy to characterize thermal diffusivity and porosity of caprocks, Eng. Geol. 220 (2017) 183–195.
- [67] R. Manwar, K. Kratkiewicz, K. Avnaki, Investigation of the effect of the skull in transcranial photoacoustic imaging: a preliminary ex vivo study, Sensors 20 (15) (2020) 4189.
- [68] J.P. Monchalin, L. Bertrand, G. Rousset, et al., Photoacoustic spectroscopy of thick powdered or porous samples at low frequency, J. Appl. Phys. 56 (1) (1984) 190–210.
- [69] C.F. Ramirez-Gutierrez, J.D. Castano-Yepes, M.E. Rodriguez-Garcia, In situ photoacoustic characterization for porous silicon growing: detection principles, J. Appl. Phys. 119 (18) (2016) 185103.
- [70] Zhou J., He D., Shang X., et al. Photoacoustic Microscopy with Sparse Data Enabled by Convolutional Neural Networks for Fast Imaging[J]. arXiv preprint arXiv:2006.04368, 2020.
- [71] S. Jeon, J. Kim, D. Lee, et al., Review on practical photoacoustic microscopy, Photoacoustics 15 (2019) 100141.
- [72] L.V. Wang, J. Yao, A practical guide to photoacoustic tomography in the life sciences, Nat. Methods 13 (8) (2016) 627–638.
- [73] Ledig C., Theis L., Huszár F., et al. Photo-realistic single image super-resolution using a generative adversarial network[C]/Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4681–4690.
- [74] Mathieu M., Couprie C., LeCun Y. Deep multi-scale video prediction beyond mean square error[J]. arXiv preprint arXiv:1511.05440, 2015.

- [75] H. Zhao, Z. Ke, F. Yang, et al., Deep learning enables superior photoacoustic imaging at ultralow laser dosages, *Adv. Sci.* 8 (3) (2021) 2003097.
- [76] H. Zhao, K. Li, N. Chen, et al., Multiscale vascular enhancement filter applied to in vivo morphologic and functional photoacoustic imaging of rat ocular vasculature, *IEEE Photonics J.* 11 (6) (2019) 3900912.
- [77] Q. Yao, Y. Ding, G. Liu, et al., Low-cost photoacoustic imaging systems based on laser diode and light-emitting diode excitation, *J. Innov. Opt. Health Sci.* 10 (04) (2017) 1730003.
- [78] R. Manwar, M. Hosseinzadeh, A. Hariri, et al., Photoacoustic signal enhancement: towards utilization of low energy laser diodes in real-time photoacoustic imaging, *Sensors* 18 (10) (2018) 3498.
- [79] S. Wang, J. Lin, T. Wang, et al., Recent advances in photoacoustic imaging for deep-tissue biomedical applications, *Theranostics* 6 (13) (2016) 2394.
- [80] A. Hariri, M. Hosseinzadeh, S. Noei, et al., Photoacoustic signal enhancement: towards utilization of very low-cost laser diodes in photoacoustic imaging[C]// *Photons Plus Ultrasound: Imaging and Sensing 2017*, SPIE 10064 (2017) 822–826.
- [81] Y. Cao, R. Wang, J. Peng, et al., Humidity enhanced N2O photoacoustic sensor with a 4.53  $\mu\text{m}$  quantum cascade laser and Kalman filter, *Photoacoustics* 24 (2021) 100303.
- [82] M. Alaeian, H.R.B. Orlande, B. Lamien, Kalman filter temperature estimation with a photoacoustic observation model during the hyperthermia treatment of cancer, *Comput. Math. Appl.* 119 (2022) 193–207.
- [83] S.C. Rutan, S.D. Brown, Pulsed photoacoustic spectroscopy and spectral deconvolution with the Kalman filter for determination of metal complexation parameters, *Anal. Chem.* 55 (11) (1983) 1707–1710.
- [84] R. Manwar, M. Zafar, Q. Xu, Signal and image processing in biomedical photoacoustic imaging: a review, *Optics* 2 (1) (2020) 1–24.
- [85] S. Telenkov, A. Mandelis, Signal-to-noise analysis of biomedical photoacoustic measurements in time and frequency domains, *Rev. Sci. Instrum.* 81 (12) (2010) 124901.
- [86] C. Hide, T. Moore, M. Smith, Adaptive Kalman filtering for low-cost INS/GPS, *J. Navig.* 56 (1) (2003) 143–152.
- [87] M. Zhou, H. Xia, H. Lan, et al., Wavelet de-noising method with adaptive threshold selection for photoacoustic tomography[C]//2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE (2018) 4796–4799.
- [88] E.Y. Park, H. Lee, S. Han, et al., Photoacoustic imaging systems based on clinical ultrasound platform, *Exp. Biol. Med.* 247 (7) (2022) 551–560.
- [89] M. Lassen, A. Brusch, D. Balslev-Harder, et al., Phase-sensitive noise suppression in a photoacoustic sensor based on acoustic circular membrane modes, *Appl. Opt.* 54 (13) (2015) D38–D42.
- [90] D. He, J. Zhou, X. Shang, et al., De-noising of photoacoustic microscopy images by attentive generative adversarial network, *IEEE Trans. Med. Imaging* (2022).
- [91] S. Cheng, Y. Zhou, J. Chen, et al., High-resolution photoacoustic microscopy with deep penetration through learning, *Photoacoustics* 25 (2022) 100314.
- [92] Goodfellow I. Nips 2016 tutorial: Generative adversarial networks[J]. arXiv preprint arXiv:1701.00160, 2016.
- [93] Arjovsky M., Bottou L. Towards principled methods for training generative adversarial networks[J]. arXiv preprint arXiv:1701.04862, 2017.
- [94] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein generative adversarial networks [C]//International conference on machine learning, PMLR (2017).
- [95] Z. Zhang, H. Jin, W. Zhang, et al., Adaptive enhancement of acoustic resolution photoacoustic microscopy imaging via deep CNN prior, *Photoacoustics* 30 (2023) 100484.
- [96] J. Zhang, B. Chen, M. Zhou, et al., Photoacoustic image classification and segmentation of breast cancer: a feasibility study, *IEEE Access* 7 (2018) 5457–5466.
- [97] Q. Huang, Z. Miao, J. Li, et al., Classification of breast ultrasound with human-rating BI-RADS scores using mined diagnostic patterns and optimized neuro-network, *Neurocomputing* 417 (2020) 536–542.
- [98] Q. Huang, L. Ye, Multi-task/single-task joint learning of ultrasound BI-RADS features, *IEEE Trans. Ultrason., Ferroelectr., Freq. Control* 69 (2) (2021) 691–701.
- [99] Y. Luo, Q. Huang, L. Liu, Classification of tumor in one single ultrasound image via a novel multi-view learning strategy, *Pattern Recognit.* (2023) 109776.
- [100] Q. Huang, L. Zhao, G. Ren, et al., NAG-Net: nested attention-guided learning for segmentation of carotid lumen-intima interface and media-adventitia interface, *Comput. Biol. Med.* 156 (2023) 106718.
- [101] Q. Huang, H. Luo, C. Yang, et al., Anatomical prior based vertebra modelling for reappearance of human spines, *Neurocomputing* 500 (2022) 750–760.
- [102] P. Wu, Z. Wang, B. Zheng, et al., AGGN: attention-based glioma grading network with multi-scale feature extraction and multi-modal information fusion, *Comput. Biol. Med.* 152 (2023) 106457.
- [103] H. Li, P. Wu, Z. Wang, et al., A generalized framework of feature learning enhanced convolutional neural network for pathology-image-oriented cancer diagnosis, *Comput. Biol. Med.* 151 (2022) 106265.
- [104] N. Awasthi, G. Jain, S.K. Kalva, et al., Deep neural network-based sinogram super-resolution and bandwidth enhancement for limited-data photoacoustic tomography, *IEEE Trans. Ultrason., Ferroelectr., Freq. Control* 67 (12) (2020) 2660–2673.
- [105] Awasthi N., Pardasani R., Kalva S.K., et al. Sinogram super-resolution and denoising convolutional neural network (SRCN) for limited data photoacoustic tomography[J]. arXiv preprint arXiv:2001.06434, 2020.
- [106] F. Zhang, J. Zhang, Y. Shen, et al., Photoacoustic digital brain and deep-learning-assisted image reconstruction, *Photoacoustics* (2023) 100517.
- [107] C. Dehner, I. Olefir, K.B. Chowdhury, et al., Deep-learning-based electrical noise removal enables high spectral optoacoustic contrast in deep tissue, *IEEE Trans. Med. Imaging* 41 (11) (2022) 3182–3193.
- [108] M. Liu, Z. Wang, H. Li, et al., AA-WGAN: attention augmented Wasserstein generative adversarial network with application to fundus retinal vessel segmentation, *Comput. Biol. Med.* 158 (2023) 106874.
- [109] H. Li, N. Zeng, P. Wu, et al., Cov-Net: a computer-aided diagnosis method for recognizing COVID-19 from chest X-ray images via machine vision, *Expert Syst. Appl.* 207 (2022) 118029.
- [110] Kingma D.P., Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.
- [111] S. Gutta, V.S. Kadimesetty, S.K. Kalva, et al., Deep neural network-based bandwidth enhancement of photoacoustic data, *J. Biomed. Opt.* 22 (11) (2017) 116001.
- [112] Y. Zhou, F. Zhong, S. Hu, Temporal and spectral unmixing of photoacoustic signals by deep learning, *Opt. Lett.* 46 (11) (2021) 2690–2693.
- [113] Y. Gao, T. Feng, H. Qiu, et al., 4D spectral-spatial computational photoacoustic dermoscopy, *Photoacoustics* (2023) 100572.
- [114] P.J.G. Lisboa, S. Saralajew, A. Vellido, et al., The coming of age of interpretable and explainable machine learning models, *Neurocomputing* 535 (2023) 25–39.
- [115] Z. Salahuddin, H.C. Woodruff, A. Chatterjee, et al., Transparency of deep neural networks for medical image analysis: a review of interpretability methods, *Comput. Biol. Med.* 140 (2022) 105111.
- [116] Y. Dai, G. Wang, Analyzing tongue images using a conceptual alignment deep autoencoder, *IEEE Access* 6 (2018) 5962–5972.
- [117] P.W. Koh, T. Nguyen, Y.S. Tang, et al., Concept bottleneck models[C]// International conference on machine learning, PMLR (2020) 5338–5348.
- [118] H. Ma, Z. Cheng, Z. Wang, et al., Switchable optical and acoustic resolution photoacoustic dermoscope dedicated into in vivo biopsy-like of human skin, *Appl. Phys. Lett.* 116 (7) (2020).



**Xiang Wei** is a master student from Nanjing University of Science and Technology. He is now in the second year of his master's degree and his current research focuses on applications of photoacoustic imaging in biomedicine.



**Ting Feng** received her bachelor's degree, master's degree and Ph.D. degree from Nanjing University in 2010, 2012 and 2016, respectively. She is currently working at Fudan university in China. She was the visiting scholar at the University of Michigan in 2018 and 2019, and she was the joint-Ph.D. student at the University of Michigan in 2013–2015. Her current research interest includes photoacoustic imaging and measurements. A major part of her research is clinical application of photoacoustic techniques for bone health assessment.



**Qinghua Huang** received the Ph.D. degree in biomedical engineering from the Hong Kong Polytechnic University, Hong Kong, in 2007. Now he is a full professor in School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, China. His research interests include multi-dimensional ultrasonic imaging, medical image analysis, machine learning for medical data, and intelligent computation for various applications.



**Qian Chen** as a leading expert in the National Key Discipline of “Optical Engineering” at Nanjing University of Science and Technology. As the primary contributor, he has won a second-class State Technological Invention Award, a second-class State Scientific and Technological Progress Award and five first-class provincial and ministerial-level science and technology awards. As the first inventor, he has obtained 74 granted invention patents, 16 PCT international patents, and 6 U.S. patents. He has authored three books and 374 SCI papers, among which 27 have been featured on the cover. Currently, he serves as a Fellow and Executive Director of the Chinese Society of Optical Engineering and Executive Director of the Chinese Institute of Electronics.



**Haigang Ma** received the Ph.D. degree in Optics from the South China Normal University, China, in 2020. Now he is an associate professor in School of Electronic and Optical Engineering, Nanjing University of Science and Technology, China. His research interests include photoacoustic imaging, ultrasonic imaging, photoelectric detection and processing, and photoacoustic imaging for various biomedical applications.



**Chao Zuo** is a professor in optical engineering, Nanjing University of Science and Technology (NJUST), China. He leads the Smart Computational Imaging Laboratory (SCILab: [www.scilaboratory.com](http://www.scilaboratory.com)) at the School of Electronic and Optical Engineering, NJUST. He has long been engaged in the development of novel Computational Optical Imaging and Measurement technologies, with a focus on Phase Measuring Imaging Metrology such as Holographic Interferometric Microscopy, Non-interferometric Quantitative Phase Imaging (QPI), Fringe Projection Profilometry (FPP), and Structured Illumination Microscopy (SIM). He has authored > 200 peer-reviewed publications in prestigious journals with over 11,000 citations.



## 深度学习在超分辨显微成像中的研究进展(特邀)

鲁心怡<sup>1,2</sup>, 黄昱<sup>3</sup>, 张梓童<sup>4</sup>, 吴天筱<sup>1,2</sup>, 吴洪军<sup>1,2</sup>, 刘永焘<sup>1,2\*</sup>, 方中<sup>3\*\*</sup>, 左超<sup>1,2\*\*\*</sup>, 陈钱<sup>1,2</sup><sup>1</sup>南京理工大学电子工程与光电技术学院智能计算成像实验室, 江苏 南京 210094;<sup>2</sup>南京理工大学江苏省光谱成像与智能感知重点实验室, 江苏 南京 210094;<sup>3</sup>南京理工大学机械工程学院, 江苏 南京 210094;<sup>4</sup>深圳萨米医疗中心(深圳市第四人民医院)感染管理科, 广东 深圳 518118

**摘要** 超分辨显微成像技术打破了传统显微镜存在的衍射极限限制, 提供了前所未有的细节观察能力, 使人们得以观察到衍射极限以下的微观世界, 有力地推动了生物医学、细胞学、神经科学等领域的发展。然而, 现有的超分辨显微成像技术存在成像速度慢、重建图像含有伪影、对生物样品光损伤大、轴向分辨率低等缺陷。近年来, 得益于人工智能技术的快速发展, 深度学习被用于研究克服超分辨显微技术的各种缺陷, 突破了超分辨显微成像技术的发展瓶颈。聚焦于主流超分辨显微成像技术存在的缺陷, 总结了深度学习对超分辨显微技术的优化效果, 并根据超分辨显微成像技术原理的特异性, 介绍了不同网络在超分辨显微技术上的应用成效, 最后对深度学习在超分辨显微成像领域应用中存在的问题进行了分析, 并对其发展进行了展望。

**关键词** 深度学习; 图像重建; 显微成像; 超分辨

中图分类号 O436

文献标志码 A

DOI: 10.3788/LOP241455

## Advances in Deep Learning for Super-Resolution Microscopy (Invited)

Lu Xinyi<sup>1,2</sup>, Huang Yu<sup>3</sup>, Zhang Zitong<sup>4</sup>, Wu Tianxiao<sup>1,2</sup>, Wu Hongjun<sup>1,2</sup>, Liu Yongtao<sup>1,2\*</sup>,  
Fang Zhong<sup>3\*\*</sup>, Zuo Chao<sup>1,2\*\*\*</sup>, Chen Qian<sup>1,2</sup><sup>1</sup>Smart Computational Imaging Laboratory, College of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, Jiangsu, China;<sup>2</sup>Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing University of Science and Technology, Nanjing 210094, Jiangsu, China;<sup>3</sup>School of Mechanical Engineering, Nanjing University of Science and Technology, Nanjing 210094, Jiangsu, China;<sup>4</sup>Infection Management Department of Shenzhen Sami Medical Center (Shenzhen Fourth People's Hospital), Shenzhen 518118, Guangdong, China

**Abstract** Super-resolution microscopy imaging technology surpasses the diffraction limit of traditional microscopes, thereby offering unprecedented detail and allowing observation of the microscopic world below this limit. This advancement remarkably promotes developments in various fields such as biomedical, cytology, and neuroscience. However, existing super-resolution microscopy techniques have certain drawbacks, such as slow imaging speed, artifacts in reconstructed images, considerable light damage to biological samples, and low axial resolution. Recently, with advancements in artificial intelligence, deep learning has been applied to address these issues, overcoming the limitations of super-resolution microscopy imaging technology. This study examines the shortcomings of mainstream super-resolution microscopy imaging technology, summarizes how deep learning optimizes this technology, and evaluates the effectiveness of various networks based on the principles of super-resolution microscopy. Moreover, it analyzes the challenges of applying deep learning to this technology and explores future development prospects.

**Key words** deep learning; image reconstruction; microscopic imaging; super-resolution

收稿日期: 2024-06-07; 修回日期: 2024-06-27; 录用日期: 2024-07-02; 网络首发日期: 2024-07-05

基金项目: 国家自然科学基金(62275125, 62201267, 62275121, 12204239, 62175109)、国家重大科研仪器研制项目(62227818)、江苏省自然科学基金青年项目(BK20220946)、江苏省基础研究计划前沿引领专项(BK20192003)、江苏省科技计划重点国别产业技术研发合作项目(BZ2022039)、中央高校基本科研业务费专项资金(30922010313, 2023102001)、江苏省光谱成像与智能感知重点实验室开放基金(JSGP202201)、深圳 Sami 医疗中心高级别临床研究启航团队(SSMC-2024-TB5)、南京理工大学青年人才培养专项(30922010313)

通信作者: \*Yongtao.Liu@njust.edu.cn; \*\*fangzhong@njust.edu.cn; \*\*\*zuocho@njust.edu.cn

1611002-1

# 1 引言

在人类探索生命奥秘的漫长过程中,显微镜发挥着不可忽视的作用。自 16 世纪显微镜诞生起,人类对生命活动的探索踏入了微观领域。1873 年,阿贝提出了阿贝衍射极限,指出传统光学显微镜分辨率不可超过入射波长的一半,因此对于生物成像而言,基于可见光的传统显微镜分辨率局限于 200 nm,难以分辨如线粒体、肌动蛋白、活细胞微管等微小结构。为了进一步探索微观世界,突破传统光学显微镜衍射极限的超分辨显微成像技术应运而生<sup>[1]</sup>。超分辨显微成像技术的出现,成功将光学显微镜的观测分辨率由微米级带入了纳米级,分辨率提升至 20~70 nm。如今,超分辨显微成像技术已经成为了人类探索和发现微观世界的重要手段<sup>[2]</sup>。

随着超分辨显微成像技术的不断发展,诞生了诸多超分辨方法,例如受激发射损耗荧光显微镜(STED)<sup>[3]</sup>、随机光学重构显微镜(STORM)<sup>[3]</sup>、光活化定位显微镜(PALM)<sup>[4]</sup>、结构光照明显微镜(SIM)<sup>[4]</sup>、多光子非线性超分辨率成像(MPUM)<sup>[5-8]</sup>等。每种超分辨显微技术利用了不同的手段突破了衍射极限,根据其成像原理,不同的超分辨技术存在着不同先天优势和固有缺陷<sup>[9]</sup>。在生物样品的显微成像过程中,空间分辨率决定了是否能区分两个精细结构,时间分辨率决定了显微镜是否能完整捕捉生命活动过程,成像深度决定了对生物深组织成像的质量,光漂白、光毒性决定了成像对生物样品所带来的损伤程度。因此对于不同的超分辨成像技术,应根据其特性以及生物成像

的需求进行针对性优化。

深度学习是机器学习的一个分支,其通过多层人工神经网络对数据进行计算处理,根据是否需要已标注数据进行训练进行区分,网络可以分为有监督学习和无监督学习两种学习方式。近年来,深度学习快速发展,产生了诸多神经网络框架,如:卷积神经网络(CNN)<sup>[10]</sup>、生成对抗网络(GAN)<sup>[11]</sup>、U-Net<sup>[12]</sup>、ResNet<sup>[13]</sup>、Faster R-CNN<sup>[14]</sup>网络等。神经网络具有极强的自适应性,可以通过反向传播算法对自身不断优化,实现更好的数据拟合,同时其具备端到端的学习能力,实现输入端到输出端的直接映射。基于深度学习的优秀特性,使其在图像超分辨、图像去噪、图像分割等方面表现出色,这吸引了超分辨显微成像领域研究者的注意,他们将深度学习与超分辨显微技术进行结合,进一步提升成像分辨率,并克服了传统方法的超分辨显微技术缺陷。

本文讨论了深度学习在超分辨显微成像领域中的研究进展,以主流超分辨成像技术为脉络,介绍了不同超分辨显微成像技术的成像原理,分析其存在的优点以及缺陷,利用深度学习的方法应对不同技术所存在的缺陷。对于 STED,着重降低其光毒性,提升其成像速度以及轴向分辨率;对于 STORM 和 PALM,着重提升其成像速度,以及分子定位精度,提高重建分辨率;对于 SIM,着重提升其成像质量,去除重建图像伪影,并进一步减少图像重建所需帧数,降低光损伤。图 1<sup>[15-27]</sup>总结了不同神经网络对超分辨成像技术的优化方法与优化效果。

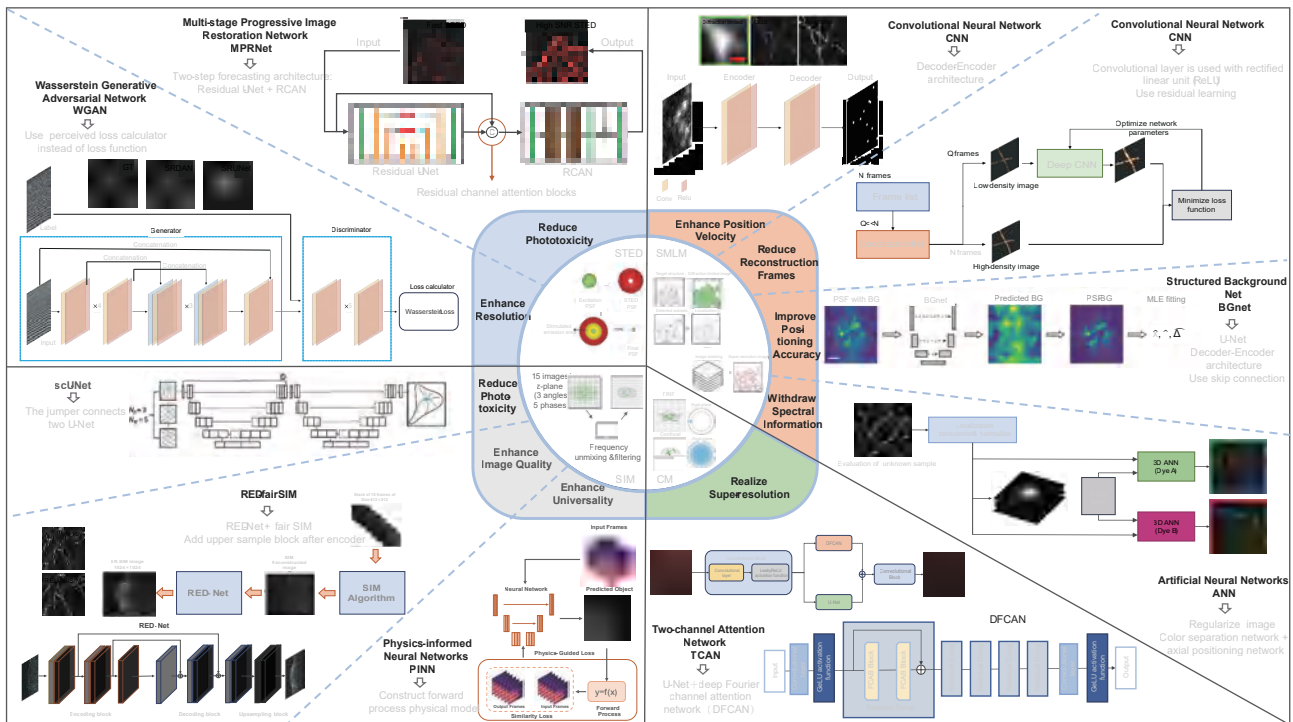


图 1 深度学习在超分辨率成像领域应用框架图<sup>[15-27]</sup>

Fig. 1 Framework of deep learning in super-resolution imaging<sup>[15-27]</sup>

## 2 深度学习在受激发射损耗显微术中的应用

STED 是目前主流的超分辨技术之一,如图 2<sup>[15,16,28]</sup>所示,STED 是先一束高斯激发光经由物

镜照射到样品上,同时引入一束环形光束通过受激辐射将外围区域的荧光分子淬灭,两束激光精准对齐进而产生环形光中心未被淬灭的发射轮廓<sup>[29]</sup>。STED 凭借其空间分辨率高、实时成像、无需后期图像重建的优点得到科学家们的广泛关注。

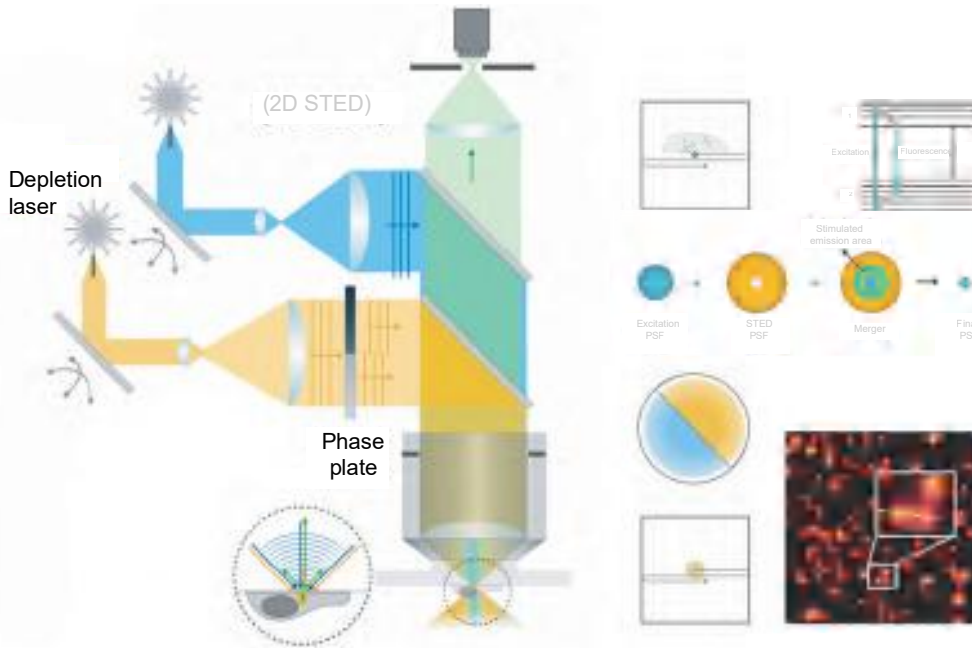


图 2 STED 光路原理图<sup>[15,16,28]</sup>

Fig. 2 Schematic diagram of STED<sup>[15,16,28]</sup>

STED 实现超分辨的关键在于损耗光的功率以及受激辐射与自发荧光相互竞争中的非线性效应,淬灭光功率越强,激发光斑的光斑外围受到的抑制越强,产生的有效荧光光斑越小,空间分辨率越高,但使用强损耗光的同时会带来光漂白、光毒性、光损伤等问题,这限制了 STED 活细胞中的应用<sup>[30-32]</sup>。此外,STED 使用点扫描成像,因此其成像速度慢,时间分辨率较低。同时,在应对厚样品时,其轴向分辨率仍有待进一步提升。

理论上,减少对生物样品的曝光时间即可降低光损伤<sup>[33]</sup>,但较短的像素驻留时间会导致较差的信噪比(SNR),进而降低图像分辨率<sup>[34]</sup>。为实现低 SNR 下的高质量超分辨率成像,美国佛罗里达大学的 Ebrahimi 等<sup>[17]</sup>在 2023 年提出使用多阶段渐进图像恢复(MPRNet)的方法,实现了 STED 的像素停留时间减小 1~2 个数量级,进而减少对样品的光漂白与光损伤。MPRNet 基于 U-Net 和残差通道注意力网络(RCAN)架构,使用 U-Net 网络架构进行上下采样,绕过低频信号,获取特征图像,然后使用 RCAN 网络进行图像重建,其网络框架图如图 3(a)所示,该方法可以精准重建低曝光图像,获得高信噪比超分辨图像,实现在保证原有分辨率的前提下将 STED 的像素停留时间减少至原本的 3.125%,成像速度极大提升。传统

STED 在 1.0 mm×0.78 mm 大小的区域记录 744 张大小为 2048 pixel×2048 pixel 的图像需要 12 h,而使用 MPRNet 仅需要 21 min。图 3(b)为 MPRNet 对噪声 STED 图像的恢复效果图。

2023 年 Chen 等<sup>[35]</sup>将寿命调谐分离(SPLIT)技术与 STED 技术相结合,使用时间分辨采集和相量分析成功区分有效荧光区域中心和外围发射的光子,实现了在不增加光损伤的同时提高分辨率。在此基础上,他们进一步将 SPLIT-STED<sup>[11]</sup>与 GAN 网络框架相结合,形成了基于 GAN 估计的荧光寿命成像网络(flimGANE),利用 GAN 网络对光子匮乏的相量图形去噪,以提供更高质量的相量图像,提升了 SPLIT 的重建水平,进而提升了系统的分辨率和鲁棒性,较单一 SPLIT-STED 实现了 1.45 倍的分辨率增强。图 3(c)、(d)为 flimGANE 网络框架图以及 Chen 等在实验中获得的两束耗尽激光强度下共聚焦、pSTED、pSTED、SPLIT 和 STED flimGANE 图像的比较。

理论上,STED 可以达到的最佳横向分辨率为 20 nm,但由于光毒性的限制,在实际活细胞的应用中横向分辨率往往只能达到 100 nm。为进一步提升横向分辨率,2020 年伦敦理工大学的 Li<sup>[18]</sup>提出深度对抗网络(DAN-based),根据 STED 原理,对较低分辨率图像使用物理建模进行计算,并输出对应的高分辨率



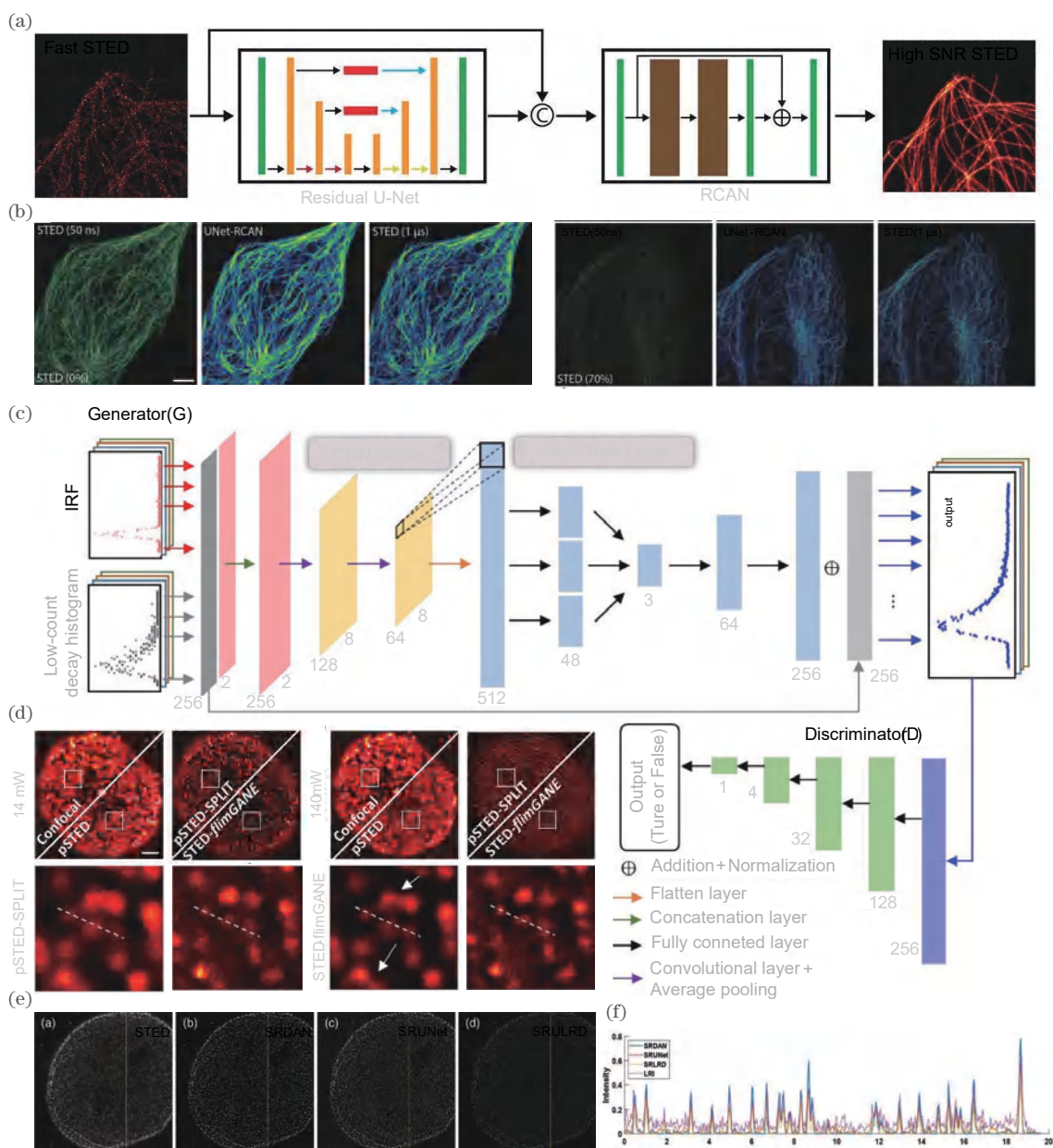


图 3 不同网络框架图和实验结果。(a) MPRNet 网络架构<sup>[17]</sup>；(b) U2OS 细胞中  $\beta$ -微管蛋白 (STAR635P) 的 MPRNet 重建图像<sup>[17]</sup>；(c) STED-flimGANE 网络结构图<sup>[35]</sup>；(d) 功率耗损极端条件下的强度图像<sup>[35]</sup>；(e) 低噪声 SRDAN 与其他方法的核孔成像对比图<sup>[18]</sup>；(f) 各算法核孔图像的线强度分布<sup>[18]</sup>

Fig. 3 Different network framework diagrams and experimental results. (a) MPRNet network architecture<sup>[17]</sup>；(b) MPRNet reconstruction images of  $\beta$ -tubulin (STAR635P) in U2OS cells<sup>[17]</sup>；(c) STED-flimGANE network structure diagram<sup>[35]</sup>；(d) intensity images under extreme conditions of rate depletion<sup>[35]</sup>；(e) comparison of nuclear hole imaging between low noise SRDAN and other methods<sup>[18]</sup>；(f) line intensity distribution of core hole images in each algorithm<sup>[18]</sup>

图像,并以此对网络进行训练,极大提高了网络模型的重建分辨率。同时,使用感知损失计算器替代了传统的均方误差损失函数,增强了网络对图像细节以及图像结构的提取能力,该网络模型可实现将 60 nm 的 STED 图像分辨率优化至 30 nm,图 3(e)、(f)为其成像效果图。

STED 成像作为一种点扫描成像方式,其实现三维 (3D) 成像需要使用 3D 扫描完成, Bessel-Bessel

STED (BB-STED) 实现了 STED 的 3D 扫描,但其图像仅是二维 (2D) 投影的结果,并没有实现轴向超分辨。目前,提升轴向分辨率、获取轴向信息的常用手段是使用螺旋点扩散函数,如单螺旋点扩散函数 (SH-PSF)<sup>[36]</sup> 和双螺旋点扩散函数 (DH-PSF)<sup>[37]</sup>,螺旋点扩散函数可将不同深度的荧光点扩展到不同的方位角,进而获取荧光点的轴向位置信息,提升轴向分辨率,但该方法无法准确分辨轴向密集荧光点。为



此, Ji 等<sup>[38]</sup>在 2024 年基于 BB-STED, 使用单螺旋点扩散函数并结合深度学习算法实现轴向超分辨。单螺旋点扩散函数相较于双螺旋点扩散函数具有更高的信噪比和更大的有效深度, 使用模拟密集单螺旋点图像对基于 CNN 的解码器-编码器网络进行训练, 将原始图像输入训练模型, 精确定位荧光点相对于光斑中心的方位角。在保证横向分辨率 50 nm 的同时, 该方法的轴向分辨率可达到 50 nm, 高于 4Pi-STED。同时, 在荧光点密度  $n=20$  的情况下, 该方法的轴向分辨率可达到 63 nm。

### 3 深度学习在单分子定位显微镜中的应用

单分子定位显微镜(SMLM)技术通过在多次循环中重复随机地激发稀疏分布的荧光分子, 并结合相关定位算法, 实现对不重叠荧光分子的精确定位, 最后

将定位图像集进行整合重建, 生成超分辨图像<sup>[39]</sup>, SMLM 成像原理如图 4<sup>[19]</sup>所示。SMLM 主要分为两种, 一种是 2006 年由哈佛大学庄小威团队首次提出的使用荧光小分子的随机光学重构显微镜(STORM)<sup>[3]</sup>; 另一种是同年由 Betzig 团队提出的使用荧光蛋白的光活化定位显微镜(PALM)<sup>[40]</sup>。由 SMLM 原理成像可知, SMLM 图像重建, 需要对样品进行大量的图像采集, 因此传统 SMLM 存在着时间分辨率低、无法实现活体动态成像、对样品存在光毒性和光漂白的问题。此外, 由于需要对采集的图像集进行分子定位和重建处理, 传统 SMLM 成像质量受制于分子定位算法精度, 成像时间受制于分子定位算法速度。而对于多色单分子定位成像, 则需要消除其在图像采集时的光谱串扰以及采集速度慢的问题。综上所述, 对于 SMLM 的优化主要集中在降低定位所需帧数、提高定位精度及速度、提升多色成像颜色区分水平上<sup>[41]</sup>。

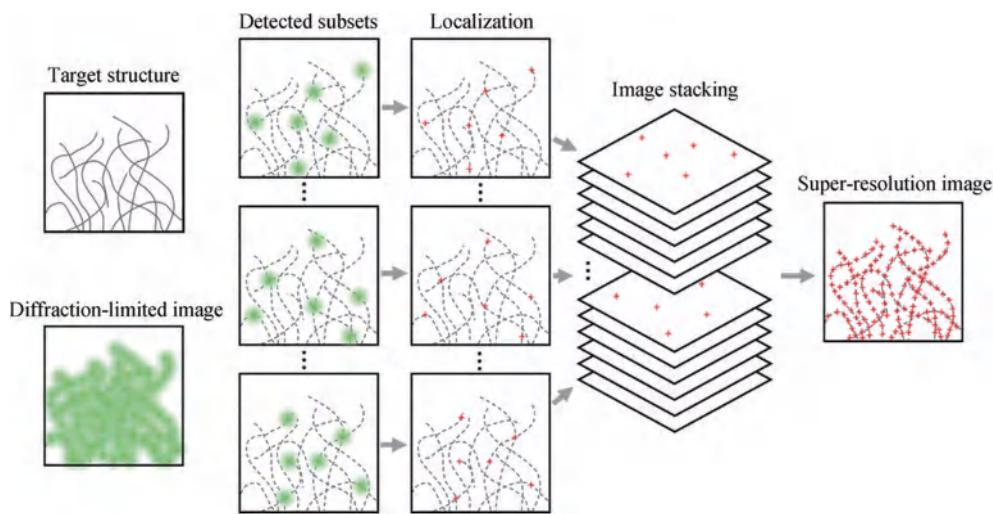


图 4 单分子定位显微成像原理<sup>[19]</sup>

Fig. 4 Principle of single molecule localization microscopy<sup>[19]</sup>

#### 3.1 提高重建速度

在单分子定位显微镜中, 单个荧光点的位置通常由高斯函数对 PSF 拟合来获取, 对于荧光点重叠密度高的区域, 其拟合过程更为复杂, 导致 SMLM 图像重建过程产生大量数据。因此, 传统 SMLM 计算成本极高, 成像速度缓慢, 且拟合重建精度依赖于对拟合参数调试, 专业性要求高。为了降低重建图像所需的计算成本, 提高成像速度, Nehme 等<sup>[20]</sup>在 2018 年提出了无参数超分辨率图像重建方法 Deep-STORM, 其网络基于传统的编码器(encoder)-解码器(decoder)架构和 CNN, 通过简单网络结构实现快速图像重建与无参数化。在荧光点与微管的重建实验中, Deep-STORM 已被证明比传统算法更快、更准确, Deep-STORM 对微管的重建效果如图 5(a)所示。但 Deep-STORM 算法高度依赖于训练数据, 因此在成像参数存在实质性差异时会产生伪影, 为此 Sahel 和 Eldar<sup>[42]</sup>提出了 Self-

STORM, 该网络使用自监督学习方案代替监督学习, 不再仅由原始图像映射到重建图像训练, 而是利用单个图像中信息的内部递归来组成编码器, 该编码器仅从低分辨率图片中学习, 不需要频繁调整优化参数, 降低了对外部训练样本的需求。该模型可以训练出与 Deep-STORM 成像质量相当的图像, 且泛用性更强, 其网络框架以及对比重建效果如图 5(b)、(c)所示。2021 年 Li 等<sup>[43]</sup>受 Deep-STORM 启发, 结合递归神经网络(RNN)提出了 Deep Recurrent-Supervised Network-STORM(DRSN-STORM), RNN 能够捕捉图像时序信息, 可以在不增加网络深度的情况下提取时间数据中的额外特征, 生成高效率模型。与经典的 Deep-STORM 方法相比, DRSN-STORM 的运行时间可以至少节省 40%, 其网络框架图和微管实验结果如图 5(d)、(e)所示。

为实现快速单分子 3D 定位, Boyd 等<sup>[44]</sup>于 2018 年

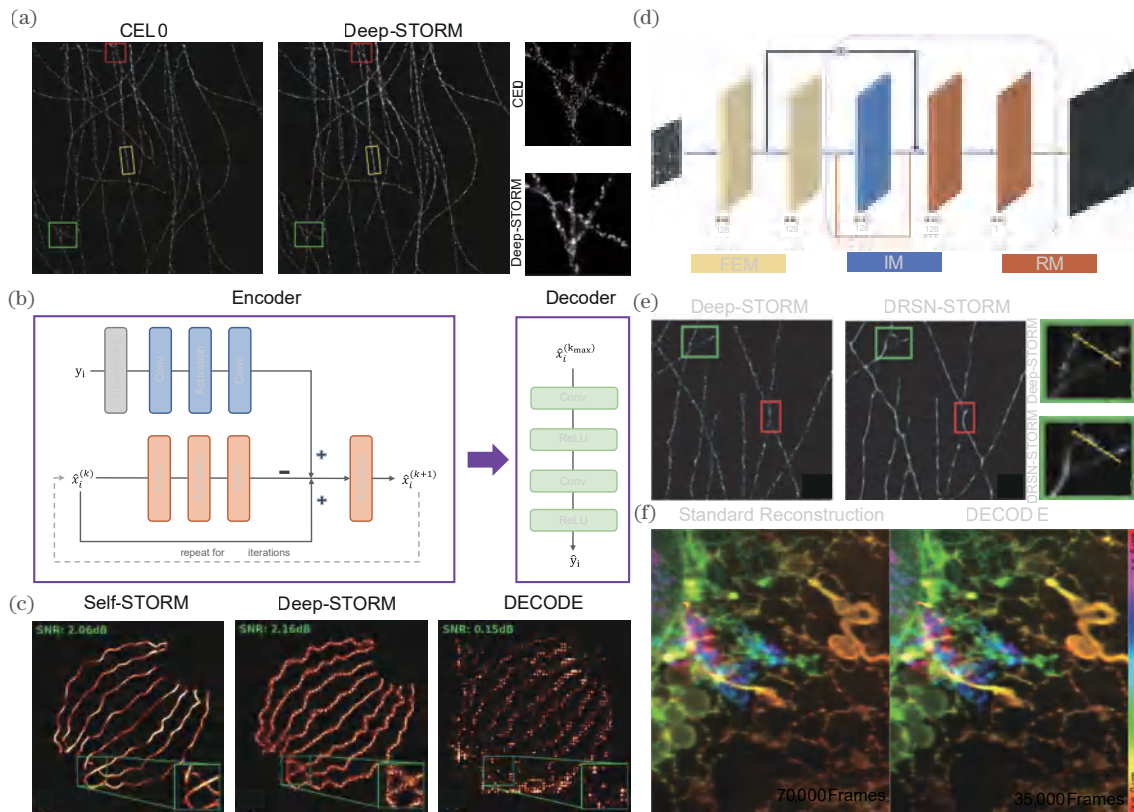


图5 提高图像重建速度相关网络与重建结果。(a) Deep-STORM 重建效果对比图<sup>[20]</sup>；(b) Self-STORM 网络架构<sup>[42]</sup>；(c) Self-STORM 重建效果对比图<sup>[42]</sup>；(d) DRSN-STORM 网络架构, 特征提取模块(FEM)、推理模块(IM)、重建模块(RM)<sup>[43]</sup>；(e) DRSN-STORM 微管图像重建效果<sup>[43]</sup>；(f) DECODE 重建效果图<sup>[46]</sup>

Fig. 5 Network and reconstruction results related to improving image reconstruction speed. (a) Deep-STORM reconstruction effect comparison<sup>[20]</sup>; (b) self-STORM network architecture<sup>[42]</sup>; (c) self-STORM rebuild effect comparison diagram<sup>[42]</sup>; (d) DRSN-STORM Network architecture, feature extracting module (FEM), inference module (IM), and reconstruction module (RM)<sup>[43]</sup>; (e) DRSN-STORM microtubule image reconstruction effect<sup>[43]</sup>; (f) DECODE reconstruct effect image<sup>[46]</sup>

提出了 DeepLOCO 网络架构。该网络基于经典 CNN 网络, 使用残差连接防止产生梯度问题, 并利用含不同种类噪声的模拟 PSF 训练网络, 创新性地以最小化贝叶斯风险作为优化目标。该设计使得 DeepLOCO 具备了快速定位分子位置的能力, 较传统算法实现了几个数量级的提升, 且适用于不同噪声的图像。使用长序列微管蛋白数据集的进行重建测试, DeepLOCO 在保证定位精度的前提下, 在 1 s 内分析了 20000 帧的 3D SMLM 数据。

传统算法使用的最大似然估计 (MLE) 在重建过程中需要不断地调整参数以确保定位的精准度, 在分子高密度标记的情况下极大地增加了重建难度。针对该问题, Zelger 等<sup>[45]</sup> 基于 VGG16 架构, 对单帧图像实现了 3D 分子定位。该网络实现了与 MLE 相当的定位精度, 定位速度可达  $22000 \text{ s}^{-1}$ , 相较于传统的 MLE 算法速度提升了 3 个数量级以上。此外, Speiser 等<sup>[46]</sup> 于 2021 年利用深度学习对 MLE 方法进行了优化, 提出了一种基于 U-Net 网络的 Deep Context Dependent (DECODE) 架构, 通过 U-Net 网络将每一帧的荧光分子特征与前后帧特征相融合, 获得更强的网络表达能

力, 加快了网络重建图像的速度, 减少了高密度分子标记下成像所需时间。Speiser 等使用 DECODE 实现了超高标记密度的微管成像。在检测精度和定位误差方面, DECODE 在 12 个数据集上的表现均优于其他常用模型, 其在 LLS-PAINT 图像上的重建效果如图 5(f) 所示。

### 3.2 少帧重建

单分子定位显微镜重建对图像帧的稀疏性要求使得在分子高密度情况下, 需要采集大量帧进行图像重建, 为缩短图像采集时间, 提升成像速度, 应尽可能减少重建所需图像帧数。为此 2018 年 Ouyang 等<sup>[47]</sup> 基于 pix2pix<sup>[48]</sup> 网络开发出了 ANNA-PALM, 相较于传统算法, ANNA-PALM 利用生物图像的结构冗余, 从采样帧数不足的 SMLM 数据中重建高质量图像。ANNA-PALM 网络使用了三层损失函数设计: 第一层使用多尺度结构相似指数 (MS-SSIM) 评价网络输出图像与密集 PALM 图像重建之间的差异; 第二层引入 CNN 预测网络, 生成模型输出的超分辨率图像对应的低分辨率宽场图像, 并与实际宽场图像进行对比, 评价二者的一致性; 第三层使用 cGAN 鉴别器比较稀疏 PLAM、



宽场图像与密集 PLAM 和网络输出图像之间的差异。该设计使得网络训练只需要少量的密集 PLAM 图像,且不会过度拟合。该网络模型实现了短时间采集情况下的高质量超分辨率图像重建,并且可以预测出重建可能出错的位置。在免疫染色微管成像实验中,

ANNA-PALM 使用宽场图像与 9 s 采集的 300 帧图像相结合作为输入,其重建图像达到了与使用 10 min 采集的 60000 帧图像重建的 PALM 相同水平,ANNA-PALM 实现了帧数减少两个数量级情况下的高质量图像重建,其架构与重建结果如图 6(a)、(b)所示。

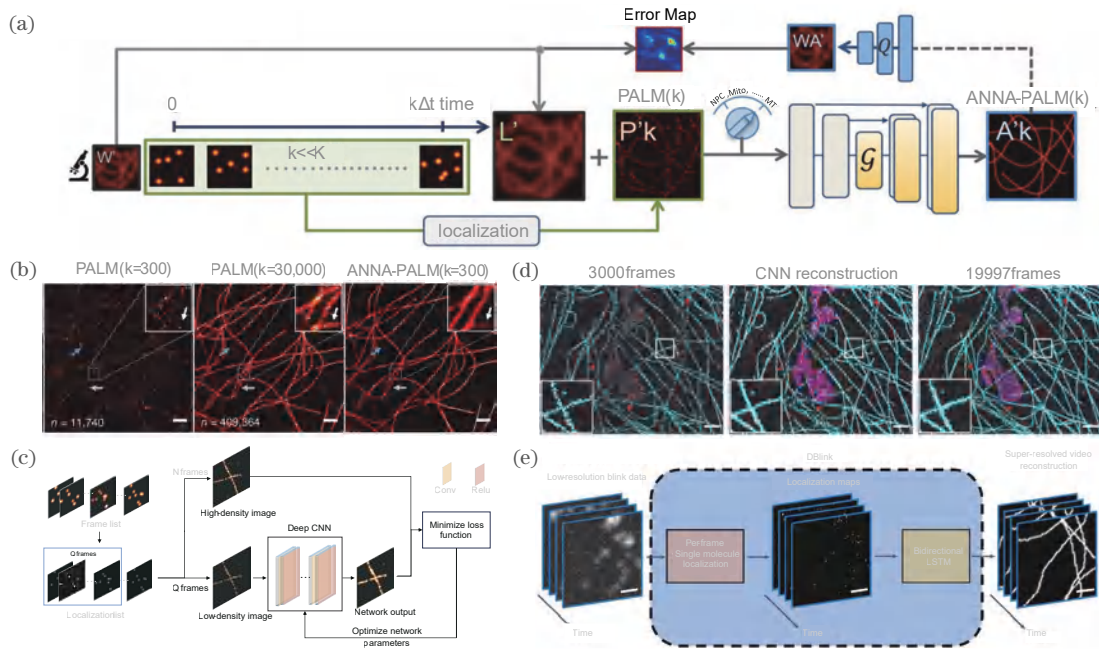


图 6 减少图像重建帧数相关的网络架构以及重建结果。(a) ANNA-PAM 网络架构<sup>[47]</sup>; (b) ANNA-PAM 重建与 PALM 对比<sup>[47]</sup>; (c) 多色单分子图像重建 CNN 框架<sup>[21]</sup>; (d) 多色单分子网络重建效果<sup>[21]</sup>; (e) DBlink 网络架构<sup>[49]</sup>

Fig. 6 Reduce the network architecture and reconstruction results related to the frame rate of image reconstruction. (a) ANNA-PAM network architecture<sup>[47]</sup>; (b) ANNA-PAM reconstruction compared with PALM<sup>[47]</sup>; (c) multi-color single molecule image reconstruction CNN framework<sup>[21]</sup>; (d) multicolor single molecule network reconstruction effect<sup>[21]</sup>; (e) DBlink network architecture<sup>[49]</sup>

光谱单分子定位显微镜 (sSMLM) 可以同时提供单分子的位置信息和光谱信息,为单个样品提供纳米级的多色超分辨成像,然而这需要大于  $10^4$  的连续衍射帧才能实现高分辨率图像重建,长时间的图像采集,不仅会影响活细胞成像,同时会带来光漂白降低图像质量。针对该问题, Gaire 等<sup>[21]</sup> 基于 CNN 网络,首先使用分子定位算法对图像分子进行定位,再利用网络模型恢复光谱分类后的稀疏图像,生成高分辨率图像,最后再将不同光谱的图像合成,生成彩色高分辨率图像。在 COS-7 细胞双色成像实验中,该方法使用 3000 帧和 23300 个定位点进行图像重建,实现了与传统方法使用 19997 帧和 134900 个定位点相同的成像质量,极大减少了重建所需帧数,其架构与重建结果如图 6(c)、(d)所示。

单分子定位显微镜对重建帧数的高需求使其时间分辨率较低,极大地限制了其在样品动态成像中的应用。2023 年 Saguy 等<sup>[49]</sup> 将双向卷积神经网络 (CNN) 与长短期存储器 (CNN-LSTM) 相结合,提出了一种 Data Base Link (DBlink) 网络。该网络将超分辨定位

图的连续帧作为输入,通过捕捉不同输入帧之间的长期相关性,输出动态超分辨结构视频,实现了超时空分辨率的视频重建。在活细胞时空分辨率重建上,实现了微管和内质网 30 nm 空间分辨率和 15 ms 时间分辨率的超分辨率重建,其网络结构如图 6(e)所示。

### 3.3 提升定位精度

在单分子定位显微镜中,单分子的定位精度决定了图像的空间分辨率,而定位精度取决于拟合算法的优越性,因此使用深度学习优化拟合算法,提高单分子的定位精度,进而实现 SMLM 的空间分辨率提升。实现单分子的精确定位需要消除背景噪声的影响,对背景信号参数进行精确估计。然而,由于局部背景信号的变化,背景信号参数的估计过程往往较为困难<sup>[12]</sup>。为此, Möckl 等<sup>[22]</sup> 基于 U-Net 架构提出了 Background Net (BGNet) 网络。利用 U-Net 网络强大的图像分割能力,剥离原始图像背景,同时模拟生成不同轴向位置的不同形状 PSF 训练网络,提升背景分割的精准度。该网络模型对于标准开孔径 (OA) 点扩散函数、 $2 \mu\text{m}$  轴向范围的双螺旋 (DH) 点扩散函数和  $6 \mu\text{m}$  轴向范围的

四足点扩散函数 (Tetra6 PSF) 均可实现精准的背景去除,显著提升了分子定位精度。在微管成像实验中,用 BGNet 估计的结构化背景与恒定的背景相比,克服了严重伪影、虚假定位、细节丢失等缺陷,重建了高质量微管超分辨图像,其网络架构与去背景效果如图 7(a)、(b) 所示。此外, Cascarano 等<sup>[50]</sup> 基于 DeepSTORM 和正则化反卷积 CEL0, 提出了 DeepCEL0 网络, 该网络将 CNN 网络作为骨架, 并使用 CEL0 作为训练损失函数。该设计使得 DeepCEL0 网络成功集合了 CEL0 高精度分子定位和 DeepSTORM 重建速度快, 无需参数计算的优势。与标准方法相比, DeepCEL0 可以在不影响计算成本的情况下提供高精度的定位图像, DeepCEL0 对 IEEE ISBI 微管数据集的重建结果如图 7(c) 所示。对于密集标记样品中单个荧光点的精确 3D 定位, Nehme 等<sup>[51]</sup> 提出了

DeepSTORM3D, 利用 CNN, 对高密度荧光点进行精确定位, CNN 在密度情况下的定位精度远远优于基于拟合删减的匹配追踪法, 同时, Nehme 等联合优化 PSF 和定位网络, 引入了一个可微的物理模拟层, 该模拟层利用相位掩模调控显微镜 PSF, 并将其 3D 荧光图像编码为对应的低分辨率 2D 图像, 再将该图像传输给 CNN 网络, 使用反向传播算法不断优化相位掩模和 CNN 参数, 其流程图如图 7(d) 所示。该方法所学习的产生 PSF 在低密度情况下与四足 PSF 效果相当, 但当荧光点密度大于  $0.2 \mu\text{m}^{-2}$  时, 学习 PSF 重建效果优于四足 PSF, 荧光点密度为  $0.197 \mu\text{m}^{-2}$  时的四足 PSF 和学习 PSF 的定位结果如图 7(e) 所示。对 U2OS 细胞核内端粒进行成像实验, 在  $20 \mu\text{m}^2$  的细胞核内含有数十个端粒, 在轴向范围  $3 \mu\text{m}$  内, 四足 PSF 的 CNN 网络仅能恢复 62 个端粒中的 49 个, 而使用学习 PSF 的

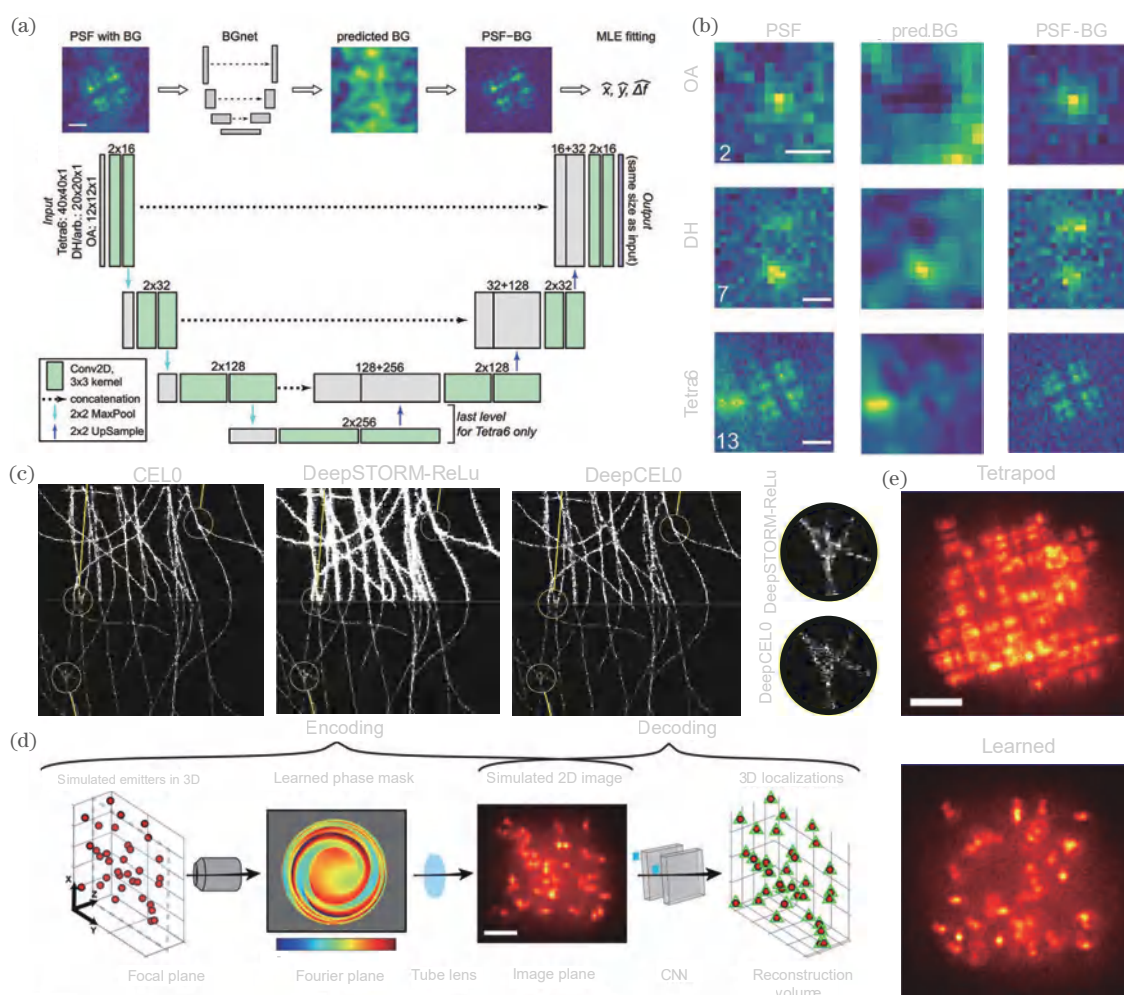


图 7 高精度分子定位相关网络与重建结果。(a) BGNet 网络架构<sup>[22]</sup>; (b) BGNet 对三种 PSF 成像方式的背景估计<sup>[22]</sup>; (c) 各算法在 IEEE ISBI 微管数据集的重建结果对比, 上半部分为二值化图像, 下半部分为标准化图像<sup>[50]</sup>; (d) 物理模拟反馈流程<sup>[51]</sup>; (e) 四足 PSF 和学习 PSF 的定位结果<sup>[51]</sup>

Fig. 7 High precision molecular localization related networks and reconstruction results. (a) BGNet network architecture<sup>[22]</sup>; (b) background estimation of three PSF imaging methods using BGNet<sup>[22]</sup>; (c) reconstruction results of IEEE ISBI microtubule data set are compared, the top half is binary image, and the bottom half is standardized image<sup>[50]</sup>; (d) physical simulation feedback flow<sup>[51]</sup>; (e) tetrapod and learned PSF localization results<sup>[51]</sup>



CNN 网络能够恢复 57 个, 仅有两个呈假阳性。

### 3.4 提取光谱信息

点扩散函数中包含了丰富的信息, 如: 分子位置、荧光发射波长等。同时, 分析单分子发光模式对于探索分子标记目标的结构和生理信息, 以及进一步清楚地了解它们的相互作用和细胞环境至关重要。近年来, 人们通过深度学习, 实现了从衍射受限图像中提取 PSF 的隐藏信息。2018 年, Zhang 等<sup>[52]</sup>提出了一种用于多路单分子分析的深度神经网络 Single-Molecule Net (smNet), 该网络由卷积层、残差块和全连接层组成, 由于光子在小区域内的分布体现了 PSF 的额外特征, 为准确提取特征信息, Zhang 等在初始层

内使用较大内核, 堆叠多层卷积层和残差块, 尽可能准确地捕获特征信息。同时, 设定单分子测量误差与每个训练图像的 Cramér-Rao lower bound (CRLB) 理论极限的相对差值作为系统的损失函数, 使得 smNet 能够提取到大范围的探测光子以及背景水平的相关信息。在三维单分子开关纳米显微镜图像重建实验中, 传统高斯方法的重建结构存在大量的伪影, 而 smNet 通过读取 PSF 中的额外信息, 获得像差等额外参数并依次建模, 使得其在样品不同深度均实现了高精度, 无伪影的图像重建, 其网络架构与重建的 COS-7 细胞中线粒体蛋白 TOM20 的超分辨率成像结果如图 8(a) 所示。

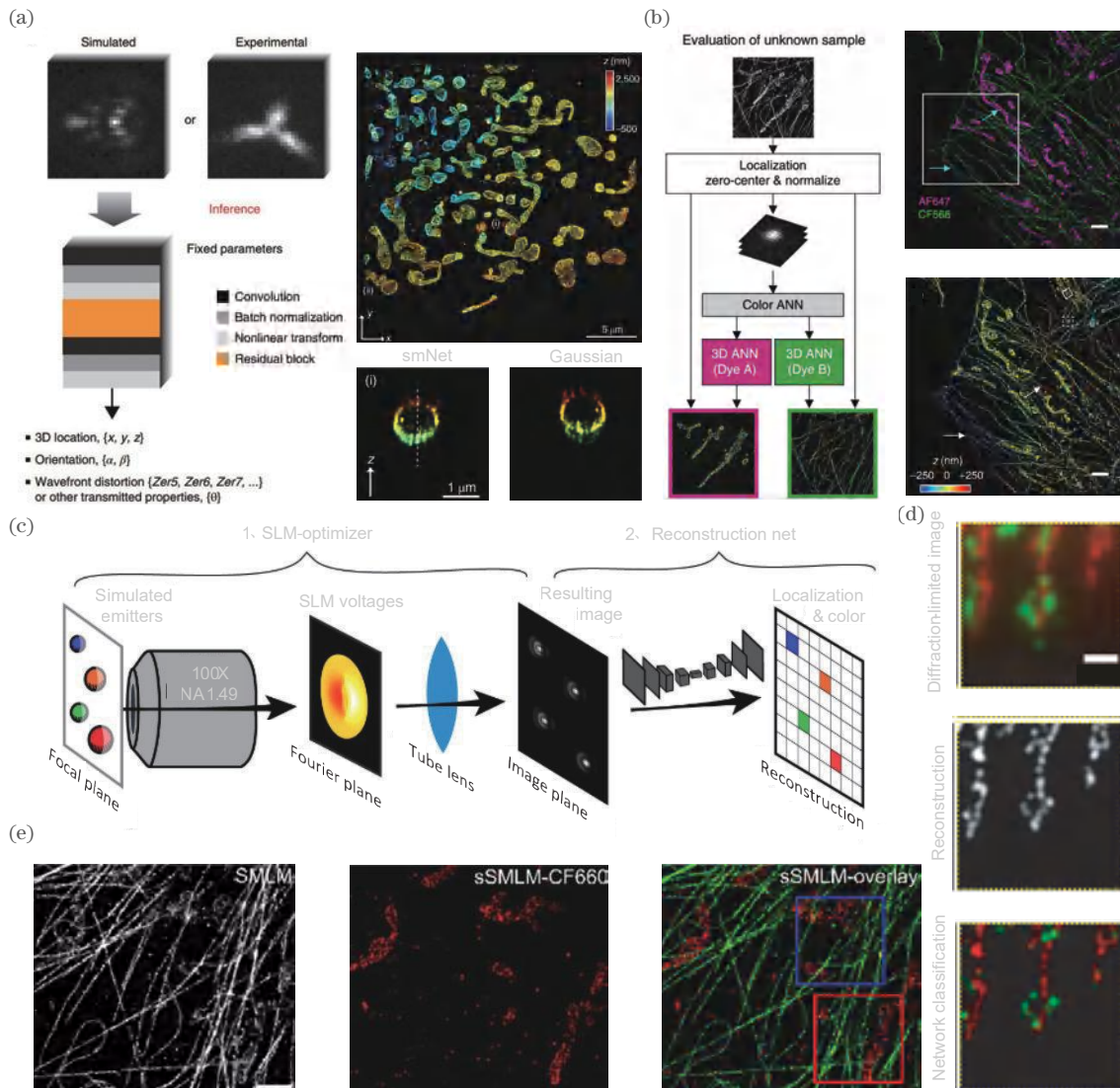


图 8 深度学习从 PSF 中提取附加光谱信息的网络及成像结果。(a) smNet 网络架构和重建效果<sup>[52]</sup>; (b) 颜色分离与轴向定位架构与重建效果<sup>[23]</sup>; (c) 优化相位掩模的图像颜色分类过程图<sup>[53]</sup>; (d) 荧光标记的 HeLa 细胞成像图<sup>[53]</sup>; (e) COS-7 细胞的颜色分类图像<sup>[57]</sup>

Fig. 8 Deep learning extracts additional spectral information from PSF networks and imaging results. (a) smNet network architecture and reconstructed image<sup>[52]</sup>; (b) color separation and axial positioning architecture and reconstruction effects<sup>[23]</sup>; (c) image color classification process diagram of optimized phase mask<sup>[53]</sup>; (d) fluorescent-labeled HeLa cell imaging<sup>[53]</sup>; (e) color classification image of COS-7 cells<sup>[57]</sup>

多色显微成像常使用的方法是在多个传感器之间划分发射光谱,使用光谱滤波片进行光谱分类需要对通道之间进行精确配准,否则将影响重建图像精度,而使用单色光源对样品依次成像进行光谱分类则无法对生物样品进行动态成像,且其较慢的采集过程易使重建图像产生伪影。为实现多色超分辨显微成像, Kim 等<sup>[23]</sup>使用颜色区分网络和轴向定位神经网络从 PSF 中分别预测单个分子的轴向位置和荧光颜色,每当输入一个单分子图像,模型会先将单分子图像定位在二维平面中并进行颜色区分,然后对颜色分离的单分子图像进行轴向定位,再将所得颜色信息和轴向位置信息与每个分子的二维定位相结合,最终生成多维 SMLM 数据。该网络输入的全部数据通过一次成像采集,因此该方法规避了传统多色图像重建过程中的对准问题,其方法架构与重建效果如图 8(b)所示。在 COS-7 细胞双色染色实验中,该方法在光子数为 3000 的情况下,实现了图像的高精度分色重建,且优于传统方法在 5000 光子数下的重建水平,其重建结果如图 8(c)所示。此外,2019 年 Hershko 等<sup>[53]</sup>利用神经网络对多色超分辨成像进行了两步优化:首先,其利用神经网络构建灰度图像颜色分类模型,由灰度图像实现高精度的颜色分类;然后,利用神经网络并结合相位调制算法,予以不同波长不同的相位延迟,生成编码 PSF,以便颜色分类模型对于荧光点进行精确的颜色区分,提高颜色分类模型的分类精度。该网络在四色荧光点分类实验中,正确预测分类了  $96.8\% \pm 2.1\%$  的荧光点,在对荧光标记的 Hela 细胞成像实验中,成功区分了不同荧

光标记的微管和线粒体结构,其网络架构及成像效果如图 8(c)、(d)所示。

为获得额外的光谱信息,研究人员将单分子显微镜与光谱收集通道相结合,构成了光谱单分子定位显微镜(sSMLM)。每个被系统收集到的光子只能随机进入一个光谱通道,因此光谱收集通道的增加会导致每个光谱通道内的光子数减少<sup>[54-56]</sup>,进而导致光谱间分类错误率较高,光谱信息获取困难。为解决该问题, Zhang 等<sup>[57]</sup>利用机器学习分析荧光分子的全谱图,其网络使用全连接层,并将交叉熵作为损失函数,并使用 Adam 算法对网络进行优化。与传统的光谱质心(SC)法相比,该方法提高了 10 倍的分正确率和 2 倍的光谱数据利用率。在微管蛋白与线粒体的双色成像实验中,传统 SMLM 方法难以识别出线粒体结构,而该网络模型清晰分辨了两种结构,且未发生光谱串扰的现象。其网络架构与对光谱分类效果如图 8(e)所示。

#### 4 深度学习在结构光照明显微术的应用

结构光照明显微镜(SIM)使用图案照明激发荧光,通过横向相移和旋转不同的离散角度,获取一系列原始图像,并利用图像重建算法生成超分辨图像。这种超分辨显微方法利用了摩尔条纹的拍频原理,将难以观测的高频信息转移到低频空间中,并在傅里叶域中分离低频信息和高频信息,将分离的信息移动到正确位置并重新组合,从而实现两倍的分辨率增强,其原理图如图 9<sup>[58-60]</sup>所示。SIM 相较于其他的超分辨显微

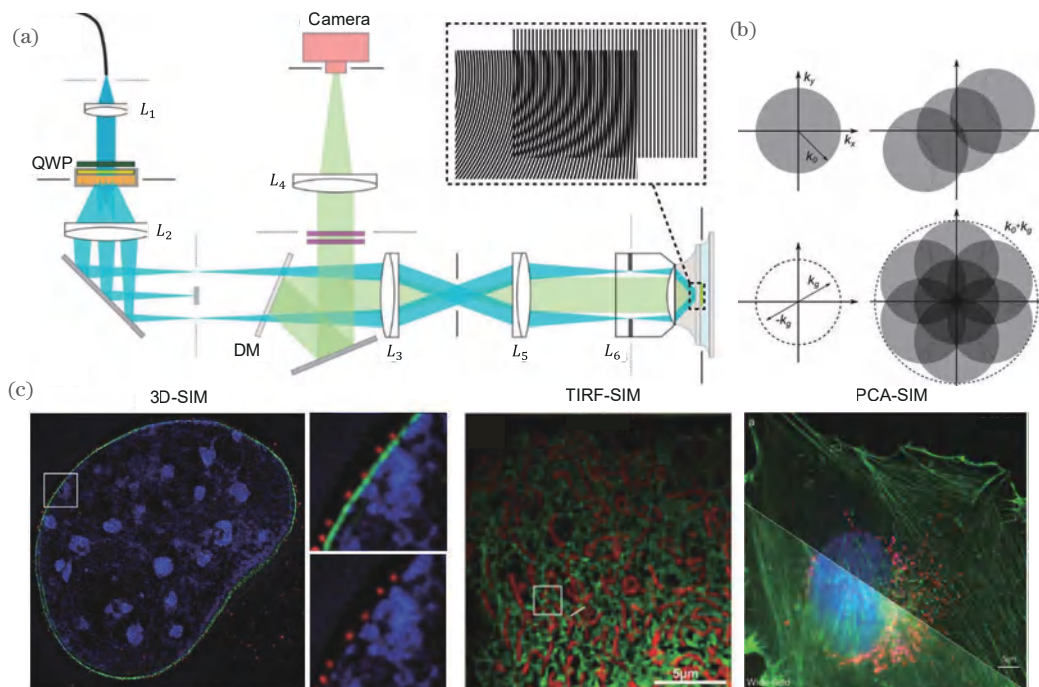


图 9 SIM 成像光路以及原理图。(a) SIM 成像光路图<sup>[58]</sup>; (b) SIM 的光谱扩展<sup>[58]</sup>; (c) SIM 成像图<sup>[58-60]</sup>

Fig. 9 SIM imaging optical path and schematic diagram. (a) SIM imaging optical path diagram<sup>[58]</sup>; (b) spectrum extension of SIM<sup>[58]</sup>; (c) SIM imaging image<sup>[58-60]</sup>



技术,具有极高的光子利用效率,因此其可以降低激发荧光所需的光功率,且成像速度快,这些特性使得 SIM 非常适合应用于活细胞成像。但是,由于多次采集不同角度和相位图像的需求, SIM 仍存在着光漂白和光损伤的问题,影响荧光光强稳定性及其图像重建效果,限制了其在生物样品中的进一步应用。此外, SIM 图像重建算法对计算资源要求高,对噪声敏感,因此在成像过程中的光照畸变、采样速率不足、参数估计错误等会导致重建图像产生伪影。综上所述,对于 SIM 的优化主要集中在进一步减少成像所需帧数、提高定位精度以及定位速度、去除重建图像伪影上<sup>[61, 62]</sup>。

#### 4.1 降低光毒性

在传统的 SIM 方法中,为了分离原始图像中的低频信息和高频信息,往往需要在同一方向上的三张移动照明图像,同时,为了提高同性分辨率还需要至少在三个角度上获得移动照明图像,因此重建超分辨率 SIM 图像至少需要 9 张图像,而对于 3D-SIM 图像重建每个轴向切片则需要 15 张图像,因此传统 SIM 方法仍存在较强的光漂白和光毒性。为进一步减少成像所需要的光剂量,增强 SIM 在生物样品中的应用,最直接的办法就是减少 SIM 重建所需的原始图像。在过去,研究人员开发了多种算法对 SIM 的图像重建进行提升,但这些方法需要对图像形成进行假设,受环境及噪声影响

较大,不仅需要使用者具备先验知识,且泛用性较差。因此,2020 年 Jin 等<sup>[24]</sup>利用深度学习的方法,基于 U-Net 网络,使用堆叠 U-Net 将 15 张原始 SIM 图像作为输入,并使用传统 SIM 重建的超分辨率图像作为真值进行训练,得到了 U-Net-SIM15,使用该网络对未输入过网络的陌生细胞结构进行成像得到了与传统 SIM 重建水平相当的结果。此外, Jin 等在此基础上进一步减少了原始图像的输入量,通过使用三张原始 SIM 图像对网络进行训练,得到了与 U-Net-SIM15 相当的成像质量,该方法被称为 U-Net-SIM3。为最大化降低光损伤, Jin 等降低了所使用的激光功率以及曝光时间,并训练了另一个 U-Net 网络用于获取低曝光图像中的信息,并将该网络的输出结果导入 U-Net-SIM15,利用 skip layer 连接这两个网络形成了新的网络架构,得到了 scU-Net-SIM。该网络可恢复低曝光原始图像,获得与传统 SIM 成像质量相当的成像结果,相较于传统 SIM 其所用原始图像减少为 1/5,光子数减少为 1/100,其网络架构与 scU-Net 重建对比如图 10(a)、(b)所示。类似地,2020 年 Ling 等<sup>[63]</sup>提出使用 CycleGAN 网络实现由三张具有单向相移的原始结构光图像重建出与传统算法相同水平的超分辨率图像。该网络的特点是在训练时不需要训练集与真值集一一对应,相较于 Jin 等使用的 U-Net 网络训练效率更高,极大地减少了成像使用的光子

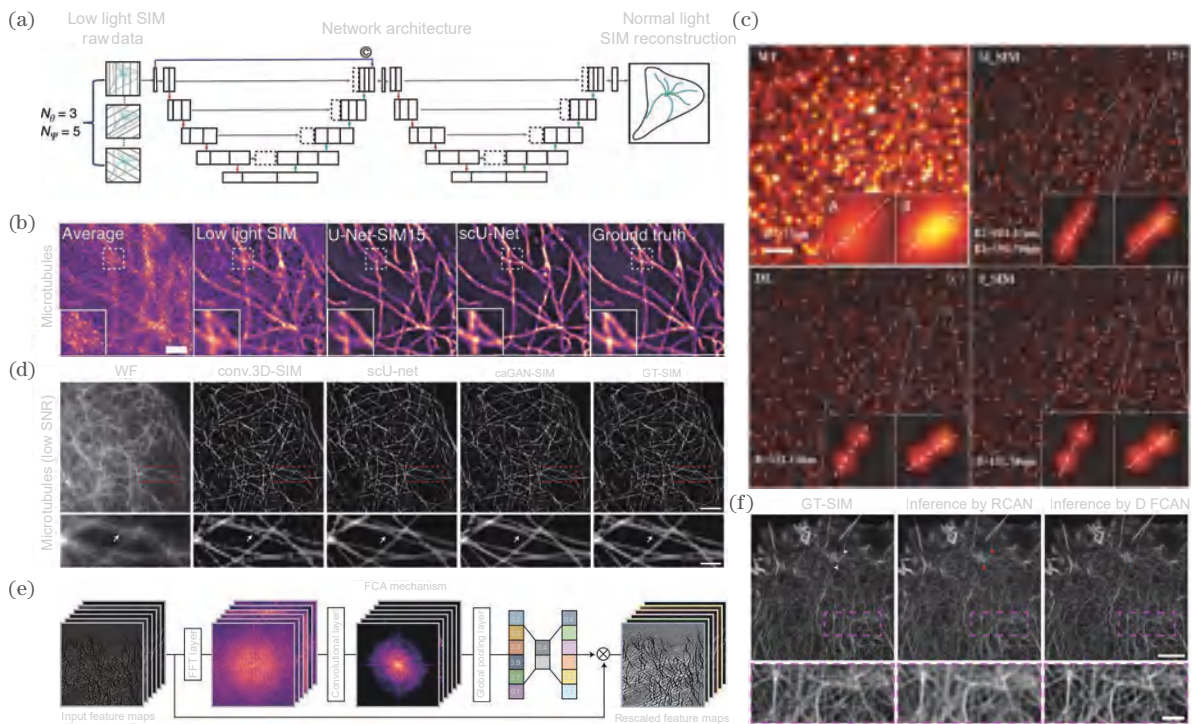


图 10 深度学习在结构光超分辨显微成像中的成像结果。(a) U-Net-SIM3 网络架构<sup>[24]</sup>; (b) 弱光条件下 U-Net-SIM5 和 scU-Net 的重建结果<sup>[24]</sup>; (c) 使用 CycleGAN 网络的 SIM 纳米珠成像图<sup>[63]</sup>; (d) 低信噪比下 caGAN 微管成像<sup>[64]</sup>; (e) DFCAN 的傅里叶注意机制原理示意图<sup>[65]</sup>; (f) DFCAN 重建 f-肌动蛋白细胞骨架图像<sup>[65]</sup>

Fig. 10 Imaging results of deep learning in structured light super-resolution microscopy. (a) U-Net-SIM3 network architecture<sup>[24]</sup>; (b) reconstruction results of U-Net-SIM5 and scU Net under low light conditions<sup>[24]</sup>; (c) SIM nanobead imaging using CycleGAN network<sup>[63]</sup>; (d) caGAN microtubule imaging under low signal-to-noise ratio<sup>[64]</sup>; (e) schematic diagram of the Fourier attention mechanism principle of DFCAN<sup>[65]</sup>; (f) DFCAN reconstruction of f-actin cytoskeleton images<sup>[65]</sup>

数,其对纳米珠重建效果如图 10(c)所示。

对于 3D-SIM, Qiao 等<sup>[64]</sup>于 2021 年基于 GAN 网络架构,建立了一种通道注意力网络(caGAN)用于进行 3D-SIM 重建。使用该方法对微管和溶酶体的动态相互作用进行成像,在记录了 200 个时间点后,样品没有出现明显的光漂白现象,在生物样品上表现优秀。caGAN-SIM 最大的优势在于它可以在图像数量相较于传统 SIM 减少为 1/7.5、总光子数减少为 1/15 的情况下依旧保证良好的成像质量,其微管图重建效果如图 10(d)所示。同年, Qiao 等<sup>[65]</sup>基于傅里叶域中不同特征映射的功率谱特征提出了傅里叶通道注意力网络(DFCAN)并结合生成对抗网络构建 DFGAN,傅里叶通道注意力机制可以依照功率谱中包含的频率分量贡献自适应地调整每个特征映射,相较于空间通道注意力机制,可以更精确地推断出精细结构。对于传统 SIM,线粒体的动态成像过程中需要长时间曝光或高强度照明获得多张原始图像,这会对线粒体造成强光毒性,难以获取完整线粒体的运动过程。而 DFCAN 和 DFGAN 成功实现了对线粒体的超微结构运动的成像。因为 DFCAN 对于单帧的光子数要求远低于传统 SIM,可以长时间对活细胞进行成像,可获取的成像帧数超过 1000 帧,成像时间较传统 SIM 提高了 10 倍,其原理与 f-肌动蛋白细胞骨架重建如图 10(e)、(f)所示。

在 2022 年, Cheng 等<sup>[66]</sup>利用深度学习进一步减少了 SIM 重建所需的图像帧数,他们提出的快速轻量级 SIM 超分辨率网络(FLSN)可以实现将任何角度或相位拍摄的原始帧 SIM 图像转化成相应的超分辨率结果,该方法称为 SF-SIM。SF-SIM 相较于传统的 SIM,其成像速度提升了 14 倍,其中 FLSN 中设计的多内核多尺度网络可以帮助 SF-SIM 适应不同的样本,同时基于 Haar 小波设计,可以使网络更好地去除图像的噪声。

## 4.2 提升成像质量

优秀可靠的算法是获得高分辨 SIM 重建图像的重要途径, SIM 的图像重建算法是 SIM 领域的研究热点,研究者们开发了诸多算法用于对原始图像进行高分辨率重建,例如 fairSIM<sup>[67]</sup>、CC-SIM<sup>[68]</sup>、OpenSIM<sup>[69]</sup>等,但这些算法对重建原始图像的信噪比要求较高。在实际成像过程中,由于样品光漂白、成像时间过短、激发光功率较低等问题,原始图像信噪比较低,重建结果分辨率下降,重建图像出现伪影。针对这一问题, 2021 年 Shah 团队<sup>[25]</sup>提出将经典的计算重建方法 fairSIM 与残差编码器-解码器网络(RED-Net)相结合,并将该工作流程称为 RED-fairSIM。他们对 RED-Net 网络进行了改进,在其编码块后增加了一个上采样块,使其能够对输入的图像在较低维度时进行去噪,减少了训练时间。研究发现,将 fairSIM 重建后图像输入使用了高噪声原始图像训练后的 RED-Net 网络,最终的成像结果解决了 fairSIM 无法重建高分辨率的图

像和重建图像存在伪影的问题。该方法在不同噪声程度的骨肉瘤细胞微管上得以验证,无论是在低噪声水平还是高噪声水平的原始图像上都出色地完成了高质量图像重建任务,其网络架构与对 U2OS 骨肉瘤细胞重建效果如图 11(a)、(b)所示。

传统的 SIM 图像重建,需要原始图像的高对比度,这对实验所使用的光源以及实验的光学精度要求较高,为了克服该问题, Chen 等<sup>[70]</sup>在 2024 年于 *Optics Express* 上提出基于残差神经网络构建的 CR-SIM。该网络的训练集包含了低对比度图像,在对具有较大背景噪声以及厚度的样品测试时, CR-SIM 的图像重建质量显著高于传统的 fairSIM、IM-SIM<sup>[71]</sup>。除此之外, Chen 等还发现该网络可以很好地弥补数字微镜器件结构照明显微镜(DMD-SIM)所存在的缺陷。DMD-SIM 具有结构紧凑、成本低,成像质量高度依赖于原始图像对比度的特点,将 CR-SIM 网络用于 DMD-SIM,在 huFIB 细胞微管极低对比度成像实验中, Fair-SIM、IM-SIM 算法均因为投影系统的低通滤波特性,其调制曲线抑制了高频条纹对比度,无法正确计算出图像的相位以及频率,但 CR-SIM 的特殊设计可以使其避开高频调制的限制,提高成像质量。此外,在不同样品的成像测试中, CR-SIM 表现出了极好泛用性,在微观、网格蛋白包覆凹坑、f-肌动蛋白和线粒体结构中均成功实现了高质量图像重建,且没有引入伪影的现象,其对细胞微管重建效果如图 11(c)所示。

对于多焦点结构照明显微镜(MSIM),其重建过程包含了针孔成像、局部缩放、求和与反卷积,因此对于 MSIM 而言,高效的重建算法非常重要。针对这一特点, 2023 年 Liao 等<sup>[72]</sup>提出使用 CNN 网络,直接建立原始 MSIM 图像与重建后的 MSIM 的映射关系,对图像重建过程进行加速。该方法实现了将 MSIM 的原始图像帧数减少 3/4,在不增加光毒性和光漂白情况下提高 MSIM 体内成像的时间分辨率,且网络对低噪声和高噪声的原始图像都具有一定的重建能力,使得网络只需要训练一次,所得训练权重可以适用于不同的噪声水平的原始数据。在深度为 100  $\mu\text{m}$  的斑马鱼活体内成像实验中, Deep-MSIM 相较于传统的 MISM,实现了成像时间由 162.63 s 到 72.87 s 的提升,其对细胞微管重建效果如图 11(d)所示。

传统的 SIM 重建仅在横向分辨率上实现了超分辨,轴向分辨率仍有待提升, 2008 年 Gustafsson 等<sup>[73]</sup>提出使用三束相干光束实现 SIM 重建图像的轴向分辨率两倍提升的 3D-SIM,在此基础上, Burns 等<sup>[26]</sup>于 2021 年提出使用 RCAN 网络实现从 2D-SIM(两束相干光,三个角度)重建达到 3D-SIM 的轴向分辨率,不仅去除了高图像计数引起的伪影,同时降低了样品光漂白。该网络以 2D-SIM 图像作为输入,以波长减小



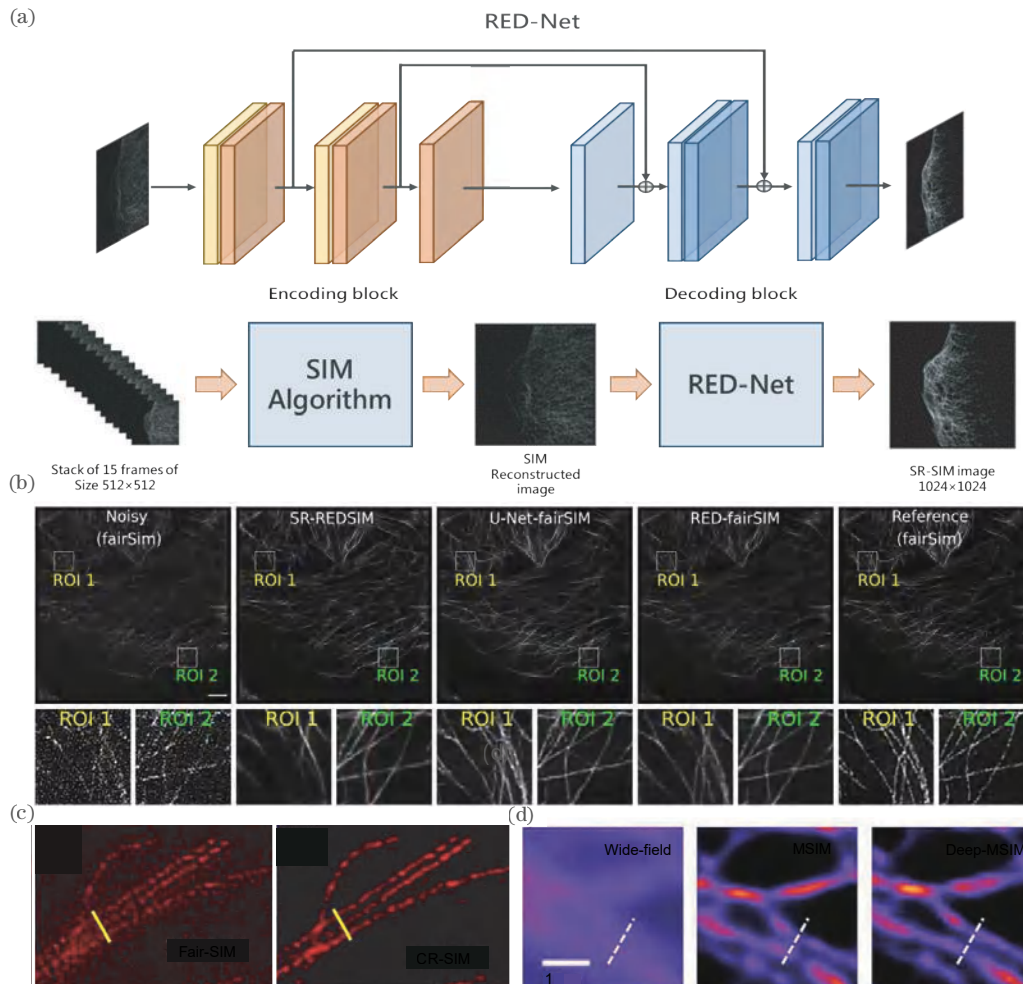


图 11 提升 SIM 图像重建质量的网络架构以及重建结果图。(a) RED-fairSIM 深度学习网络架构图<sup>[25]</sup>；(b) U2OS 骨肉瘤细胞 RED-fairSIM 成像结果图<sup>[25]</sup>；(c) CR-SIM huFIB 细胞微管成像图<sup>[70]</sup>；(d) Deep-MSIM 微管成像效果图<sup>[72]</sup>

Fig. 11 Network architecture and reconstruction result graph for improving the quality of SIM image reconstruction. (a) RED-fairSIM deep learning network architecture diagram<sup>[25]</sup>；(b) RED-fairSIM imaging results of U2OS osteosarcoma cells<sup>[25]</sup>；(c) CR-SIM huFIB cell microtubule imaging image<sup>[70]</sup>；(d) Deep-MSIM microtubule imaging effect<sup>[72]</sup>

$1/\sqrt{2}$  的共聚焦图像作为真值图像,并通过减少共聚焦图像堆栈中的切片厚度来实现真值图像的轴向分辨率提升。该方法在 24 个看不见的模拟染色质结构的测试中,RCAN 重建的半峰全宽(FWHM)为 436.8 nm (标准差为 21.8),对比传统 SIM 平均值的 669.67 nm (标准差为 207.2),轴向分辨率得到了显著提升。

#### 4.3 增强神经网络的泛用性

深度学习在 SIM 领域被广泛应用,但神经网络的泛用性仍有待提升,一是深度学习在各研究小组里的广泛应用,二是神经网络在不同样品上的泛用性。

2021 年 Qiao 等在 *Nature Methods* 上发文,讨论了深度学习在具体什么水平的原始图像重建时,深度学习的方法要优于传统 SIM 技术,同时他们搭建了多模态 SIM 系统,该系统包含了 TIRF-SIM、GI-SIM、非线性 SIM 以及在宽信噪比内获得一一匹配的低分辨率和高分辨率图像。Qiao 等<sup>[65]</sup>使用该系统对网格蛋白包覆凹坑、内质网、微管和肌动蛋白成像,对每一样品

记录 10 种不断提升的激发强度下的 50 组原始 SIM 图像,并保证在最高激发光强下所有原始 SIM 图像具有足够高的信噪比,足以重建高质量 SIM 图像。该成像数据集被命名为 BioSR,如图 12(a)所示,并可以公开获取。在此基础上,Qiao 等对卷积神经网络(SRCNN)、增强型深度神经网络(EDSR)、条件图像转换生成对抗神经网络(Pix2Pix)和跨模态生成对抗神经网络(CMGAN)进行了重建分析,并以归一化均方根误差、多尺度结构相似性指数以及分辨率作为评价标准,划定了在不同样品下传统 SIM 的适用范围,以及每一神经网络重建效果优于传统 SIM 的范围。他们的工作为后续深度学习 SIM 网络提供了训练数据,为评估深度学习超分辨显微网络提供了高质量基准。

目前应用于超分辨显微成像领域的深度学习神经网络大多数采用有监督学习的方法,这些网络结构存在两个缺陷:1)需要训练集和真值集,且大部分网络需

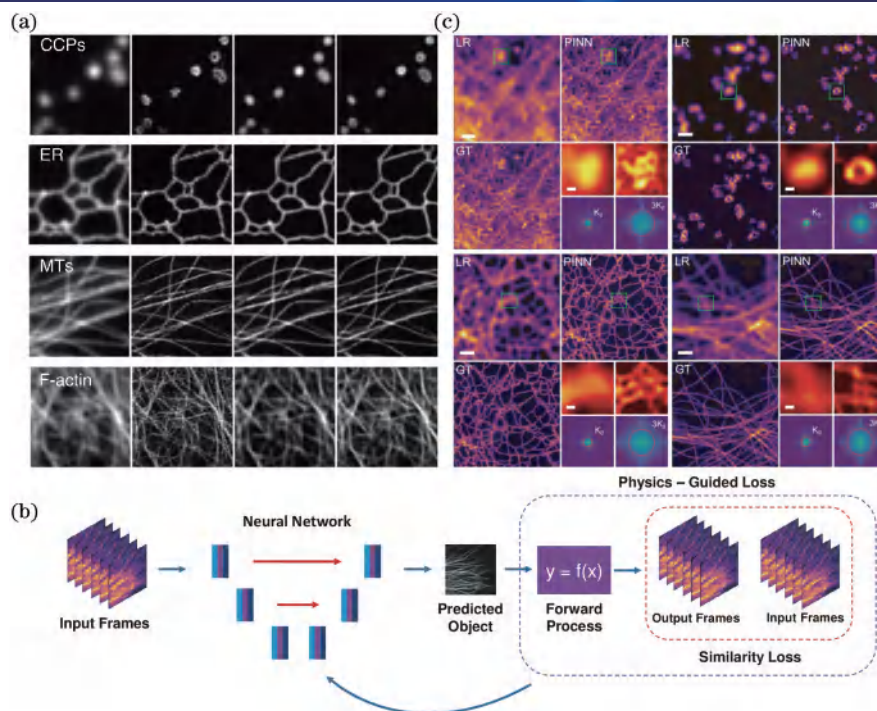


图 12 BioSR 数据集和 PINN 网络相关图像。(a) BioSR 数据集<sup>[65]</sup>; (b) PINN 网络架构<sup>[26]</sup>; (c) 基于 PINN 的非线性 SIM 分辨率在多种对象类型上的优化结果图<sup>[26]</sup>

Fig. 12 BioSR data set and PINN network related images. (a) BioSR data set<sup>[65]</sup>; (b) PINN network architecture<sup>[26]</sup>; (c) optimization results of nonlinear SIM resolution based on PINN for multiple object types<sup>[26]</sup>

要配对的训练集和真值集,但高分辨率的真值图像往往获取困难,通常需要使用其他的超分辨率技术获取; 2) 监督训练的网络对应用对象类别的敏感度较高,当训练网络应用于不同对象时,重建图像可能会表现出训练集特征。为进一步提升深度学习模型的泛用性,2023 年加利福尼亚大学的 Burns 等<sup>[26]</sup>在 *Optics Express* 上提出将深度学习网络与 SIM 正演模型相结合,构建出物理信息神经网络(PINN),其网络架构与在多种对象类型上的优化效果如图 12(b)、(c) 所示,在不使用训练集和真值集的情况下,对有限分辨率的图像直接进行优化。该网络为一个类似于 U-Net 网络框架,低分辨率图像输入网络后,通过 SIM 照明调制后输出一个图像,该图像再经过 SIM 的正向过程生成一系列的子图像,并对比该图像与之前的输入图像,直至二者之间的损失函数趋于稳定,实现高分辨率图像输出。该网络在线性 SIM、非线性 SIM,以及 f-肌动蛋白、网格蛋白包被凹坑(CCP)、内质网和微管等不同结构上均表现优秀,提高了成像分辨率,且不需要对网络进行多次训练,具有鲁棒性。

## 5 跨模态超分辨成像

除去在现有超分辨率技术使用深度学习克服其缺陷使其达到更好的成像效果的方法外,研究者们还尝试使用深度学习将非超分辨成像技术所成图像直接转换为超分辨图像,即跨模态超分辨成像(CM)。由于在

相同的成像质量下,共聚焦显微镜的激发功率比 STED 所需的激发功率低 1/10~1/3,如果能在该情况下使共聚焦显微镜获得与 STED 相同的分辨率,则可以极大地降低成像所需的光剂量,进而减少光漂白和光毒性。为此 Wang 等<sup>[74]</sup>于 2018 年在 *Nature Methods* 上提出利用深度学习实现荧光显微镜的跨模态转换,他们使用 GAN 实现了共聚焦图像与 STED 相匹配的分辨率,全内反射荧光显微镜(TIRF)图像获得与基于 TIRF 的结构光照明显微镜相匹配的分辨率。该方法降低了超分辨成像的门槛,使超分辨显微成像可以在更多系统上普及,其网络框架、共聚焦到 STED 的跨模态转换图像以及对纳米珠重建效果如图 13(a)、(b) 所示。

2023 年 Huang 等<sup>[27]</sup>提出一种双通道注意力网络(TCAN),该网络基于条件生成对抗网络(cGAN),生成器使用了 U-Net 和深度傅里叶通道注意力网络(DFCAN),这样的设计使得该网络不仅可以提高跨不同数据集的预测性能,还可以精准地学习图像的高频信息,获得更精准的映射结果。该方法在直径 23 nm 的纳米珠测试实验中由 TCAN 输出的共聚焦网络图像 PSF 的 FWHM 达到了  $(58 \pm 1)$  nm,而传统 STED 图像仅为  $(83 \pm 9)$  nm。在 HeLa 细胞核成像实验中,TCAN 比 STED 图像能更好地解析密集标记的核孔复合物(NPC),并降低了背景噪声,在保留有用信息和去噪之间实现了折中。最后,在多种生物结构以



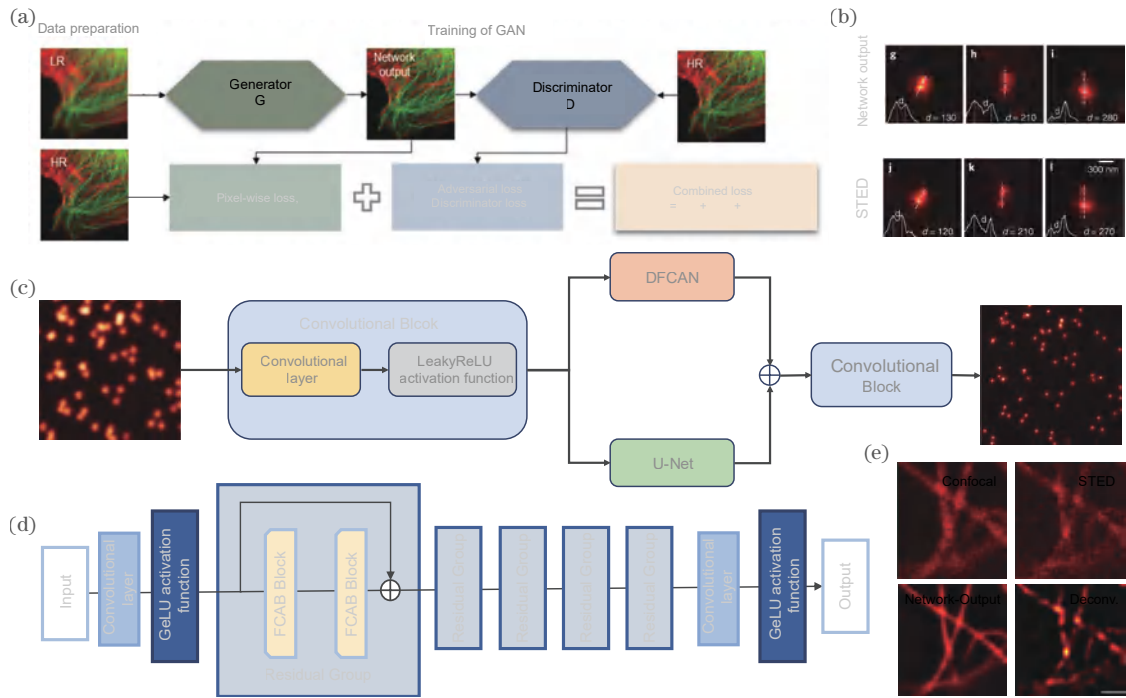


图 13 深度学习实现跨模态转换网络结构图以及结果图。(a)深度学习实现荧光显微镜的跨模态转换网络框架<sup>[74]</sup>;(b)从共聚焦到 STED 的跨模态图像转换结果图<sup>[74]</sup>;(c) TCAN 网络生成器部分网络结构图<sup>[27]</sup>;(d) DFCAN 机制网络图<sup>[27]</sup>;(e) TCAN 对微管进行超分辨率成像结果图<sup>[27]</sup>

Fig. 13 Deep learning implements cross modal transformation network structure diagram and result diagram. (a) Deep learning implementation of cross modal transformation network framework for fluorescence microscopy<sup>[74]</sup>; (b) results of cross modal image conversion from confocal to STED<sup>[74]</sup>; (c) partial network structure diagram of TCAN network generator<sup>[27]</sup>; (d) DFCAN mechanism network diagram<sup>[27]</sup>; (e) TCAN super-resolution imaging results of microtubules<sup>[27]</sup>

及肌动蛋白微观双色成像实验中,将双色共聚焦图像的分辨率从 230 nm 提升到 110 nm,其网络架构以及对微管重建效果如图 13(d)、(e)所示。

## 6 结束语

现有的超分辨显微技术由于其成像原理而存在着成像速度慢,重建图像存在伪影,对生物样品光损伤大等问题,使用传统物理方法或算法解决这些问题往往较为复杂,或者难以取得期望效果。深度学习以其优秀的解决图像重建逆问题能力而极大地克服了超分辨显微成像技术中存在的各种缺陷,但深度学习方法在超分辨显微成像领域中的应用仍面临着一系列的挑战:1) 目前在超分辨显微成像领域中使用的神经网络主要还是以有监督学习为主,为获取精确高质量模型,需要使用大量一一对应的低分辨率图像和高分辨率图像,但高分辨率图像往往获取困难,这限制了深度学习方法在超分辨显微成像领域中的广泛应用。2) 深度学习模型训练成本高,神经网络越深越复杂,其训练模型的准确性越高,但这需要使用大量的计算资源,训练时间较长。3) 深度学习模型重建图像的可信度有待提升。深度学习模型不同于传统的具有严密理论推导与证明的光学理论模型,其直接建立图像间的特征映射的特点使其更像一个“黑盒”,因此在学界缺乏理论说服力。

针对以上几种挑战,可以从如下几个方面寻求解决方法:1) 建立高质量、样本种类丰富的公共数据集,促进深度学习的广泛应用。如 Qiao 等<sup>[65]</sup>建立的 BioSR 的 SIM 数据集,包含多样的样本和对应的高低分辨率图像,以供研究人员共享和使用。2) 根据训练任务设计神经网络,轻量化网络模型,使用特殊设计减少神经网络的层数,减少计算资源的使用和训练所需要的时间。3) 提高网络的泛用性。推进无监督网络模型在超分辨显微成像领域的进一步应用,降低模型训练的高分辨率图像需求,同时增强网络鲁棒性,也可以借鉴迁移学习的方法,将在一个样本上学到的特征映射迁移到另一个相关任务上,使得网络适用于不同的应用场景。4) 提高模型的可解释性。通过可解释性机制,深度学习模型的决策过程更加透明。采用可视化技术、注意力机制和可解释性网络设计等方法,使用户可以更清晰地了解到模型网络的机制,提高网络的可信度。

综上所述,深度学习目前已经在超分辨显微成像领域取得了诸多优秀的成果,尽管其仍存在一定的问題,但可以预见的是,随着深度学习理论进一步完善以及实践的不断进步,这些问题将被逐步解决,深度学习技术将大力推动着超分辨显微成像技术的持续进步,创造更多优秀的成果!



## 参 考 文 献

- [1] Chen X D, Zou C L, Gong Z J, et al. Subdiffraction optical manipulation of the charge state of nitrogen vacancy center in diamond[J]. *Light: Science & Applications*, 2015, 4(1): e230.
- [2] Yang T J, Luo Y R, Ji W, et al. Advancing biological super-resolution microscopy through deep learning: a brief review[J]. *Biophysics Reports*, 2021, 7(4): 253-266.
- [3] Rust M J, Bates M, Zhuang X W. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)[J]. *Nature Methods*, 2006, 3(10): 793-795.
- [4] Betzig E, Patterson G H, Sougrat R, et al. Imaging intracellular fluorescent proteins at nanometer resolution [J]. *Science*, 2006, 313(5793): 1642-1645.
- [5] Chen C H, Wang F, Wen S H, et al. Multi-photon near-infrared emission saturation nanoscopy using upconversion nanoparticles[J]. *Nature Communications*, 2018, 9: 3290.
- [6] Denk W, Strickler J H, Webb W W. Two-photon laser scanning fluorescence microscopy[J]. *Science*, 1990, 248 (4951): 73-76.
- [7] Liu Y T, Wang F, Lu H X, et al. Super-resolution mapping of single nanoparticles inside tumor spheroids[J]. *Small*, 2020, 16(6): e1905572.
- [8] Liu Y T, Wen S H, Wang F, et al. Population control of upconversion energy transfer for stimulation emission depletion nanoscopy[J]. *Advanced Science*, 2023, 10 (20): e2205990.
- [9] Cao Y Y, Li X P, Gu M. Super-resolution nanofabrication with metal-ion doped hybrid material through an optical dual-beam approach[J]. *Applied Physics Letters*, 2014, 105(26): 263102.
- [10] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. *Communications of the ACM*, 2017, 60(6): 84-90.
- [11] Friedrich M, Gan Q, Ermolayev V, et al. STED-SPIM: stimulated emission depletion improves sheet illumination microscopy resolution[J]. *Biophysical Journal*, 2011, 100 (8): L43-L45.
- [12] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation [M]//Navab N, Hornegger J, Wells W M, et al. *Medical image computing and computer-assisted intervention-MICCAI 2015. Lecture notes in computer science*. Cham: Springer, 2015, 9351: 234-241.
- [13] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.
- [14] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(6): 1137-1149.
- [15] 王佳林, 严伟, 张佳, 等. 受激辐射损耗超分辨显微成像系统研究的新进展[J]. *物理学报*, 2020, 69(10): 108702.
- [16] Wang J L, Yan W, Zhang J, et al. New advances in the research of stimulated emission depletion super-resolution microscopy[J]. *Acta Physica Sinica*, 2020, 69(10): 108702.
- [17] Schermelleh L, Ferrand A, Huser T, et al. Super-resolution microscopy demystified[J]. *Nature Cell Biology*, 2019, 21(1): 72-84.
- [18] Ebrahimi V, Stephan T, Kim J, et al. Deep learning enables fast, gentle STED microscopy[J]. *Communications Biology*, 2023, 6(1): 674.
- [19] Li M Z. Deep adversarial network for super stimulated emission depletion imaging[J]. *Journal of Nanophotonics*, 2020, 14(1): 016009.
- [20] 安莎, 但旦, 于湘华, 等. 单分子定位超分辨显微成像技术研究进展及展望(特邀综述)[J]. *光子学报*, 2020, 49 (9): 0918001.
- [21] An S, Dan D, Yu X H, et al. Progress and prospect of research on single-molecule localization super-resolution microscopy (invited review) [J]. *Acta Photonica Sinica*, 2020, 49(9): 0918001.
- [22] Nehme E, Weiss L E, Michaeli T, et al. Deep-STORM: super-resolution single-molecule microscopy by deep learning[J]. *Optica*, 2018, 5(4): 458-464.
- [23] Gaire S K, Zhang Y, Li H Y, et al. Accelerating multicolor spectroscopic single-molecule localization microscopy using deep learning[J]. *Biomedical Optics Express*, 2020, 11(5): 2705-2721.
- [24] Möckl L, Roy A R, Petrov P N, et al. Accurate and rapid background estimation in single-molecule localization microscopy using the deep neural network BNet[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2020, 117(1): 60-67.
- [25] Kim T, Moon S, Xu K. Information-rich localization microscopy through machine learning[J]. *Nature Communications*, 2019, 10: 1996.
- [26] Jin L H, Liu B, Zhao F Q, et al. Deep learning enables structured illumination microscopy with low light levels and enhanced speed[J]. *Nature Communications*, 2020, 11: 1934.
- [27] Shah Z H, Müller M, Wang T C, et al. Deep-learning based denoising and reconstruction of super-resolution structured illumination microscopy images[J]. *Photonics Research*, 2021, 9(5): B168-B181.
- [28] Burns Z, Liu Z W. Untrained, physics-informed neural networks for structured illumination microscopy[J]. *Optics Express*, 2023, 31(5): 8714-8724.
- [29] Huang B Y, Li J, Yao B W, et al. Enhancing image resolution of confocal fluorescence microscopy with deep learning[J]. *PhotoniX*, 2023, 4(1): 2.
- [30] Wang Y F, Kuang C F, Gu Z T, et al. Time-gated stimulated emission depletion nanoscopy[J]. *Optical Engineering*, 2013, 52(9): 093107.
- [31] Hell S W, Wichmann J. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy[J]. *Optics Letters*, 1994, 19(11): 780-782.
- [32] Heller I, Sitters G, Broekmans O D, et al. STED nanoscopy combined with optical tweezers reveals protein

- dynamics on densely covered DNA[J]. *Nature Methods*, 2013, 10(9): 910-916.
- [31] Liu Y T, Zhou J J, Wen S H, et al. On-chip mirror enhanced multiphoton upconversion super-resolution microscopy[J]. *Nano Letters*, 2023, 23(12): 5514-5519.
- [32] Takasaki K T, Ding J B, Sabatini B L. Live-cell superresolution imaging by pulsed STED two-photon excitation microscopy[J]. *Biophysical Journal*, 2013, 104(4): 770-777.
- [33] Kilian N, Goryaynov A, Lessard M D, et al. Assessing photodamage in live-cell STED microscopy[J]. *Nature Methods*, 2018, 15(10): 755-756.
- [34] Tortarolo G, Castello M, Diaspro A, et al. Evaluating image resolution in stimulated emission depletion microscopy[J]. *Optica*, 2018, 5(1): 32-35.
- [35] Chen Y I, Chang Y J, Sun Y S, et al. Spatial resolution enhancement in photon-starved STED imaging using deep learning-based fluorescence lifetime analysis[J]. *Nanoscale*, 2023, 15(21): 9449-9456.
- [36] Pavani S R P, Thompson M A, Biteen J S, et al. Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2009, 106(9): 2995-2999.
- [37] Lew M D, Lee S F, Badieirostami M, et al. Corkscrew point spread function for far-field three-dimensional nanoscale localization of pointlike objects[J]. *Optics Letters*, 2011, 36(2): 202-204.
- [38] Ji Y H, Chen D N, Wu H Z, et al. Localizing axial dense emitters based on single-helix point spread function and deep learning[EB/OL]. (2024-02-10) [2024-05-06]. <http://arxiv.org/abs/2402.06863v2>.
- [39] Pan W H, Li W, Qu J H, et al. Research progress on organic fluorescent probes for single molecule localization microscopy[J]. *Chinese Journal of Applied Chemistry*, 2019, 36: 269-281.
- [40] Hess S T, Girirajan T P K, Mason M D. Ultra-high resolution imaging by fluorescence photoactivation localization microscopy[J]. *Biophysical Journal*, 2006, 91(11): 4258-4272.
- [41] Colabrese S, Castello M, Vicidomini G, et al. Machine learning approach for single molecule localisation microscopy[J]. *Biomedical Optics Express*, 2018, 9(4): 1680-1691.
- [42] Sahel Y B, Eldar Y. Self-STORM: deep unrolled self-supervised learning for super-resolution microscopy[EB/OL]. (2024-03-25) [2024-05-06]. <http://arxiv.org/abs/2403.16974v1>.
- [43] Li J Y, Tong G, Pan Y N, et al. Spatial and temporal super-resolution for fluorescence microscopy by a recurrent neural network[J]. *Optics Express*, 2021, 29(10): 15747-15763.
- [44] Boyd N, Jonas E, Babcock H, et al. DeepLoco: fast 3D localization microscopy using neural networks[EB/OL]. (2018-02-16) [2024-05-06]. <https://www.biorxiv.org/content/10.1101/267096v1>.
- [45] Zelger P, Kaser K, Rossboth B, et al. Three-dimensional localization microscopy using deep learning[J]. *Optics Express*, 2018, 26(25): 33166-33179.
- [46] Speiser A, Müller L R, Hoess P, et al. Deep learning enables fast and dense single-molecule localization with high accuracy[J]. *Nature Methods*, 2021, 18(9): 1082-1090.
- [47] Ouyang W, Aristov A, Lelek M, et al. Deep learning massively accelerates super-resolution localization microscopy[J]. *Nature Biotechnology*, 2018, 36(5): 460-468.
- [48] Isola P, Zhu J Y, Zhou T H, et al. Image-to-image translation with conditional adversarial networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5967-5976.
- [49] Saguy A, Alalouf O, Opatovski N, et al. DBlink: dynamic localization microscopy in super spatiotemporal resolution via deep learning[J]. *Nature Methods*, 2023, 20(12): 1939-1948.
- [50] Cascarano P, Comes M C, Sebastiani A, et al. DeepCELO for 2D single-molecule localization in fluorescence microscopy[J]. *Bioinformatics*, 2022, 38(5): 1411-1419.
- [51] Nehme E, Freedman D, Gordon R, et al. DeepSTORM 3D: dense 3D localization microscopy and PSF design by deep learning[J]. *Nature Methods*, 2020, 17(7): 734-740.
- [52] Zhang P Y, Liu S, Chaurasia A, et al. Analyzing complex single-molecule emission patterns with deep learning[J]. *Nature Methods*, 2018, 15(11): 913-916.
- [53] Hershko E, Weiss L E, Michaeli T, et al. Multicolor localization microscopy and point-spread-function engineering by deep learning[J]. *Optics Express*, 2019, 27(5): 6158-6183.
- [54] Jeong D, Kim D. Super-resolution fluorescence microscopy-based single-molecule spectroscopy[J]. *Bulletin of the Korean Chemical Society*, 2022, 43(3): 316-327.
- [55] Kim G H, Chung J, Park H, et al. Single-molecule sensing by grating-based spectrally resolved super-resolution microscopy[J]. *Bulletin of the Korean Chemical Society*, 2021, 42(2): 270-278.
- [56] Zhang Z Y, Kenny S J, Hauser M, et al. Ultrahigh-throughput single-molecule spectroscopy and spectrally resolved super-resolution microscopy[J]. *Nature Methods*, 2015, 12(10): 935-938.
- [57] Zhang Z Y, Zhang Y, Ying L, et al. Machine-learning based spectral classification for spectroscopic single-molecule localization microscopy[J]. *Optics Letters*, 2019, 44(23): 5864-5867.
- [58] Brunstein M, Wicker K, Hérault K, et al. Full-field dual-color 100-nm super-resolution imaging reveals organization and dynamics of mitochondrial and ER networks[J]. *Optics Express*, 2013, 21(22): 26162-26173.
- [59] Schermelleh L, Carlton P M, Haase S, et al. Subdiffraction multicolor imaging of the nuclear periphery with 3D structured illumination microscopy[J]. *Science*, 2008, 320(5881): 1332-1336.
- [60] Turcotte R, Liang Y J, Tanimoto M, et al. Dynamic super-resolution structured illumination imaging in the

- living brain[J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(19): 9586-9591.
- [61] Ma Y, Wen K, Liu M, et al. Recent advances in structured illumination microscopy[J]. Journal of Physics: Photonics, 2021, 3(2): 024009.
- [62] Saxena M, Eluru G, Gorthi S S. Structured illumination microscopy[J]. Advances in Optics and Photonics, 2015, 7(2): 241-275.
- [63] Ling C, Zhang C L, Wang M Q, et al. Fast structured illumination microscopy via deep learning[J]. Photonics Research, 2020, 8(8): 1350-1359.
- [64] Qiao C, Chen X Y, Zhang S W, et al. 3D structured illumination microscopy via channel attention generative adversarial network[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2021, 27(4): 6801711.
- [65] Qiao C, Li D, Guo Y T, et al. Evaluation and development of deep neural networks for image super-resolution in optical microscopy[J]. Nature Methods, 2021, 18(2): 194-202.
- [66] Cheng X, Li J, Dai Q, et al. Fast and lightweight network for single frame structured illumination microscopy super-resolution[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 5007711.
- [67] Müller M, Mönkemöller V, Hennig S, et al. Open-source image reconstruction of super-resolution structured illumination microscopy data in ImageJ[J]. Nature Communications, 2016, 7: 10980.
- [68] Wicker K, Mandula O, Best G, et al. Phase optimisation for structured illumination microscopy[J]. Optics Express, 2013, 21(2): 2032-2049.
- [69] Delp S L, Anderson F C, Arnold A S, et al. OpenSim: open-source software to create and analyze dynamic simulations of movement[J]. IEEE Transactions on Bio-Medical Engineering, 2007, 54(11): 1940-1950.
- [70] Chen Y B, Liu Q Q, Zhang J F, et al. Deep learning enables contrast-robust super-resolution reconstruction in structured illumination microscopy[J]. Optics Express, 2024, 32(3): 3316-3328.
- [71] Cao R Z, Chen Y H, Liu W J, et al. Inverse matrix based phase estimation algorithm for structured illumination microscopy[J]. Biomedical Optics Express, 2018, 9(10): 5037-5051.
- [72] Liao J H, Zhang C S, Xu X C, et al. Deep-MSIM: fast image reconstruction with deep learning in multifocal structured illumination microscopy[J]. Advanced Science, 2023, 10(27): e2300947.
- [73] Gustafsson M G L, Shao L, Carlton P M, et al. Three-dimensional resolution doubling in wide-field fluorescence microscopy by structured illumination[J]. Biophysical Journal, 2008, 94(12): 4957-4970.
- [74] Wang H D, Rivenson Y, Jin Y Y, et al. Deep learning enables cross-modality super-resolution in fluorescence microscopy[J]. Nature Methods, 2019, 16(1): 103-110.





# Optics Letters

## Learning-based single-shot long-range synthetic aperture Fourier ptychographic imaging with a camera array

BOWEN WANG,<sup>1,2</sup>  SHENG LI,<sup>1,2</sup> QIAN CHEN,<sup>1,2</sup>  AND CHAO ZUO<sup>1,2,\*</sup> 

<sup>1</sup>Smart Computational Imaging Laboratory (SCILab), School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China

<sup>2</sup>Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China

\*Corresponding author: zuochao@njust.edu.cn

Received 24 October 2022; revised 28 November 2022; accepted 28 November 2022; posted 29 November 2022; published 2 January 2023

In this Letter, we report a new long-range synthetic aperture Fourier ptychographic imaging technique, termed learning-based single-shot synthetic aperture imaging (LSS-SAI). LSS-SAI uses a camera array to record low-resolution intensity images corresponding to different non-overlapping spectral regions in parallel, which are synthesized to reconstruct a super-resolved high-quality image based on a physical model-based dual-regression deep neural network. Compared with conventional macroscopic Fourier ptychographic imaging, LSS-SAI overcomes the stringent requirement on a large amount of raw data with a high spectral overlapping ratio for high-resolution, high signal-to-noise imaging of reflective objects with diffuse surfaces, making single-shot long-range synthetic aperture imaging possible. Experimental results on rough reflective samples show that our approach can improve the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) by 10.56 dB and 0.26, respectively. We also demonstrate the single-shot ptychography capability of the proposed approach by the synthetic aperture imaging of a dynamic scene at a camera-limited speed (30 fps). To the best of our knowledge, this is the first demonstration of macroscopic Fourier ptychography to single-shot synthetic aperture imaging of dynamic events.

© 2023 Optica Publishing Group

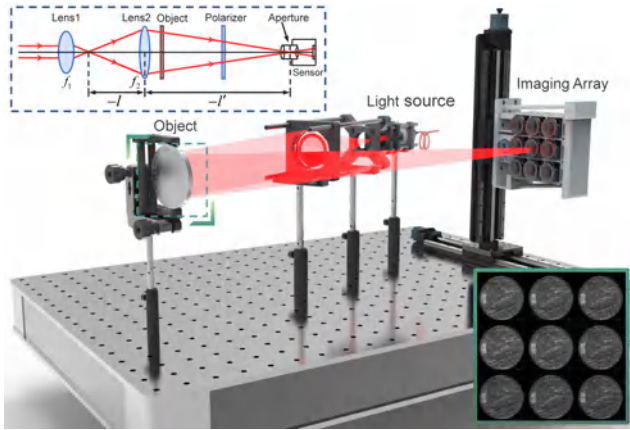
<https://doi.org/10.1364/OL.479074>

Acquiring high-resolution images has emerged as an indispensable requirement in application scenarios such as astronomy, remote sensing, and geological exploration. A major limitation of remote imaging detection is the spatial resolution, which is jointly capped by the finite aperture size and pixel size corresponding to the Nyquist sampling frequency. Astronomers attempted to extend the effective aperture of the system by either employing a well-designed large-aperture lens or splicing the primary mirror, posing significant challenges for lightweight designs. This approach also tends to introduce optical aberrations with bulky dimensions, which precludes its feasibility in practical imaging. Pioneering research has emerged to

circumvent the inherent limitations of imaging systems, e.g., coherent optical detection [1] and flat plate interference [2].

As a promising and elegant computational imaging approach, Fourier ptychographic microscopy (FPM) [3], invented in 2013, breaks the trade-off between the large field of view and high-resolution (HR) with a combination of synthetic aperture radar (SAR) [4] and optical phase retrieval [5]. Combined with the concept of Fourier optics, the imaging process can be understood as sampling the different regions of the HR Fourier domain of an object. Its application potential has been demonstrated in both microscopic biomedical imaging [6–8], and remote sensing [9,10], and meanwhile, the technology has been incorporated in the latest Fourier optics publications. Undoubtedly, during each acquisition, a certain amount of redundant information (at least 35% aperture overlapping percentage [11] in the Fourier domain) needs to be leveraged to perform the lost phase information decoupling, as the sensor can only record intensity information. The converged intensity and phase images are yielded by iterative optimization, jointly imposing both space and frequency-domain constraints on the observed data. This, in turn, is laborious, which implies it is less suitable for dynamic scenes, hampering its application in dynamic scenarios (default: the observed scene remains stationary over a time-lapse). Adaptive compensation [12] and simulated annealing correction algorithms [13] have also been proposed successively to tackle the artifact phenomenon in the reconstruction results, providing fast convergence speed with few computational overheads. Motivated by the rise of convolutional neural network (CNN) techniques [14] and their flexibility to the prior latent features as network layers, many efforts [15–17] to refine the FPM framework have been catalyzed. It has been proven that Neural networks can simultaneously restore the envisaged images by aggregating multi-scale features and nonlinear mappings [18,19], notably in the field of phase recovery.

In this Letter, we report a new super-resolution technique, termed learning-based single-shot synthetic aperture imaging, which is capable of “regenerating” the lost spatial resolution with deep learning. The proposed method leverages the advantages of deep learning data fitting to address the problem of



**Fig. 1.** Overview of the proposed LSS-SAI framework. The object is illuminated by a fiber laser, and the Fourier spectrum is formed at the aperture plane. The specific optical path tracing diagram is shown in the upper left-hand corner.

speckles and artifacts in reconstruction, rekindling sparse aperture, single shot, and ambiguity-free super-resolved imaging. Inspired by a priori knowledge in image processing, we could impute the extension of the Fourier spectrum to the physical prior of four elaborate-designed parallel network architectures. CNN can excavate more texture features from the original multiple encoded images, refining the details while balancing the speckle interference. The iterative non-negativity constraint was further employed to compute the filling of the missing information, yielding optimal outcomes. As a similar concept, the learning-based network [20,21] has achieved tremendous success in searching the map functions (statistical model of desirable target and the observational data) for various underdetermined imaging problems, verifying the feasibility of constraining the inverse problem through the network.

In the construction system, as shown in Fig. 1, a coherent laser is introduced to illuminate the  $3 \times 3$  camera array on the receiver side. The sub-aperture will receive the wave vector of the incident beam from different angles, and the specific imaging process can be expressed as follows:

$$o_{\text{output}}(x, y) = h(x, y) \otimes [o_{\text{input}}(x, y) e^{i(u_m, n, x + v_m, n, y)}], \quad (1)$$

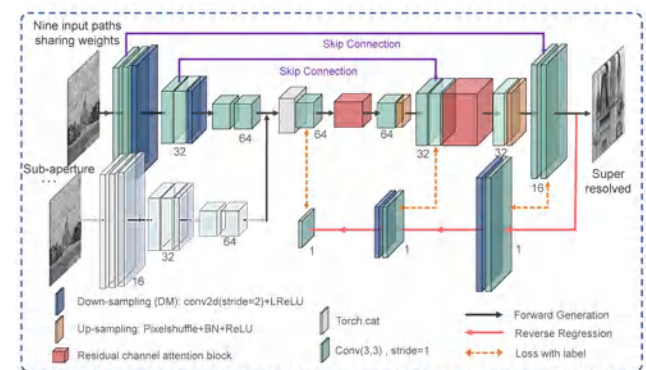
where  $o_{\text{input}}(x, y)$  and  $o_{\text{output}}(x, y)$  represent the complex amplitudes of the input and output optical fields, respectively;  $\otimes$  represents the two-dimensional convolution process;  $e^{i(u_m, n, x + v_m, n, y)}$  denotes the wave vector of the  $m$ th row and  $n$ th column of the angled incident plane light wave. The light field at the imaging aperture plane can be described as  $\Psi(u, v)$  for the sensor, which can only record intensity information, and the measurements are phaseless, whereupon the intensity information recorded is formulated as follows:

$$I(x, y) = |\mathcal{F}^{-1} [\Psi(u, v) \cdot P(u - u_c, v - v_c)]|^2, \quad (2)$$

where  $(u_c, v_c)$  is the center of the aperture;  $P(u, v)$  is the Fourier transform of  $h(x, y)$ , i.e., the coherent transfer function of the optical imaging system ( $NA/\lambda$ ). It is also the intrinsic concept of FP to increase the resolution by obtaining the sub-spectrum at different locations, thus extending the range of the equivalent spectrum and widening the size of the equivalent aperture. We mainly focus on the reconstruction quality enhancement of

the long-range rough reflective samples. Therefore, it is necessary to consider the influence of phase fluctuations on the rough surface, which implies that a random phase distribution will be integrated. The ingenious exploitation of angular illumination to mitigate the influence of speckle noise is a thoughtful approach to coherent synthetic aperture imaging. Object information is encrypted in disordered-seeming speckles, especially those related to the phase fluctuations of the surface, which can be decoded by inverse transmission matrices and CNN. From the perspective of information optics, the whole reflection process can be analogized to the dot product between the scattering layer and the smooth non-diffuse target [9]. We set the scattering intensity  $A$  and the corresponding phase  $\varphi$  randomly distributed in the interval  $[0, 1]$  and  $[0, 2\pi]$ , respectively. Thereby, the complex amplitude distribution  $S$  can be expressed as  $S = A \exp(i\varphi)$ .

The proposed verification platform contains nine imaging sensors (pixel size of  $1.85 \mu\text{m}$ ) equipped with the FUJINON lens (75 mm focal length,  $F^\#$  from 2.8 to 16). Coherent illumination conditions are required to be satisfied in FP imaging. The illumination source employed in the demonstration is a semiconductor laser with a wavelength of 632 nm and a maximum power of 5 mW. The distance between the measured object and our system is 2.3 m. The specific imaging optical path is shown in Fig. 1, and we can observe that each sub-camera is tightly arranged (non-overlapping) with the corresponding acquired images presented in the right-hand corner. For captured image pairs, the sequence captured by the camera with a lens (F-number 4) is recorded as the ground truth HR sequence, and the sequence captured by the camera with a lens (F-number 12) is adapted to generate the corresponding LR sequence, generating a dataset for  $3\times$  super-resolved imaging. Based on this setup, we built a dataset containing 1000 raw data tailored by off-the-shelf detectors for the proposed network. Figure 2 reveals the intrinsic model of the proposed network, which is a nine-path CNN. It is noteworthy that the method takes the form of a “forward generation–reverse regression” procedure. The network attempts to recover an estimate of the envisioned object from the degraded image by prior mapping knowledge (e.g., the system transfer function). Physics-informed learning seamlessly incorporates both data and mathematical models to address the under-determined problem, even in noisy and high-dimensional contexts. Nine non-overlapping intensity data were fed into the network simultaneously to derive a high-resolution super-resolved image with high signal-to-noise. We hypothesize that the training process can be regarded as a prior learning



**Fig. 2.** Proposed network follows the form of a “forward generation–reverse regression” procedure. The proposed method is a nine-input, single-output supervised network.



process, including illumination angle, speckle noise, and acquisition position, which will further bolster the interpretability of the model.

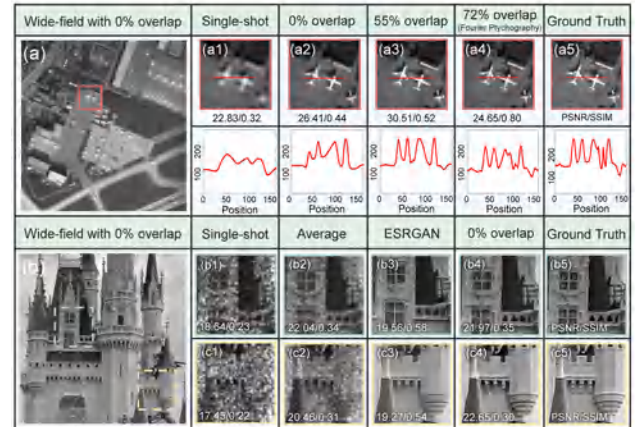
The proposed algorithm is flexible since once the training task is completed, no further manual adjustment of parameters is required to optimize the reconstruction performance. In order to yield reasonable predictions, the forward generation process and the reverse regression process are simultaneously constrained in the network model, and the dual loss functions compensate each other to produce the entire loss function balance. In the constructed network environment, the Adam optimizer is installed to implement structure feedback, with initial learning rate, batch size, and epoch setting to  $10^{-4}$ , 4, and 200, respectively. The model was operated on the configured computer with an Nvidia RTX2080Ti graphics card and an Intel Core™ i7-9700K CPU @ 3.60GHz×8 processor. The training and testing times for the entire network are 7.7 hours and 0.2 seconds, respectively. The network feeds back/refines the fitting errors between the data through the optimizer to avoid producing contrived results, thus, effectively retaining latent texture information. The mathematical expression of the loss function is presented as

$$Loss = \sum_{i=1}^N Loss_1 [F(x_i), y_i] + \lambda Loss_2 [D(y_i), x_i], \quad (3)$$

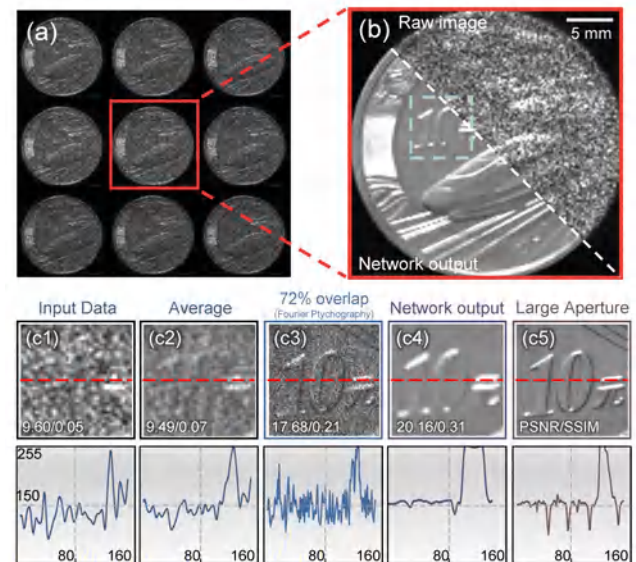
where  $x_i$  and  $y_i$  represent the input LR and output HR images, respectively;  $Loss_1 [F(x_i), y_i]$  and  $Loss_2 [D(y_i), x_i]$  describe the loss functions of forward regression and inverse regression tasks, respectively. The super-resolved image  $F(x_i)$  is constantly approaching the similarity with its corresponding HR image in the training process. Simply put, the similarity between the predicted map  $D(y_i)$  and the forward-fed map is continuously approached during the regression process. Hinting that the forward loss value is preferred, hereby, we set the weight distribution  $\lambda$  of the hybrid loss function to 0.1.

We evaluated the proposed method on both synthetic and real-world datasets. To test the effectiveness of the proposed method, we first reconstructed the low-resolution scene without coherent speckles (dataset was created by DIV2K [22]). To establish the unique advantages of the proposed method over traditional ptychography imaging, we conduct quantitative analyses in terms of input images having different overlaps, as illustrated in Figs. 3(a1)–3(a4). The bottom row of Figs. 3(a1)–3(a5) presents the line profile along the red dotted line. As one would expect, the robustness of the network is boosted, and more texture components of the image are reproduced with the increasing amount of data fed into the network. We also perform the reconstruction of diffuse reflective objects (rough paper), as shown in Fig. 3(b), and the corresponding zoomed-in areas are shown in Figs. 3(b1)–3(b5) and Figs. 3(c1)–3(c5). It is noted that the proposed network results still defeat the other network method [23], such as Generative Adversarial Network, in terms of the maximum improvement of 3.38 dB in peak signal-to-noise ratio (PSNR). The learning-based single-shot synthetic aperture imaging (LSS-SAI) approach supports significant improvement (extreme reduction from 50 minutes to 0.2 seconds) in imaging speed with a negligible decline in reconstruction quality against the related methods.

Furthermore, we selected a coin made of metal alloy with a diameter of 27 mm as the object, as shown in Fig. 4. Figures 4(c1)–4(c5) demonstrate the different reconstruction results



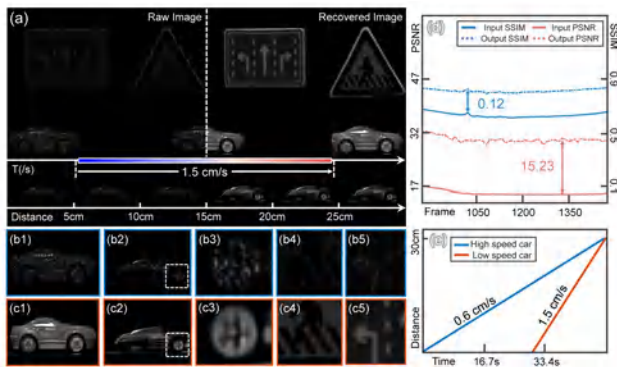
**Fig. 3.** Comparison of network reconstruction results. (a) Simulation reconstruction results of specular objects. (b) Performance of the LSS-SAI platform for the smooth object. (a1)–(a5), (b1)–(b5), (c1)–(c5) Corresponding region reconstruction results.



**Fig. 4.** Comparison of reconstruction results in the case of diffuse reflection. (a) The original images fed into the network. (b) Predicted reconstruction results of the network. (c1)–(c5) Results of image quality evaluations for the region of interest in different cases/methods (the reference image with an aperture F-number of 4 is selected as the label).

in the magnified region of interest for the commemorative coin. The gray scale profile of the magnified region is plotted below the corresponding one, for which the smooth profile indicates lower scattering noise. Figure 4(c1) illustrates that the signal-to-noise ratio and the resolution of the coin are too inferior to distinguish the features. As shown in Fig. 4(c2), although the noise is partially suppressed as a result of taking the cumulative average of the nine sub-aperture maps, the detailed components of the images are not yet reproducible. Theoretically, with enough low-resolution images, this method is able to increase the resolution of the image by a factor of two. Although the high-frequency components of the image are a super-resolved reconstruction, there is still significant speckle





**Fig. 5.** Constructed vehicle dynamic pursuit imaging results. (a) Comparison of the recorded low-resolution image (under F-number 12) with the predicted super-resolved image. (b1)–(b5) Magnification of the regions in the low-resolution images. (c1)–(c5) Super-resolution reconstruction results of the corresponding regions. (d) Comparison of the PSNR and SSIM curves of the original image and the reconstructed image in the dynamic experiment. (e) The corresponding curves of displacement distance versus time for the two vehicles.

noise in Fig. 4(c3). Moreover, the same trade-off between massive data volume (temporal resolution) and spatial resolution has to be compromised. On the contrary, the reconstruction result of the proposed method is presented in Fig. 4(c4), improving 10.56 and 0.26 in both PSNR and structural similarity (SSIM) indexes, respectively. The proposed method demonstrates efficient noise suppression capabilities while improving image resolution, elegantly solving the problems of conventional methods.

Moreover, we demonstrate the high temporal resolution of LSS-SAI by performing super-resolved videography of a dynamic scenario containing two isolated samples (See Visualization 1 for the whole video recording). The established system is depicted in Fig. 5, which employs a linear displacement stage to push the movement of both toy vehicles separately and places a stationary sign at the rear. Figure 5(b) shows that the reconstructed result from raw low-resolution data is too coarse to identify the model tire and steering sign details, which is capped by the combined limitation of aperture diffraction and laser speckle noise. In contrast, the proposed LSS-SAI yielded the best super-resolved reconstruction with well-reproduced surface details, as shown in Fig. 5(c), which is almost reproduced to the ground-truth data. The reconstruction method improves the PSNR and SSIM metrics by up to 0.12 and 15.23, respectively, compared with the original image for 1500 consecutive frames. Furthermore, we also performed the corresponding linear fits for the two moving targets, which can be inversely calculated as 0.6 cm/s and 1.5 cm/s for the two vehicles, respectively. Experiments show that the proposed algorithm is a powerful approach for improving the performance of Fourier ptychography even if containing complex speckle noise.

In this Letter, we have presented a learning-based single-shot synthetic aperture imaging, endowing the capability to overcome the reconstruction quality deterioration and stringent overlapping ratio constraints in conventional FP. Moreover, thanks to its single-shot nature, LSS-SAI is fundamentally immune to arti-

fact induced by object motion. The proposed method has great potential for performing super-resolution imaging of macroscopic diffuse reflectance observations. More modifications and innovations remain to be implemented in further, e.g., whether it is promising to reconstruct the phase information of far-field diffuse scattering objects.

**Funding.** National Natural Science Foundation of China (U21B2033, 61905115, 62105151, 62175109); Leading Technology of Jiangsu Basic Research Plan (BK20192003); National Major Scientific Instrument Development Project (62227818); Space Optoelectronic Measurement and Perception Lab (No.LabSOMP-2022-05); Youth Foundation of Jiangsu Province (BK20190445, BK20210338).

**Disclosures.** The authors declare no conflicts of interest.

**Data availability.** Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

## REFERENCES

1. T. S. Ralston, D. L. Marks, P. S. Carney, and S. A. Boppart, *Nat. Phys.* **3**, 129 (2007).
2. Q. Chu, Y. Shen, M. Yuan, and M. Gong, *Opt. Commun.* **405**, 288 (2017).
3. G. Zheng, R. Horstmeyer, and C. Yang, *Nat. Photonics* **7**, 739 (2013).
4. A. Moreira, P. Prats-Iraola, M. Younis, G. Krieger, I. Hajnsek, and K. P. Papathanassiou, *IEEE Geosci. Remote Sens. Mag.* **1**, 6 (2013).
5. L.-H. Yeh, J. Dong, J. Zhong, L. Tian, M. Chen, G. Tang, M. Soltanolkotabi, and L. Waller, *Opt. Express* **23**, 33214 (2015).
6. P. Song, S. Jiang, T. Wang, C. Guo, R. Wang, T. Zhang, and G. Zheng, *Photonics Res.* **10**, 1624 (2022).
7. G. Zheng, C. Shen, S. Jiang, P. Song, and C. Yang, *Nat. Rev. Phys.* **3**, 207 (2021).
8. L. Tian, Z. Liu, L.-H. Yeh, M. Chen, J. Zhong, and L. Waller, *Optica* **2**, 904 (2015).
9. J. Holloway, Y. Wu, M. K. Sharma, O. Cossairt, and A. Veeraraghavan, *Sci. Adv.* **3**, e1602564 (2017).
10. S. Dong, R. Horstmeyer, R. Shiradkar, K. Guo, X. Ou, Z. Bian, H. Xin, and G. Zheng, *Opt. Express* **22**, 13586 (2014).
11. J. Sun, Q. Chen, Y. Zhang, and C. Zuo, *Opt. Express* **24**, 15765 (2016).
12. Y. Shu, J. Sun, J. Lyu, Y. Fan, N. Zhou, R. Ye, G. Zheng, Q. Chen, and C. Zuo, *Photonix* **3**, 24 (2022).
13. J. Sun, Q. Chen, Y. Zhang, and C. Zuo, *Biomed. Opt. Express* **7**, 1336 (2016).
14. C. Zuo, J. Qian, S. Feng, W. Yin, Y. Li, P. Fan, J. Han, K. Qian, and Q. Chen, *Light: Sci. Appl.* **11**, 39 (2022).
15. S. Jiang, K. Guo, J. Liao, and G. Zheng, *Biomed. Opt. Express* **9**, 3306 (2018).
16. L. Bominathan, M. Maniparambil, H. Gupta, R. Baburajan, and K. Mitra, "Phase retrieval for fourier ptychography under varying amount of measurements," arXiv:1805.03593 (2018).
17. C. Wang, M. Hu, Y. Takashima, T. J. Schulz, and D. J. Brady, *Opt. Express* **30**, 2585 (2022).
18. J. Tang, K. Wang, Z. Ren, W. Zhang, X. Wu, J. Di, G. Liu, and J. Zhao, *Opt. Lasers Eng.* **139**, 106463 (2021).
19. K. Wang, J. Dou, Q. Kema, J. Di, and J. Zhao, *Opt. Lett.* **44**, 4765 (2019).
20. B. Wang, Y. Zou, L. Zhang, Y. Li, Q. Chen, and C. Zuo, *Opt. Lasers Eng.* **156**, 107078 (2022).
21. K. Wang, M. Zhang, J. Tang, L. Wang, L. Hu, X. Wu, W. Li, J. Di, G. Liu, and J. Zhao, *Photonix* **2**, 8 (2021).
22. E. Agustsson and R. Timofte, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2017).
23. X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (2018).



# Model-based deep learning for fiber bundle infrared image restoration

Bo-wen Wang<sup>a, b, 1</sup>, Le Li<sup>a, b, 1</sup>, Hai-bo Yang<sup>c</sup>, Jia-xin Chen<sup>c</sup>, Yu-hai Li<sup>c</sup>, Qian Chen<sup>a, b</sup>,  
Chao Zuo<sup>a, b, \*</sup>

<sup>a</sup> Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense, Nanjing University of Science and Technology, Nanjing, Jiangsu Province, 210094, China

<sup>b</sup> Smart Computational Imaging (SCI) Laboratory, Nanjing University of Science and Technology, Nanjing, Jiangsu Province, 210094, China

<sup>c</sup> Science and Technology on Electro-Optical Information Security Control Lab, Tianjin, 300308, China

## ARTICLE INFO

### Article history:

Received 10 June 2022

Received in revised form

19 October 2022

Accepted 7 December 2022

Available online 20 December 2022

### Keywords:

Fiber bundle

Deep learning

Infrared imaging

Image restoration

## ABSTRACT

As the representative of flexibility in optical imaging media, in recent years, fiber bundles have emerged as a promising architecture in the development of compact visual systems. Dedicated to tackling the problems of universal honeycomb artifacts and low signal-to-noise ratio (SNR) imaging in fiber bundles, the iterative super-resolution reconstruction network based on a physical model is proposed. Under the constraint of solving the two subproblems of data fidelity and prior regularization term alternately, the network can efficiently “regenerate” the lost spatial resolution with deep learning. By building and calibrating a dual-path imaging system, the real-world dataset where paired low-resolution (LR) - high-resolution (HR) images on the same scene can be generated simultaneously. Numerical results on both the United States Air Force (USAF) resolution target and complex target objects demonstrate that the algorithm can restore high-contrast images without pixilated noise. On the basis of super-resolution reconstruction, compound eye image composition based on fiber bundle is also embedded in this paper for the actual imaging requirements. The proposed work is the first to apply a physical model-based deep learning network to fiber bundle imaging in the infrared band, effectively promoting the engineering application of thermal radiation detection.

© 2022 China Ordnance Society. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## 1. Introduction

The fiber bundle imaging technique has demonstrated its success in military periscope, life detection, and endoscopic imaging [1], owing to the inherent flexibility of fiber optics. In the military field, the fiber bundle periscope combines the flexible passive fiber optic image transmission system with the internal sighting scope, enabling the shooter to conveniently utilize corners, tree trunks, trenches, high platforms, and other terrain features for periscopic hidden observation and rapid aiming and firing on targets. By constructing a compact set of lenses, multiple LR sub-images are formed in the compound eye system [2,3], and composited sub-eye images are achieved by post-processing. In addition, combining the

advantages of infrared imaging [4] and fiber-optic sensing can provide more valuable technical approaches to the field of detection and open up a variety of applications. Meanwhile, in turn, limited by its geometric nature (irregular layouts of fiber cores), images taken by such systems have penetrated honeycomb-like fixed pattern noises [5,6]. Infrared images are accompanied by a significant non-uniformity effort, which damages the imaging performance rather than helping it [7,8]. In addition, the effective imaging resolution of a fiber bundle system is capped by the inherent physical fiber core diameter and fiber density rather than the optical system or the detector pixel size. Image restoration possibility is also considered the potential of the compound-eye system, which is yet to be explored fully. Therefore, there is an urgent requirement for an effective algorithm to improve the spatial imaging resolution while separating the honeycomb patterns, which is the motivation of this paper.

Most current imaging applications of fiber bundles are applied in micro-endoscopic, which are typically accompanied by poor light transmission (incoming light intensity information) due to the minor incident numerical aperture of the fiber bundle. Therefore,

\* Corresponding author. Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense, Nanjing University of Science and Technology, Nanjing, Jiangsu Province, 210094, China

E-mail address: [zuochao@njust.edu.cn](mailto:zuochao@njust.edu.cn) (C. Zuo).

Peer review under responsibility of China Ordnance Society

<sup>1</sup> The authors contributed equally to this work.

the low SNR in far-field imaging is also a huge challenge, which progressively hinders the accurate analysis of objects. In order to optimize the imaging perception of the optical fiber bundle technique, it is necessary to explore implementable methods to eliminate the impact of honeycomb patterns and compensate for the mixed texture information. Corresponding optimization methods have been presented successively. Earlier computational methods were generally implemented a priori by the regularity of fiber pattern arrangement. Rupp [9] borrowed the interpolation method in the spatial domain to establish the exact position relationship of each fiber center pixel point and subsequently interpolates the cladding pixels based on the values of the neighboring pixel points. Shinde [10] proposed that the honeycomb structure image is regularly arranged in the frequency domain, and the envisioned data can be recovered using the band-stop filtering technique, yet it is hugely challenging to determine the threshold of the band-stop filter, which normally results in part of the valuable information being filtered as well. However, previous computational methods only eliminated undesired pixelated patterns without substantial improvement in spatial resolution.

Super-resolution is an ill-posed problem [11–14] that deals with restoring an HR image from a single or a series of raw images based on either specific a priori knowledge or just an assumed generic notion about the tighter correspondence imaging model. The deep learning technique [15–17] breaks the dependence of traditional methods on prior knowledge and efficiently utilizes the raw information “hidden” in the original honeycomb patterns. Minimizing the optimization problem by mapping massive data samples [18,19] (deep learning methods gradually reduce the loss function through multiple epochs and update the weight parameters through feedback), is conducive to precisely learning the high-resolution image. In particular, U-Net [20] has achieved tremendous success in searching the mapping functions (observation models and noise statistics) for various underdetermined medical imaging problems, verifying the feasibility of constraining the inverse problem through the network. Ravi [21] estimated the pseudo ground truth image by a video alignment algorithm and then tried to recover the fiber bundle image by three different convolutional neural networks. Simultaneously, Shao [22] implemented a generative adversarial restoration neural network (GARNN) or a 3D convolution network to remove the foveal effect and restore the “hidden” features. The feasibility of super-resolution reconstruction of infrared images through the network was also verified in previous work [23–25]. Perhaps not surprisingly, conventional deep-learning algorithms lack interpretability to some extent and heavily rely on abundant examples to train the network without incorporating any physical degradation model constraints [26–28]. Each model training can only focus on a single situation-specific image reconstruction project and lacks the flexibility to cope with different tasks or different scale factors. To address the above issues, we propose an image-resolved algorithm based on the physical-model deep learning network, which is promising in regenerating high-resolution and non-honeycomb pattern images. In this research work, the primary options focused on the infrared radiation band, yet actual visible light images could also be employed. Based on this research work, it seems reasonable to pursue a similar set of fiber-optic bundle system configurations in future military research.

The remaining structures of this paper are as follows. Section 2 depicts the basic principle of our proposed method and presents the details of the proposed network for infrared fiber bundle super-resolution. Abundant experimental results and analysis are demonstrated in Section 3. Finally, Section 4 enforces a discussion and summarizes the paper.

## 2. Methods

### 2.1. Theoretical analysis

The optical fiber image bundle is composed of multiple fibers, and each core conveys an individual image element. The current fiber bundle has a trade-off between the field of view and the core sampling rate. In order to combine the field of view and luminous flux of a single fiber, the diameter of currently manufactured fibers is empirically chosen to be slightly larger. The inherent fiber bundle physical limitation interrupts the information and weakens the imaging resolution, leading to a honeycomb pixelation (fixed pattern noise) of the sample.

In order to quantitatively investigate the correspondence between image quality and parameters of the optical fiber bundle imaging system, the concept of the average modulation transfer function (MTF) [29–31], commonly employed in the discrete sampling system, is introduced as the evaluation metric. The average MTF of the system can be indicated by the product of the Fourier transforms of each discrete sampling distribution function, which is expressed as follows:

$$MTF_{\text{sys}} = MTF_{\text{object}} \cdot MTF_{\text{fib}} \cdot MTF_{\text{relay}} \cdot MTF_{\text{detector}} \quad (1)$$

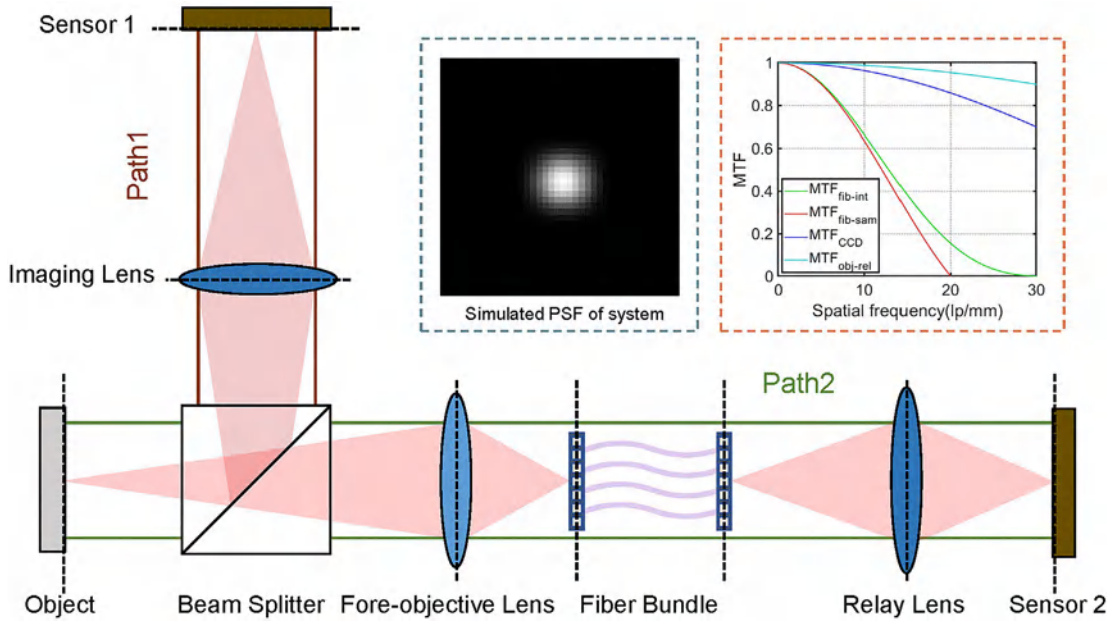
where  $MTF_{\text{sys}}$ ,  $MTF_{\text{object}}$ ,  $MTF_{\text{fib}}$ ,  $MTF_{\text{relay}}$ , and  $MTF_{\text{detector}}$  represent the MTFs of the overall system, the fore-objective lens, the optical fiber imaging bundle, the relay lens, and the detector, respectively. The dual sampling of both fiber bundle and detector exists in the proposed system, and the imaging process of the fiber bundle includes low-pass filtering followed by the integral sampling of the fiber core and the discrete decimation of each fiber in the dense arrangement.  $MTF_{\text{fib}}$  is composed of the fiber integral function  $MTF_{\text{fib-int}}$  and the fiber sampling function  $MTF_{\text{fib-sam}}$ . Therefore,  $MTF_{\text{sys}}$  can be expressed as

$$\begin{aligned} MTF_{\text{sys}} &= MTF_{\text{object}} \cdot MTF_{\text{fib-int}} \cdot MTF_{\text{fib-sam}} \cdot MTF_{\text{relay}} \cdot MTF_{\text{CCD}} \\ &= MTF_{\text{object}} \cdot MTF_{\text{rel}} \cdot \left[ \frac{2J_1(\pi df)}{\pi df} \right]^2 \cdot |\text{sinc}(\Delta f)| \cdot |\text{sinc}(pf)| \end{aligned} \quad (2)$$

Among them,  $J_1$  refers to the first-order Bessel function,  $d$  is the diameter of the fiber core,  $f$  is the spatial frequency,  $\Delta$  represents the fiber core center distance, and  $p$  represents the detector pixel size.

Our system adopts the chalcogenide glass infrared fiber bundle with a core diameter of 40  $\mu\text{m}$  and a core center distance of 50  $\mu\text{m}$ . Furthermore, infrared sensors with a pixel size of 15  $\mu\text{m}$  and a center wavelength of 4.2  $\mu\text{m}$  are chosen for data collection. As shown in Fig. 1, under the premise of this invariant optical system, the system MTF mainly relies on the fiber bundle MTF. In this case, the cutoff frequency and spatial resolution of the system are not limited by the pixel size of the sensor or the diffraction effect. Instead, they are dramatically trapped by the fiber core center distance and core diameter ( $MTF_{\text{fib}}$ ). Due to the dual discrete sampling mechanism of the optical fiber bundle imaging system, honeycomb-like fixed patterns are imposed on its output images. In the optical imaging system, the multiplication of MTF in the frequency domain is equivalent to the convolution of the point spread function (PSF) in the spatial domain. The PSF (blur kernel) of the system is also a crucial indicator of imaging resolution capability. In light of the above analysis, the system degradation model is mainly affected by the fiber bundle, and the simulated blur kernel of the fiber bundle imaging system with the hexagonal structure is shown in Fig. 1. The proposed method is also dedicated to solving the





**Fig. 1.** Schematic diagram of the dual-path imaging system. The blue dashed line illustrates the PSF of the simulated system, and the red dashed line represents the MTF of the simulated system.

dilemma of incompatibility between sensor and fiber bundle sampling in fiber optic imaging systems.

The “one-to-one” dual-path imaging system is established to obtain HR images and degraded LR images simultaneously, as shown in Fig. 1. To obtain the true observation image (label) and fiber image (input image) simultaneously, we introduced a beam splitter to split the light from the object into two paths. By physically adjusting the distance between the lens and the object, the image alignment error of the two imaging paths is at the sub-pixel level. Further, the post-alignment algorithm accomplishes the tiny alignment of the dual-channel images.

### 2.2. Proposed algorithm and network architecture

In this section, we investigate how to adapt the deep-learning method for honeycomb artifacts removal and far-field image restoration. As such, we resort to theoretical analysis and formulate the image restoration problem in a framework. Generally, image super-resolution is an inverse problem [32] where the objective is to recover the latent HR image  $\mathbf{x}$  from its blurred, decimated, and noisy observation  $\mathbf{y} = \mathbf{S}\mathbf{H}\mathbf{x} + \mathbf{n}$ , where  $\mathbf{S}$  denotes the standard down-sampler,  $\mathbf{H}$  represents the convolution operation with blur kernel, and  $\mathbf{n}$  is the additive noise. The model-based deep learning method with degradation constraint is interpretable compared with conventional deep learning methods. According to the maximum a posteriori (MAP) framework, the HR image can be performed by solving the following optimization problem:

$$\hat{\mathbf{x}} = \operatorname{argmin}_{\mathbf{x}} \frac{1}{2} \|\mathbf{y} - \mathbf{S}\mathbf{H}\mathbf{x}\|^2 + \lambda \Phi(\mathbf{x}) \quad (3)$$

where  $\frac{1}{2} \|\mathbf{y} - \mathbf{S}\mathbf{H}\mathbf{x}\|^2$  is the data fidelity term related to the model likelihood,  $\Phi(\mathbf{x})$  is the prior regularization term associated with the image prior information, and the role of the  $\lambda$  is to weigh the importance of the prior regularization term relative to the data fidelity term. For the purpose of acquiring the further unfolding inference, the half-quadratic splitting (HQS) algorithm introduces an auxiliary variable  $\mathbf{z}$  and transforms Eq. (3) into an approximate equivalent problem

$$L_{\mu}(\mathbf{x}, \mathbf{z}) = \frac{1}{2} \|\mathbf{y} - \mathbf{S}\mathbf{H}\mathbf{z}\|^2 + \lambda \Phi(\mathbf{x}) + \frac{\mu}{2} \|\mathbf{z} - \mathbf{x}\|^2 \quad (4)$$

where  $\mu$  is a regularization parameter associated with the quadratic penalty term, and such a problem can be solved via the following iterative scheme:

$$\mathbf{z}_k = \operatorname{argmin}_{\mathbf{z}} \|\mathbf{y} - \mathbf{S}\mathbf{H}\mathbf{z}\|^2 + \mu_k \|\mathbf{z} - \mathbf{x}_{k-1}\|^2 \quad (5)$$

$$\mathbf{x}_k = \operatorname{argmin}_{\mathbf{x}} \frac{\mu_k}{2} \|\mathbf{z}_k - \mathbf{x}\|^2 + \lambda \Phi(\mathbf{x}) \quad (6)$$

Mathematically, the data fidelity term and prior regularization term are decoupled into two individual subproblems, which can facilitate each other to realize blur elimination and detail recovery. Therefore, the model-based iterative network (MBIN) can be designed, whose framework consists of a data fidelity module and a prior regularization module iteratively. Fig. 2 illustrates the overall architecture of MBIN with  $k$  iterations.

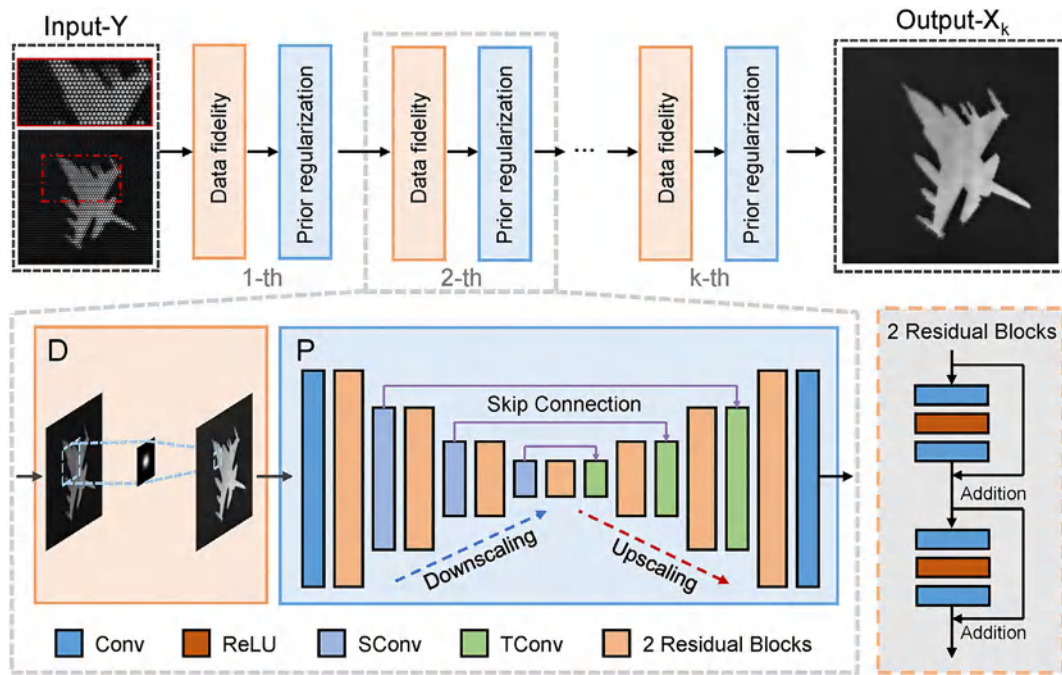
The data fidelity module imposes a degradation model constraint on the solution, which can be leveraged and incorporated into the network construct to guarantee more precise and reliable reconstruction. Specifically, the data fidelity module is related to a quadratic regularized least-squares problem which has various solutions for different degradation kernels. A direct solution is given by

$$\mathbf{z}_k = \left( \mathbf{H}^T \mathbf{S}^T \mathbf{S} \mathbf{H} + \mu_k \mathbf{I} \right)^{-1} \left( \mathbf{H}^T \mathbf{S}^T \mathbf{y} + \mu_k \mathbf{x}_{k-1} \right) \quad (7)$$

We assume that the convolution is performed under circular boundary conditions. Hence, the fast Fourier transform can be adapted to efficiently implement Eq. (7).

For the prior regularization module, Eq. (6) can be reformulated as

$$\mathbf{x}_k = \operatorname{argmin}_{\mathbf{x}} \frac{1}{2 \left( \sqrt{\lambda/\mu_k} \right)^2} \|\mathbf{z}_k - \mathbf{x}\|^2 + \Phi(\mathbf{x}) \quad (8)$$



**Fig. 2.** The overall architecture of MBIN with iterations  $k = 8$ . MBIN consists of two main iterative modules, the data fidelity module to guarantee the solution complies with the degradation process and the prior regularization module to enforce desired properties of the output.

Treating  $z_k$  as the “blurred” image, Eq. (8) minimizes the residue between  $z_k$  and the “clear” image  $x$  using the prior  $\Phi(x)$  [33]. The corresponding minimization function is also named the so-called loss function. The latent mapping between “blurred” and “clear” maps can be learned by training the prior module, which acts as a detail enhancer for high-frequency recovery.

Inspired by the prior knowledge in information optics, we may impute the expansion of high-frequency components to the image prior rooted in the elaborately designed architectures. Physics-informed learning seamlessly incorporates both data and mathematical models to address the under-determined problem, even in noisy and high-dimensional contexts. In fact, the physical prior was integrated into the forward generation process in earlier investigations. The optimized values of the physical prior are derived by gradient descent applying a back-propagation algorithm of derivatives, which is akin to the deep learning optimization process.

The process of cross introducing the optimization iterations of the physical model together with the fitting function of deep learning will significantly strengthen the interpretability of the network. Notably, in Fig. 2, we introduce the physical iterative process and further incorporate the system’s PSF into the network model. Quite the contrary, if only end-to-end learning is done without any physical model intervention, the image quality will undoubtedly suffer degradation. In previous super-resolved reconstruction projects, numerous efforts neglected the introduction of physical models as the most critical aspect.

U-Net, widely formed in multi-scale image-to-image transforms, is adopted to construct this module. As illustrated in Fig. 2, the fundamental structure of the prior regular module involves a contractive branch and an expansive branch with four folds. Moreover, the module takes advantage of Resnet [34] to enhance network capacity and performance by introducing residuals. A set of two residual blocks are integrated on each scale of the branch [35], as shown in Fig. 2. Specifically, the number of channels from the first scale layer to the end layer is set to 64, 128, 256, and 512, respectively, in that order. For the down-sampling and up-sampling

operation in the contractive and expansive branch, we adopt  $2 \times 2$  stridden convolution (SConv) and  $2 \times 2$  transposed convolution (TConv), respectively, which are not followed by the activation function. In addition, skip connection can not only transfer image feature information but also alleviate the problem of gradient disappearance, enabling convenient transmission of valuable information in the network. In respect of the loss function, we adopt the  $L_2$  loss to evaluate the peak signal-to-noise ratio (PSNR) performance. In the network, the batch size is set to 48. Adam solver [36] is adopted to optimize the parameters with the learning rate initialized as  $10^{-4}$ . The hardware platform of the network for model training is Intel Core™ i7-9700 K CPU 3.60 GHz equipped with the RTX2080Ti graphics card, and the software platform is PyTorch 1.6.0 under the Ubuntu 16.04 operating system.

### 3. Experimental results and analysis

#### 3.1. Dataset preparation

For the validation of our proposed method, the dataset is prepared by simulations, which dispenses with the available extra imaging system to provide well-registered pairs of fiber bundle images and their corresponding ground truth data. To obtain the LR counterpart from each mapped HR image, we impose the PSF for each hexagonal arranged core to the mapped HR image. The convolution operation of PSF, followed by the down-sampling operation, implements a weighted sum of HR pixels to yield an LR image pixel.

In recent years, deep learning is emerging as a powerful tool to address problems by learning from data, largely driven by the availability of massive datasets. Unfortunately, such simple degradation models could not faithfully describe the complex degradation processes in the real world. This motivates us to build a real-world SR dataset to narrow the synthetic-to-real gap. Our training data set consists of 1000 paired LR-HR images and their corresponding ground truth data. To monitor the accuracy of the

neural networks on data never seen before, we created a validation set by setting apart 50 images from the original training data. A representative dataset in our proposed network is depicted in Fig. 3. Note that our proposed network is still based on supervised learning. We consider the training process of the network as the task of learning the image prior, including fiber fixed noise, luminance bias, frequency characteristics, etc. The network attempts to recover an estimate of the envisioned sample from the degraded image by prior mapping knowledge (e.g., the system transfer function). Therefore, a sufficient variety of images should be included in our datasets to construct the network mapping function as much as possible. The reconstruction process of the network is also not a magic trick, which requires a sufficient amount of prior information to support the fitting process of the parameters. Only when the dataset is guaranteed to a certain extent the nonlinear mapping will perform well in a massive sample.

For the data collected from the dual-path imaging system, HR images captured from Sensor 1 have the same pixel dimensions of  $1024 \times 1024$ , while LR images recorded by Sensor 2 are cropped to  $256 \times 256$  pixels from  $640 \times 480$  pixels. The end-to-end networks demand that LR images are interpolated to the same size as HR images, and image pairs are aligned by sequentially applying coarse and fine registrations for the network to learn the direct mapping relations. When using experimental data as ground truth, the reconstructed performance is inevitably contaminated by noise. Geometric changes such as rotation, translation, and deformation are implemented in images by finding inter-image feature points to eliminate the distortion errors between different lenses. When the distortion error is eliminated, we consider that coarse registration

can be regarded as finished, and only the displacement between pixel levels remains for the registration error with respect to two images. Consequently, the frequency domain cross-correlation method [37] is adopted for precise registration to achieve sub-pixel error correction between image pairs. Considering the invariant characteristics of the imaging system, we could complete the registration of all datasets by performing only one-time coarse and fine registration, as shown in Fig. 4.

### 3.2. Quantitative evaluation based on the USAF resolution target

A USAF resolution target was employed to quantitatively validate the performance of the proposed method in terms of image resolution. Depending on the system parameters mentioned in Subsection 2.1, the forward model of the system is simulated with explicit degradation to acquire LR-HR image pairs in Subsection 3.1. The cellular fiber image in Fig. 5(a) is generated from the original high-resolution infrared image, which is shown in Fig. 5(e), through the dual discrete sampling of the fiber bundle imaging system. Figs. 5(b)–5(d) presents the reconstructed results using U-Net, super-resolution convolutional network (SRCNN) [38], and MBIN methods, respectively. It is clearly observed from Fig. 5 that the MBIN method addresses nearly all the major limitations of the other methods.

Although all methods exhibit removal of the honeycomb patterns, images reconstructed by U-Net and SRCNN still survive with distinct pixelated patterns along the edge of bars, where the recovered edges are jagged rather than smooth. On the contrary, MBIN eliminates these artifacts along edges, maintaining



Fig. 3. A representative dataset in our proposed network.

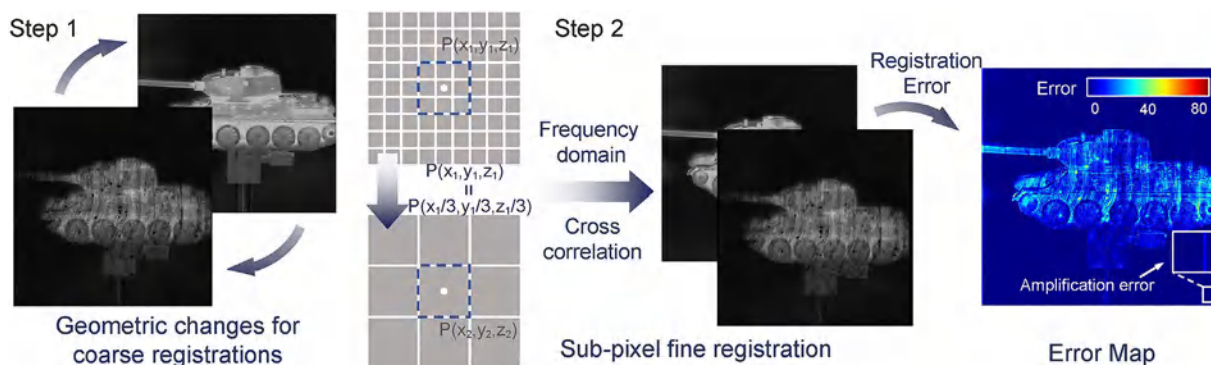
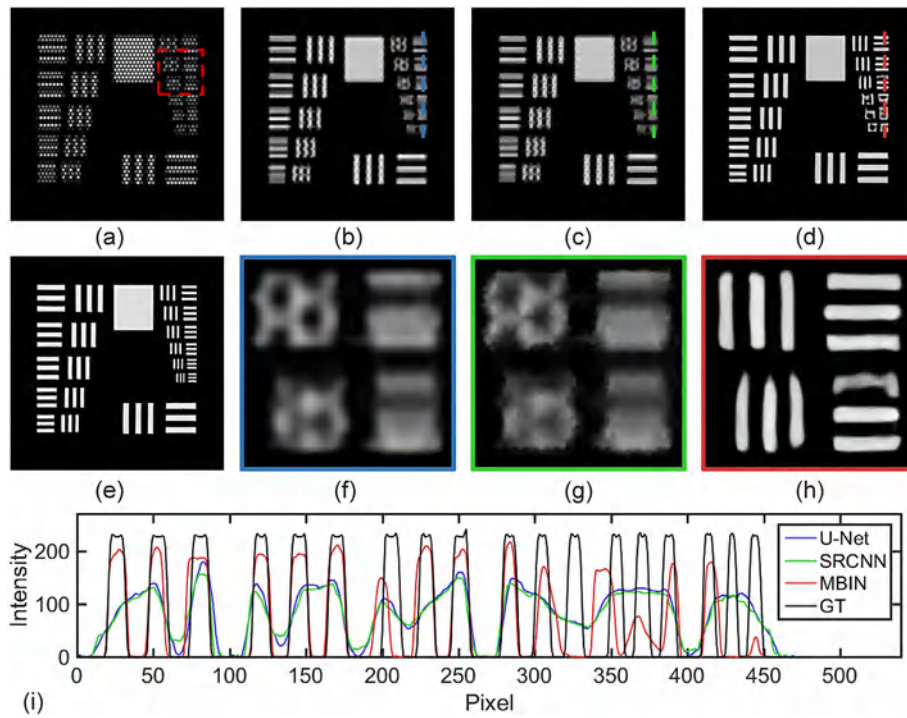


Fig. 4. Schematic diagram of the dual-path image alignment processing.

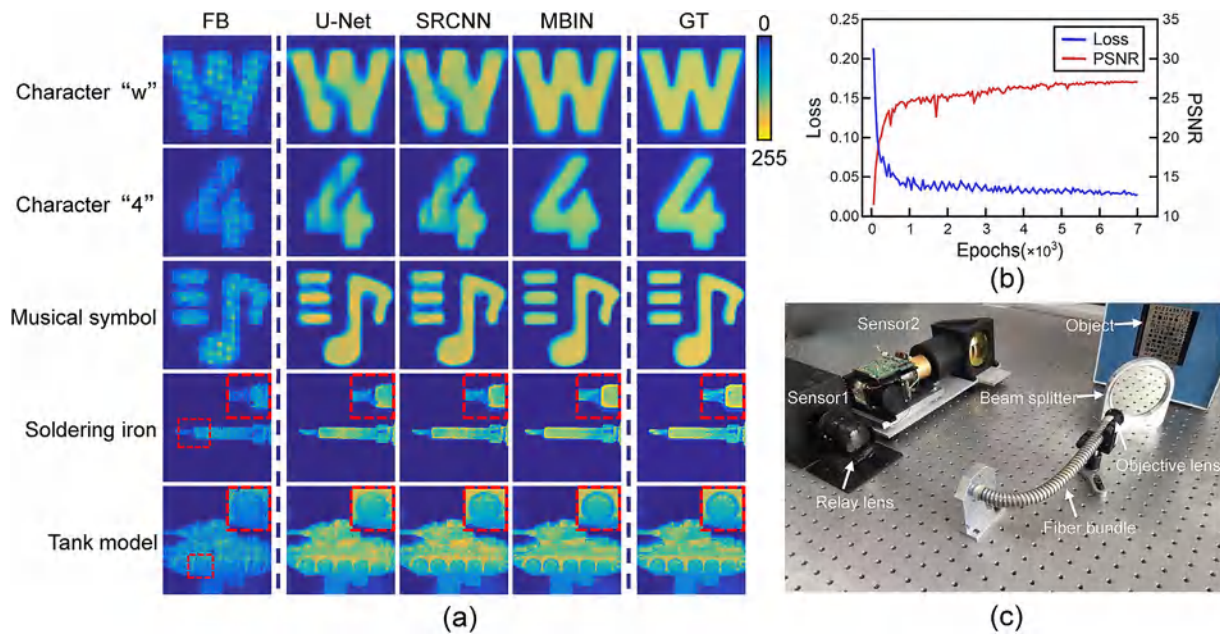




**Fig. 5.** Comparison of U-Net, SRCNN, and MBIN on the synthetic USAF resolution target: (a) Input USAF target image; (b)–(d) Reconstructed images respectively used U-Net, SRCNN and MBIN methods; (e) Original USAF target image (ground truth); (f)–(h) Zoomed-in images of the region of interest (ROI) in (b)–(d); (i) Cross-sectional profile of dash line shown in (b)–(d).

uniformity in intensity closer to the original image. As shown in Fig. 5(a), the minimum line pairs that can be resolved in the synthetic image is the element of the third group on the left, with a corresponding line width of  $0.92 \mu\text{m}$ . To intuitively compare the performance of the three methods, we intercept regions of the rectangular box for comparison, as shown in Figs. 5(f)–5(h). Note that results reconstructed by U-Net and SRCNN are blurred, and line pairs cannot be distinguished in the zoomed-in area.

Obviously, high-resolution images can be effectively reconstructed using our proposed method, and more specifically, our method can enhance the original resolution to the third group on the right, corresponding to a line pair resolution of  $0.54 \mu\text{m}$ . The proposed method extends the imaging resolution to 1.7 folds, successfully breaking through the imaging resolution limited by the physical size of the original intrinsic system. Furthermore, we can be surprised to observe that the intensity profile of MBIN in Fig. 5(i)



**Fig. 6.** (a) Comparison of fiber bundle image reconstruction; (b) The average Loss value and average PSNR value of the validation dataset against the number of training epochs; (c) The dual-path imaging system.

achieves the best performance in both contrast and resolution. Evidentially, the proposed MBIN demonstrates apparent advantages in terms of linewidth and sharpness exhibited in the reconstructed image.

### 3.3. The network performance on real datasets

To further demonstrate that the proposed MBIN indeed improves the image performance, we conducted real scene data experiments with the dual-path imaging system. For the sake of efficient learning of the end-to-end mapping relations, except for the interpolation and registrations mentioned in subsection 3.1, the LR images are pre-processed with background noise reduction and infrared image enhancement before feeding into the network model. We compared the MBIN with two state-of-the-art restorative neural networks to verify the network’s performance capabilities. In Fig. 6(a), the patterns are partial images captured by the system oriented to the blackbody radiator and different hollow boards. The soldering iron and the tank model belong to self-heating objects directly recorded by the system. Obviously, these results indicate that all three methods could exhibit favorable performance in recovering the hidden information from the honeycomb patterns, whereas the high-frequency edges of the restored images obtained by U-Net and SRCNN are blurred. In contrast, sharper images reconstructed by MBIN distinguish finer details. We plot the average loss value and the average PSNR value for the

validation dataset against the training epoch number of MBIN in Fig. 6(b). Such two curves oscillate in the early epochs and converge stably after more training epochs.

In order to further quantitatively evaluate the results obtained by different methods, the PSNR value and the structural similarity (SSIM) value are calculated for each reconstructed image relative to the corresponding ground truth, as listed in Table 1. It is obviously desirable that MBIN has the highest performance values in all cases. Specifically, it shows superior performance in both SSIM and PSNR, which are, on average, 5% and 5 dB greater than other methods, respectively. Perhaps not surprisingly, the principle behind this result is that the proposed network removes the honeycomb patterns effectively, and the hidden details are restored extremely similar to their ground truth, with higher PSNR and SSIM values representing more satisfactory results.

In the practical engineering application, multiple sub-eyes of the single-aperture fiber bundle can be arranged in the form of a compound eye array to capture multiple images at the same time, achieving high resolution and large field of view simultaneously by stitching multiple adjacent images with overlapping regions. Based on the super-resolution reconstruction method in this paper, the  $2 \times 1$  compound eye array system with infrared fiber bundles is established. The system can be shifted horizontally to physically scan corresponding areas of the hollow board in front of the blackbody radiator continuously for acquiring  $2 \times 5$  sub-images. Following that, a large-field image of the target is obtained by stitching sub-images, as shown in Fig. 7, expanding the horizontal field of view angle from  $21.48^\circ$  to  $65.31^\circ$ .

**Table 1**  
Measurement of different reconstruction methods on PSNR/SSIM.

Object/Algorithms	U-Net PSNR/SSIM	SRCNN PSNR/SSIM	MBIN PSNR/SSIM
Character “w”	23.05/0.889	23.52/0.898	28.63/0.942
Character “4”	22.45/0.904	23.08/0.896	28.54/0.963
Musical symbol	25.81/0.921	25.58/0.904	29.34/0.968
Soldering iron	30.39/0.964	31.30/0.966	36.74/0.981
Tank model	28.18/0.880	28.75/0.896	32.88/0.965

### 4. Conclusions

We have established and investigated a computational compound-eye imaging system with super-resolution reconstruction. The real-world infrared fiber bundle images and their corresponding ground truth images are both generated by the development dual-way imaging system. Image registration and rectification algorithms are developed to progressively align the



**Fig. 7.** Stitching results based on the fiber bundle and corresponding sub-eye images.



image pairs. The constructed dataset can address the real-world image super-resolution problem with better performance. We also find that with the aid of introducing a physical model-based network, the solution can be incorporated to preclude some disturbing terms in the ill-posed inverse problem and possibly comply with the imaging model. As such, we gain insights into questions concerning image restoration problems. Finally, simply utilizing larger dimension imaging sensors and coordinating with multiple sub-eye images could, in principle, further push the imaging field of view and spatial resolution. The proposed MBIN method promises to enable new measurement opportunities for military defense detection and evolve our knowledge in the photoelectric detection field.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

### Acknowledgments

The authors would like to acknowledge the National Natural Science Foundation of China (Grant Nos. 61905115, 62105151, 62175109, U21B2033), Leading Technology of Jiangsu Basic Research Plan (Grant No. BK20192003), Youth Foundation of Jiangsu Province (Grant Nos. BK20190445, BK20210338), Fundamental Research Funds for the Central Universities (Grant No. 30920032101), and Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense (Grant No. JSGP202105) to provide fund for conducting experiments.

### References

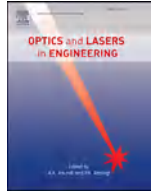
- Perperidis A, Dhaliwal K, McLaughlin S, Vercauteren T. Image computing for fibre-bundle endomicroscopy: a review. *Med Image Anal* 2020;62:101620. <https://doi.org/10.1016/j.media.2019.101620>.
- Chan W-S, Lam EY, Ng MK, Mak GY. Super-resolution reconstruction in a computational compound-eye imaging system. *Multidimens Syst Signal Process* 2007;18:83–101. <https://doi.org/10.1007/s11045-007-0022-3>.
- Ma M, Zhang Y, Deng H, Gao X, Gu L, Sun Q, et al. Super-resolution and super-robust single-pixel superposition compound eye. *Opt Laser Eng* 2021;146:106699. <https://doi.org/10.1016/j.optlaseng.2021.106699>.
- Wang Y, Dai S. Mid-infrared supercontinuum generation in chalcogenide glass fibers: a brief review. *Photonix* 2021;2:1–23.
- Reichenbach KL, Xu C. Numerical analysis of light propagation in image fibers or coherent fiber bundles. *Opt Express*, OE 2007;15:2151–65. <https://doi.org/10.1364/OE.15.002151>.
- Han J-H, Kang JU. Effect of multimodal coupling in imaging micro-endoscopic fiber bundle on optical coherence tomography. *Appl Phys B* 2012;106:635–43. <https://doi.org/10.1007/s00340-011-4847-y>.
- Sanghera JS, Aggarwal ID. Active and passive chalcogenide glass optical fibers for IR applications: a review. *J Non-Cryst Solids* 1999;256–257:6–16. [https://doi.org/10.1016/S0022-3093\(99\)00484-6](https://doi.org/10.1016/S0022-3093(99)00484-6).
- Qiu Z, Ma Y, Fan F, Huang J, Wu M, Mei X. A pixel-level local contrast measure for infrared small target detection. *Defence Technol* 2021. <https://doi.org/10.1016/j.dt.2021.07.002>.
- Rupp S, Winter C, Elter M. Evaluation of spatial interpolation strategies for the removal of comb-structure in fiber-optic images. In: 2009 annual international conference of the IEEE engineering in medicine and biology society; 2009. p. 3677–80. <https://doi.org/10.1109/IEMBS.2009.5334719>.
- Shinde A, Matham MV. Pixelate removal in an image fiber probe endoscope incorporating comb structure removal methods. *J Med Imag Health Informat* 2014;4:203–11. <https://doi.org/10.1166/jmih.2014.1255>.
- Kulkarni N, Nagesh P, Gowda R, Li B. Understanding compressive sensing and sparse representation-based super-resolution. *IEEE Trans Circ Syst Video Technol* 2012;22:778–89. <https://doi.org/10.1109/TCSVT.2011.2180773>.
- Takeda H, Milanfar P, Protter M, Elad M. Super-resolution without explicit subpixel motion estimation. *IEEE Trans Image Process* 2009;18. <https://doi.org/10.1109/TIP.2009.2023703>. 1958–75.
- Marcia RF, Willett RM. Compressive coded aperture superresolution image reconstruction. In: 2008 IEEE international conference on acoustics, speech and signal processing. Las Vegas, NV, USA: IEEE; 2008. p. 833–6. <https://doi.org/10.1109/ICASSP.2008.4517739>.
- Nguyen Nhat, Milanfar P, Golub G. A computationally efficient superresolution image reconstruction algorithm. *IEEE Trans Image Process* 2001;10:573–83. <https://doi.org/10.1109/83.913592>.
- Yang X, Xiang W, Zeng H, Zhang L. Real-world video super-resolution: a benchmark dataset and a decomposition based learning scheme n.d..vol. 10.
- Zuo C, Qian J, Feng S, Yin W, Li Y, Fan P, et al. Deep learning in optical metrology: a review. *Light Sci Appl* 2022;11:39. <https://doi.org/10.1038/s41377-022-00714-x>.
- Wang K, Zhang M, Tang J, Wang L, Hu L, Wu X, et al. Deep learning wavefront sensing and aberration correction in atmospheric turbulence. *Photonix* 2021;2:1–11.
- Hong-hai Y, Xiao-peng Y, Shao-kun L, Ping L, Xin-hong H. Radar emitter multi-label recognition based on residual network. *Defence Technol* 2022;18:410–7. <https://doi.org/10.1016/j.dt.2021.02.005>.
- Meng F, Li Y, Shao F, Yuan G, Dai J. Visual-simulation region proposal and generative adversarial network based ground military target recognition. *Defence Technol* 2021. <https://doi.org/10.1016/j.dt.2021.07.001>.
- Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical image computing and computer-assisted intervention – MICCAI 2015. Cham: Springer International Publishing; 2015. p. 234–41. [https://doi.org/10.1007/978-3-319-24574-4\\_28](https://doi.org/10.1007/978-3-319-24574-4_28). vol. 9351.
- Ravi D, Szczotka AB, Shakir DI, Pereira SP, Vercauteren T. Effective deep learning training for single-image super-resolution in endomicroscopy exploiting video-registration-based reconstruction. *Int J CARS* 2018;13:917–24. <https://doi.org/10.1007/s11548-018-1764-0>.
- Shao J, Zhang J, Huang X, Liang R, Barnard K. Fiber bundle image restoration using deep learning. *Opt Lett*, OL 2019;44:1080–3. <https://doi.org/10.1364/OL.44.001080>.
- Zou Y, Zhang L, Liu C, Wang B, Hu Y, Chen Q. Super-resolution reconstruction of infrared images based on a convolutional neural network with skip connections. *Opt Laser Eng* 2021;146:106717. <https://doi.org/10.1016/j.optlaseng.2021.106717>.
- Wang B, Zou Y, Zhang L, Li Y, Chen Q, Zuo C. Multimodal super-resolution reconstruction of infrared and visible images via deep learning. *Opt Laser Eng* 2022;156:107078. <https://doi.org/10.1016/j.optlaseng.2021.106717>.
- Wang B, Zou Y, Zhang L, Hu Y, Yan H, Zuo C, et al. Low-light-level image super-resolution reconstruction based on a multi-scale features extraction network. *Photonics*, vol. 8. Multidisciplinary Digital Publishing Institute; 2021. p. 321. <https://doi.org/10.3390/photonics8080321>.
- Perperidis A, Parker HE, Karam-Eldaly A, Altmann Y, Dhaliwal K, Thomson RR, et al. Characterization and modelling of inter-core coupling in coherent fiber bundles. *Opt Express* 2017;25:11932. <https://doi.org/10.1364/OE.25.011932>.
- Eldaly AK, Altmann Y, Perperidis A, Krstajic N, Choudhary TR, Dhaliwal K, et al. Deconvolution and restoration of optical endomicroscopy images. *IEEE Trans Comput Imaging* 2018;4:194–205. <https://doi.org/10.1109/TCI.2018.2811939>.
- Zhang K, Van Gool L, Timofte R. Deep unfolding network for image super-resolution. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). Seattle, WA, USA: IEEE; 2020. p. 3214–23. <https://doi.org/10.1109/CVPR42600.2020.00328>.
- Drougard R. Optical transfer properties of fiber bundles. *J Opt Soc Am* 1964;54:907. <https://doi.org/10.1364/JOSA.54.000907>.
- Wittenstein W, Fontanella JC, Newbery AR, Baars J. The definition of the OTF and the measurement of aliasing for sampled imaging systems. *Opt Acta: Int J Optics* 1982;29:41–50. <https://doi.org/10.1080/713820741>.
- de Luca L, Cardone G. Modulation transfer function cascade model for a sampled IR imaging system. *Appl Opt* 1991;30:1659. <https://doi.org/10.1364/AO.30.001659>.
- Park Sung Cheol, Park Min Kyu, Kang Moon Gi. Super-resolution image reconstruction: a technical overview. *IEEE Signal Process Mag* 2003;20:21–36. <https://doi.org/10.1109/MSP.2003.1203207>.
- Chan SH, Wang X, Elgendy OA. Plug-and-Play ADMM for image restoration: fixed-point convergence and applications. *IEEE Trans Comput Imaging* 2017;3:84–98. <https://doi.org/10.1109/TCI.2016.2629286>.
- He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR). Las Vegas, NV, USA: IEEE; 2016. p. 770–8. <https://doi.org/10.1109/CVPR.2016.90>.
- Lim B, Son S, Kim H, Nah S, Lee KM. Enhanced deep residual networks for single image super-resolution. In: 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW). Honolulu, HI, USA: IEEE; 2017. p. 1132–40. <https://doi.org/10.1109/CVPRW.2017.151>.
- Kingma DP, Ba J. Adam: a method for stochastic optimization. 2017.
- Guizar-Sicairos M, Thurman ST, Fienup JR. Efficient subpixel image registration algorithms. *Opt Lett* 2008;33:156. <https://doi.org/10.1364/OL.33.000156>.
- Dong C, Loy CC, He K, Tang X. Image super-resolution using deep convolutional networks. *IEEE Trans Pattern Anal Mach Intell* 2016;38:295–307. <https://doi.org/10.1109/TPAMI.2015.2439281>.





Contents lists available at ScienceDirect

## Optics and Lasers in Engineering

journal homepage: [www.elsevier.com/locate/optlaseng](http://www.elsevier.com/locate/optlaseng)

# Multimodal super-resolution reconstruction of infrared and visible images via deep learning

Bowen Wang<sup>a,b</sup>, Yan Zou<sup>c</sup>, Linfei Zhang<sup>a,b</sup>, Yuhai Li<sup>d</sup>, Qian Chen<sup>a,b</sup>, Chao Zuo<sup>a,b,\*</sup><sup>a</sup> Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China<sup>b</sup> Smart Computational Imaging (SCI) Laboratory, Nanjing University of Science and Technology, Nanjing, Jiangsu Province 210094, China<sup>c</sup> Military Representative Office of army equipment department in Nanjing, Nanjing, Jiangsu Province 210094, China<sup>d</sup> Science and Technology on Electro-Optical Information Security Control Lab, Tianjin 300000, China

## ARTICLE INFO

## Keywords:

Super-resolution  
Infrared image  
Convolutional neural network  
Multi-modal imaging  
Image fusion

## ABSTRACT

In this paper, we propose a deep-learning-based infrared-visible images fusion method based on encoder-decoder architecture. The image fusion task is reformulated as a problem of maintaining the structure and intensity ratio of the infrared-visible image. The corresponding loss function is designed to expand the weight difference between the thermal target and the background. In addition, a single image super-resolution reconstruction based on a regression network is introduced to address the issue that traditional network mapping functions are not suitable for natural scenes. The forward generation and reverse regression models are considered to reduce the irrelevant function mapping space and approach the ideal scene data through double mapping constraints. Compared with other state-of-the-art approaches, our experimental results achieve superior performance in terms of both visual effects and objective assessments. In addition, it can stably provide high-resolution reconstruction results consistent with human visual observation while bridging the resolution gap between the infrared-visible images.

## 1. Introduction

Image fusion techniques [1–3] aim to generate an informative image with specific algorithms from multiple source images. Thanks to the ability to recombine disparate information, infrared and visible image fusion technology plays a pivotal role in the detecting imaging systems. Hence, the fused result has a more distinct and complete depiction of the scene, which is beneficial to human perception and machine processing. The fusion image can synthesize a novel image with complementary information of the source images. Maximizing the integration of interest information is an essential bottleneck to reveal novel insights and fundamental scientific issues in biomedicine [4], forest fire fighting [5], and safe driving. For example, it is common to generate high dynamic range (HDR) images by applying the multiple exposure fusion (MEF) [6–8] approach. HDR imaging method can provide more prosperous image details, making reconstructed images more distinct and pleasing to human visual observation. Based on this approach, the infrared and visible fusion algorithm [9–11] can integrate the advantages of each information. Generally speaking, infrared images lack texture information and cannot effectively characterize the scene. Notwithstanding, it has been widely applied own to its inherent thermal radiation characteristics and the ability to realize cloud penetration imaging in long-wave infrared

bands. In contrast, the visible image contains texture details with high spatial resolution, which is conducive to enhancing the ability of target recognition and conforms to the human visual system. However, the visible image also has a fatal disadvantage: it is impossible to obtain a high-quality image under low illumination conditions. Therefore, visible-infrared imaging is interdependent and jointly promoted.

Although the image fusion technology has made significant improvements, the pixel size of the long-wave infrared detector has approached the physical limit (17 μm) due to limitations in software algorithms and hardware technology. Meanwhile, with the imaging resolution increasing, the manufacturing cost of the device will also dramatically expand. Therefore, the current dual-band image fusion technology is insufficient to stably realize all-weather high-resolution imaging. At this time, the traditional super-resolution (SR) models and algorithms are no longer suitable, and their computational complexity adds the pressure of massive calculation to the application. Recently, deep learning (DL) [12,13] has emerged as a powerful technique in the field of image fusion owing to its outstanding feature extraction, representation capability, strong robustness, and efficient reconstruction performance. From the artificial intelligence robot developed by Deepmind company to the powerful robot dog in Boston, promising news came one after another. Artificial intelligence [14–17] produces a familiar word around

\* Corresponding author.

E-mail address: [zuochao@njust.edu.cn](mailto:zuochao@njust.edu.cn) (C. Zuo).<https://doi.org/10.1016/j.optlaseng.2022.107078>

Received 27 December 2021; Received in revised form 31 March 2022; Accepted 10 April 2022

Available online 21 April 2022

0143-8166/© 2022 Elsevier Ltd. All rights reserved.

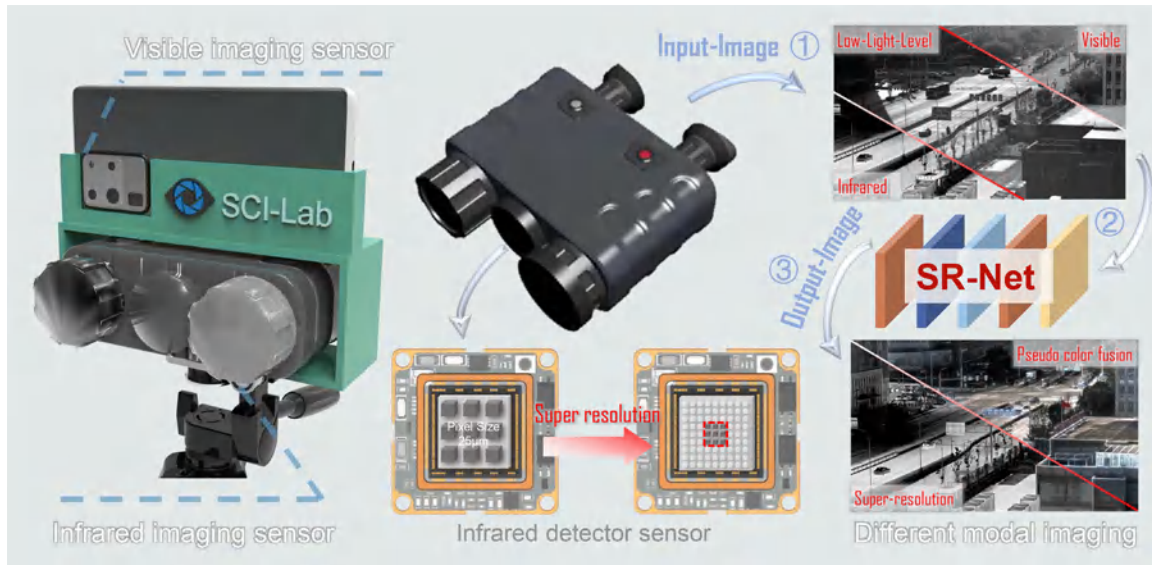


Fig. 1. Structural diagram and imaging reconstruction notion of the cross-modal fusion imaging system.

us. This is a remarkable manifestation of the gradual replacement of manual operation by intelligent machines. This trend is being driven by the increasing demand for the emergence of multi-dimensional sensors coupled with artificial intelligence computing technology. Over the past decades, deep learning technology has become a research hotspot in the era of massive data. Both academia and industry show strong interest to this technology, especially in computer vision [18,19]. As a "Data-Driven" technology that has emerged in recent years, it has achieved surpassing achievement in many applications such as image classification [20], object detection [21,22], and recognition [23,24]. And as shown in Fig. 1, overcoming the pixelation imaging problem caused by inadequate spatial sampling is also the novelty of Multi-image super-resolution fusion (Multi-SR-Fusion) technology.

The remaining structures of this paper are as follows. In Section 2, we briefly review related works on deep learning frameworks. Section 3 depicts the basic principle of our proposed method. Section 4 presents the details of the proposed Multi-SR-Fusion network for infrared and visible image fusion. Abundant experimental results and analysis are illustrated in Section 5. Finally, Section 6 provides a discussion and summarizes the paper.

## 2. Related works

At present, benefiting from the powerful feature extraction ability of DL convolution operation and learning mapping function parameters from massive data, the DL method has rapidly evolved the most potential direction in the field of image fusion. The traditional single-frame image SR [25,26] problem refers to the process of recovering from low-resolution (LR) images to high-resolution images, constantly pushing the limits to obtain higher real-world perception. In the field of computer vision, the introduction of convolutional neural networks (CNNs) [27] has extensively promoted the development of single image SR technology. The researchers continuously optimize the SR network model by introducing residual models, deep convolutional structures, and dense connectivity structures to enhance the reconstruction performance. However, due to the ill-posedness of the single image SR issue, most existing methods will generate artifacts and even lose the detailed texture under the condition of the sizeable scaling factor. Therefore, it is still a challenge to accurately reconstruct the high-frequency image details. Of the prominent DL-based methods, there are two mainstreams: convolutional neural network (CNN) [28–31] and generative adversar-

ial network (GAN) [32–34]. A majority of representative works have been proposed on this challenging problem.

In ICCV 2017, a classical fusion method, termed as DeepFuse [35], was put forward to tackle the exposure image fusion task. On this basis, Li et al. replaced the convolution network in the previous part with dense-block for improvement [36]. The fusion network is composed of the encoder, fusion layer, and decoder structure. Considering the similarity between the fused features and the original image, Zhang et al. created the proposed method better focused on the effective extraction of image features [37] by the continuous feedback of feature information from each layer. With the rapid development of the GAN network, scholars have also applied it to the field of infrared and visible images. Ma et al. proposed a detail-preserving learning-based fusion model for infrared and visible images [38]. The dual loss functions of detail loss and target edge enhancement loss are designed to improve the quality of detail information and sharpen the edges of IR targets, respectively, in the adversarial network generation framework. Nonetheless, this method does not fully consider the characteristics of infrared and visible images, and the fused images are challenging to highlight the target information. According to the aspects of infrared-visible imaging, Li et al. proposed a GAN network with a multi-scale attention mechanism [39]. The multi-scale attention mechanism generator focuses on the target information of the infrared image and the background detail information of the visible image so that the fusion network can concentrate on the specific area of the source image to reconstruct the fusion image. Generally speaking, the method based on DL can produce satisfactory results without manually designed decomposition processing and fusion rules. However, they can not highlight important targets while retaining background information, resulting in low contrast of fusion results. Due to the limitations of the manufacturing process, power consumption, or the cost of the sensor, the pixel imaging of infrared images has not been sufficiently solved. Zou et al. successfully realized the SR reconstruction of infrared images by employing the encoder-decoder network and also verified the application potential in image SR and feature extraction [40]. Therefore, if the SR structure can be added to the network, the fusion result will be predictable improved.

Gatys et al. proposed the neural style transfer method [41] and first applied the DL method to the style transfer task. The network maintains the consistency of the basic information of the two images through content loss constraints and updates the style of the input image by back-propagation iterations. By continuous forward propagation calculation loss and backpropagation optimization loss and updating the pixel value

of the reconstructed image, the optimal reconstructed image is eventually obtained. The essence of image style migration is the fusion of two different style images. In a sense, infrared and visible images can also be regarded as two separate "style" images. Therefore, this proposed method utilizes the notion of neural style transfer to alleviate the problem of infrared and visible image fusion.

As mentioned above, in recent years, infrared and visible image fusion technology based on the neural network has essential research prospects. In the task of infrared and visible image fusion, the following problems are still faced:

- (1) End-to-end imaging datasets. DL reconstruction algorithms are based on multiple datasets, while fewer datasets are available for infrared and visible image fusion tasks. How to utilize the existing data to realize the network training model is one of the challenges. And the most critical point is that the current fusion networks do not consider the resolution of infrared images, and the quality of input infrared images is too poor, resulting in unsatisfactory reconstruction results.
- (2) The resolution gap between the infrared-visible images. In the task of infrared-visible fusion, generally speaking, the resolution of the infrared detector will generally be much worse than the visible detector. Therefore, whether the infrared imaging quality can be improved through the mapping function to enhance the quality of fusion image is also one of the critical contents of this paper.
- (3) Network structure. Image fusion is a low-level task in computer vision, and the network structure should be as lightweight as possible. And how to give full play to the network ability and trade-off the weight between two images is also one of the fundamental problems.
- (4) Loss function. In the network training process, the network training parameter needs to be modified by the loss function, which puts forward more strict requirements for the loss function design.

### 3. Proposed methods

For the human visual system, the "conspicuity area" that containing essential targets is more attractive. Based on the above analysis, the problem of infrared-visible image fusion is how to maintain the high-frequency detail information and the thermal radiation information so as to realize a multi-dimensional data fusion process. The primary task of the proposed method is to improve the resolution of the infrared image and then carry out the weighted fusion of the heterologous image while obtaining a high-quality image resolution. Therefore, efficiently extracting the feature information of each image and assigning fusion weight is the focus of our research. Based on the concept of U-net semantic segmentation and style transfer [42], the thermal radiation information of the infrared image can be effectively segmented, and then the thermal image and visible texture information are transferred by style transfer structure. In our workflow, the coding-decoding fusion structure is employed for end-to-end learning, as shown in Fig. 2, so that the network can not only center on the "conspicuity area" information but also learn the image SR mapping function. The image merge problem is transformed into the issue of maintaining the structure and intensity ratio of infrared and visible images. The corresponding loss function is designed to expand the weight distinction between the thermal target and the background. Aiming at the shortage that the traditional network mapping function is ill-posed in the actual scene, the additional constraint of inverse regression is embedded to reduce the space of the possible mapping function. Lastly, the pseudo color SR reconstruction based on the scene is realized by expanding the number of channels. By doing so, the reconstructed image is more in line with the human visual effect.

Note that our method takes the infrared image and visible image as the input image and obtains the colorized fusion image through end-to-end supervised network. Multi-scale feature extraction is performed in infrared and visible images by applying the diverse dimensions kernels. Subsequently, the infrared and visible fusion image is generated through

**Table 1**  
The number of layers in the network structure.

Layer	Parameter	Numbers
Convolution layer	$1 \times 1$ , Strides = 1, padding = SAME	4
Convolution layer	$3 \times 3$ , Strides = 1, padding = SAME	21
Convolution layer	$3 \times 3$ , Strides = 2, padding = SAME	4
Convolution layer	$5 \times 5$ , Strides = 1, padding = SAME	4
ReLU layer	-	12
LReLU layer	alpha = 0.2	16
Concat layer	-	6
Deconvolution layer	$3 \times 3$	4
Element-max later	-	1
Global average pooling layer	-	4
Fully connected layer	-	8
Sigmoid layer	-	4
Pixelshuffle layer	$2 \times 2$	2
Max-pooling layer	$2 \times 2$	4

the fusion layer. The fusion structure contains multi-scale feature extraction and residual channel attention blocks (RCAB), which enables valuable feature mapping and suppresses unimportant feature mapping. The coding-decoding SR structure realizes the functions of feature extraction and reconstruction, respectively. Meanwhile, the introduction of the skip connection structure can transfer the image feature information from the encoding part to the decoding part of the network, solving the problem of gradient disappearance.

#### 3.1. Problem formulation

To express the mapping relationship of the network more clear, the network model can be defined as:

$$I_{out}(x, y) = F_{\omega, \theta} [I_{LR1}(x, y), I_{LR2}(x, y)] \quad (1)$$

where,  $F_{\omega, \theta}[\cdot]$  represents the nonlinear mapping function of the network,  $\omega$  and  $\theta$  respectively describe the weight and deviation trainable parameters in the network,  $I_{LR1}(x, y)$  describes the input long-wave infrared image,  $I_{LR2}(x, y)$  describes the input visible image, and  $I_{out}(x, y)$  is the HR image output by the network. Detailed network parameters are illustrated in Table 1.

The network structures contain the convolution, deconvolution, element-addition or multiplication, channel-fusion, max-pooling, and element-max layers. The input image of the  $X_i$  layer is represented by  $i$ , and the convolution layer and deconvolution layer are represented as:

$$F(X_i) = \max(0, W_k * X_i + B_k) \quad (2)$$

where,  $W_k$  and  $B_k$  represent filter and deviation respectively. For convenience,  $*$  represents convolution or deconvolution.

For the element-addition layer, the output is the addition of two inputs of the same size, followed by Leaky Rectified Linear Unit(LReLU) activation:

$$F(X_i, X_j) = \begin{cases} X_i + X_j, & X_i + X_j \geq 0 \\ \alpha * (X_i + X_j), & X_i + X_j < 0 \end{cases} \quad (3)$$

where,  $X_i$  and  $X_j$  represent layer  $i + 1$  and layer  $j + 1$  respectively, and  $\alpha = 0.01$ .

For the element multiplication layer, the output is the multiplication of two elements of the same size, followed by LReLU activation:

$$F(X_i, X_j) = \begin{cases} X_i \cdot X_j, & X_i \cdot X_j \geq 0 \\ \alpha * (X_i \cdot X_j), & X_i \cdot X_j < 0 \end{cases} \quad (4)$$

For the channel fusion layer, the output is the sum of two input channels of the same size:

$$F(X_i, X_j) = X_i \oplus X_j \quad (5)$$

For the max-pooling layer, the output image size is half of the input image, which is expressed by the following formula:

$$F(X_i) = \text{down}(X_i) \quad (6)$$



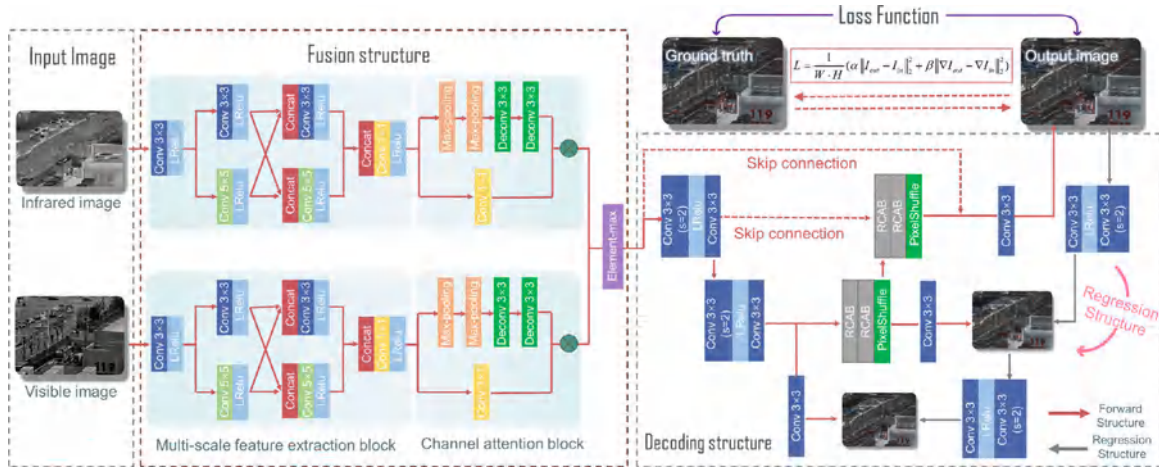


Fig. 2. Super-resolution fusion network structure of heterogeneous images based on encoding-decoding structure.

where *down* represents pooling function, and this paper adopts max-pooling.

For the element-max layer, the size of the output image is the same as the input image, which is expressed by the following formula:

$$F(X_i, X_j) = \max(X_i, X_j) \quad (7)$$

For the sub-pixel convolution layer, the output image size is twice of the input image, which is expressed by the following formula:

$$F(X_i) = \text{pixelshuffle}(X_i) \quad (8)$$

### 3.2. Loss function

Weight distribution is the core problem of image fusion, which directly determines the quality of fused image. To perform network training, we need to accurately evaluate the information similarity between the fused image and the input image pair to minimize information loss, thus effectively preserving the thermal radiation information from the infrared image and the textural detail information from the visible image. Therefore, in this paper, the image fusion problem is transformed into the issue of maintaining the structure and intensity ratio of infrared-visible images. The intensity distribution and gradient information can characterize the thermal radiation and structural information, respectively. In order to preserve the representative features of the source image to the greatest extent, a hybrid loss function is designed to retain valuable feature information. Thus, the loss function of our proposed model is set to:

$$Loss = \sum_{i=1}^N Loss_1(F(x_i), y_i) + \lambda Loss_2(D(y_i), x_i) \quad (9)$$

where  $x_i$  and  $y_i$  respectively represent the input LR and output HR images.  $Loss_1(F(x_i), y_i)$  and  $Loss_2(D(y_i), x_i)$  describe the loss functions of forward regression and inverse regression tasks, respectively. During the training process, the reconstructed images  $F(x_i)$  continuously converge to the corresponding HR images. Similarly, the similarity between the predicted image  $D(y_i)$  and the forward input LR image is continuously approached in the regression process. Here we set  $\lambda$  to 0.1 for the weight distribution of the hybrid loss function.

If  $F(x_i)$  is the accurate HR image, the image  $D(y_i)$  in the inverse regression model should be dramatically similar to the LR image. With this constraint, we can reduce the possible mapping function so as to achieve robust image reconstruction.

$$Loss_1 = \alpha \|y_i - F(x_i)\|_2^2 + \beta \|\nabla y_i - \nabla F(x_i)\|_2^2 \quad (10)$$

where,  $\|\cdot\|_2$  defines the  $L_2$  norm,  $\nabla$  represents the gradient operator.  $\alpha$  and  $\beta$  are two factors that balance these two terms,  $\alpha = \beta = 0.5$  in this

experiment.

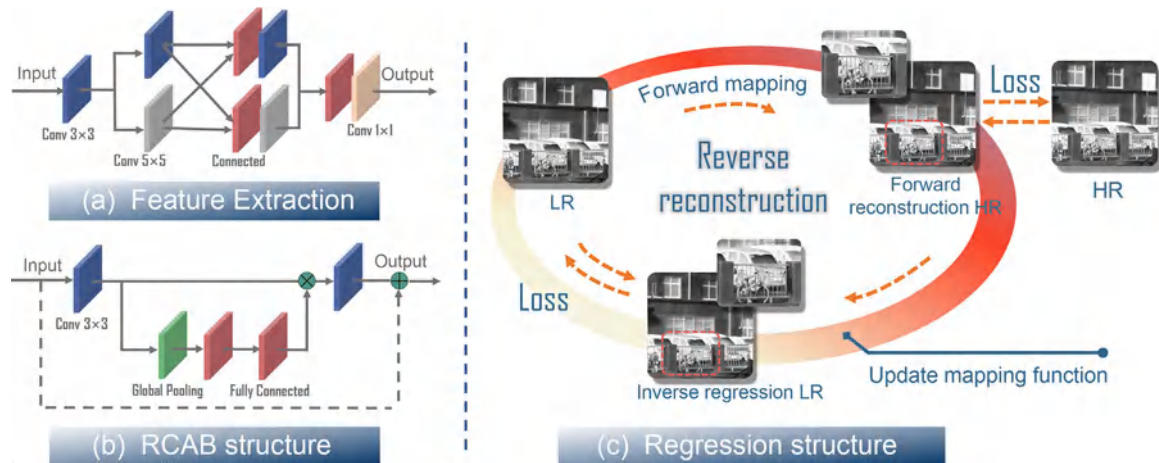
$$Loss_2 = \|D(y_i) - x_i\|_2^2 \quad (11)$$

This formulation is an improved fusion method by taking SR into account. The forward generation process and reverse regression process of the input-output image are simultaneously constrained, and the dual-loss functions compensate each other to produce the whole loss function balance. The mixed loss between the input-out images is computed to update network parameters. By minimizing the loss, the network performs accurate reconstruction of the input data in the training phase, emphasizes the valuable information, and suppresses the irrelevant information.

## 4. Network architecture

### 4.1. Multi-scale feature extraction (encoding) module

An essential part of SR reconstruction is how to extract the features of the input image. Suppose the different dimensions information can be obtained. In that case, it will conduce for signal restoration. On the other hand, the image feature information is generally extracted by a convolution kernel. Therefore, the idea of extracting the image with large convolution to obtain a more extensive receptive field has been sprouting. A larger receptive field will facilitate the reception of feature information. However, if the convolution kernel is too large, the amount of calculation will increase sharply, which is not conducive to the boost of model depth. Therefore, we can decompose the large-scale convolution into several small-scale convolutions so as to reduce the amount of calculation. Although multi-scale convolutional blocks can extract adequate features, it is also crucial to selectively focus on the essential elements and ignore the less important ones. This means that not all features are beneficial for reconstruction. Intermediate features contain valuable information, such as primary structure and details, or even irrelevant information, such as noise. Therefore, We adopt a multi-scale layer with different kernel sizes, such as  $3 \times 3$  and  $5 \times 5$ , to acquire low-frequency and high-frequency features with various receptive fields. By doing so, comprehensive image information at different scales is fetched and reused with each other. The feature fusion convolution layer virtually reduces the computational complexity and improves the convergence speed of the network. Consequently, introducing a multi-scale extraction module is profitable to obtain higher-level robust semantic features, retain more underlying details, and enrich the image feature information.



**Fig. 3.** Schematic diagram of the critical network modules. (a) Multi-scale feature extraction structure. (b) Residual channel attention blocks. (c) Dual-regression mapping structure.

#### 4.2. Super-resolution (decoding) module

The SR network adopts an encoder-decoder architecture. In the decoding layer, the pixel-shuffle method is operated to enlarge the feature map size corresponding to the convolution layer in the coding layer, and the different dimensional information is transmitted by skip connection. Skip connection can not only transfer image feature information but also alleviate the problem of gradient disappearance. We introduce the residual channel attention module to adjust the channel feature information, which is conducive to reconstructing HR images. The global average pooling layer encodes all spatial features into a whole feature on one channel. After receiving the global features, the nonlinear relationship between each channel is learned through the full connection layer. The whole operation can be regarded as learning the weight coefficients of each channel to make the model more discriminative about the features of each channel.

Currently, the mainstream network architecture model is moving in a deeper direction. A deeper network model means that it has better nonlinear expression ability. Thereby, it can learn more complex transformations and fit more complex feature inputs. However, a common accompaniment problem is that the information extracted by the middle layers is not employed thoughtfully. Therefore, the skip connection in the residual structure is worthwhile to enhance the gradient propagation and alleviate the problem of gradient disappearance caused by network deepening. In addition, the existing methods only focus on the mapping from the LR image to the HR image. However, the under-determined possible mapping space is volatile and challenging during the training process. In order to ameliorate this problem, we propose a dual regression project in the SR structure, as shown in Fig. 3(c). Through the restriction of double constraints, the robustness of the network model and its applicability to natural scenes can be promoted.

## 5. Experiment and results

### 5.1. Network and dataset settings

In the network, the batch size is 4, and the epoch is set to 200. Empirically, we use Adam optimizer to optimize the network structure, and the initial learning rate is set to  $10^{-4}$ . The network is conducted on hardware platform with an Intel Core™ i7-9700K CPU @ 3.60GHz×8, and RTX2080Ti. The software platform is running under Ubuntu 16.04 operating system. The total training time of our network is 11.20 hours, and the average test time for each image is 1.31 seconds.

The long-wave infrared (self-developed,  $800 \times 600$ ,  $25 \mu\text{m}$ ) and visible images are collected by the cross-modal image acquisition

equipment and transmitted to the network for training after registration. The corresponding images are cut into  $128 \times 128$  pieces and sent to the network for training. The infrared dataset contains 1300 images, of which 800 images are employed as the training set, and 200 images are utilized as the validation sets. The fusion dataset includes the lake, jungle, and urban imaging environments (<https://figshare.com/s/0d35b35c18c70cd3bba1>).

It is worth noting that the HR infrared images are acquired at long focal lengths, and conversely, the LR infrared images are obtained at short focal lengths (large field-of-view imaging). The pixel mapping of the HR image is yielded by partially recording the central region of the LR image, as shown in Fig. 1. Instead of creating the training dataset through simulations (bicubic down-sampling or an approximate model of the point spread function), in the presented technique, the desired target  $3 \times$  super-resolved images are accordingly obtained by tripling the focal length (25mm-75mm).

We utilized the visual saliency map (VSM) and weighted least square (WLS) to realize heterogeneous image fusion. The original image can be decomposed into the bottom and several detail layers by multi-scale decomposition (MSD). The bottom layer mainly contains low-frequency information, which determines the overall appearance and the fused image contrast. In this paper, VSM is used to merge the bottom layer to effectively extract the salient structure so as to avoid the blurring of low-frequency information. Detail layers are merged according to the traditional "maximization" rule. The absolute value of the detail layer coefficient is considerable, which corresponds to more significant features.

The monochrome display has constantly perplexed low-light-level night vision and infrared imaging systems. Therefore, it is also an essential task to employ the visible color component information to achieve pseudo-color of fused images. We map the RGB color components to the HSV color space in the color migration task. The grayscale fused image is created as the V component of the predicted image, and the chromaticity H and saturation S are kept constant to achieve the final color image output.

### 5.2. Experimental results analysis

A majority of visible texture information plays a significant role in restoring and reconstructing HR color fusion images. However, in the night vision imaging environment, the visible detector can not provide enough detailed information, so improving the SR reconstruction ability of the infrared image is also an essential research direction. In order to verify this concept, we partially modify the network structure and remove the visible image from the input structure. Fig. 4 depicts

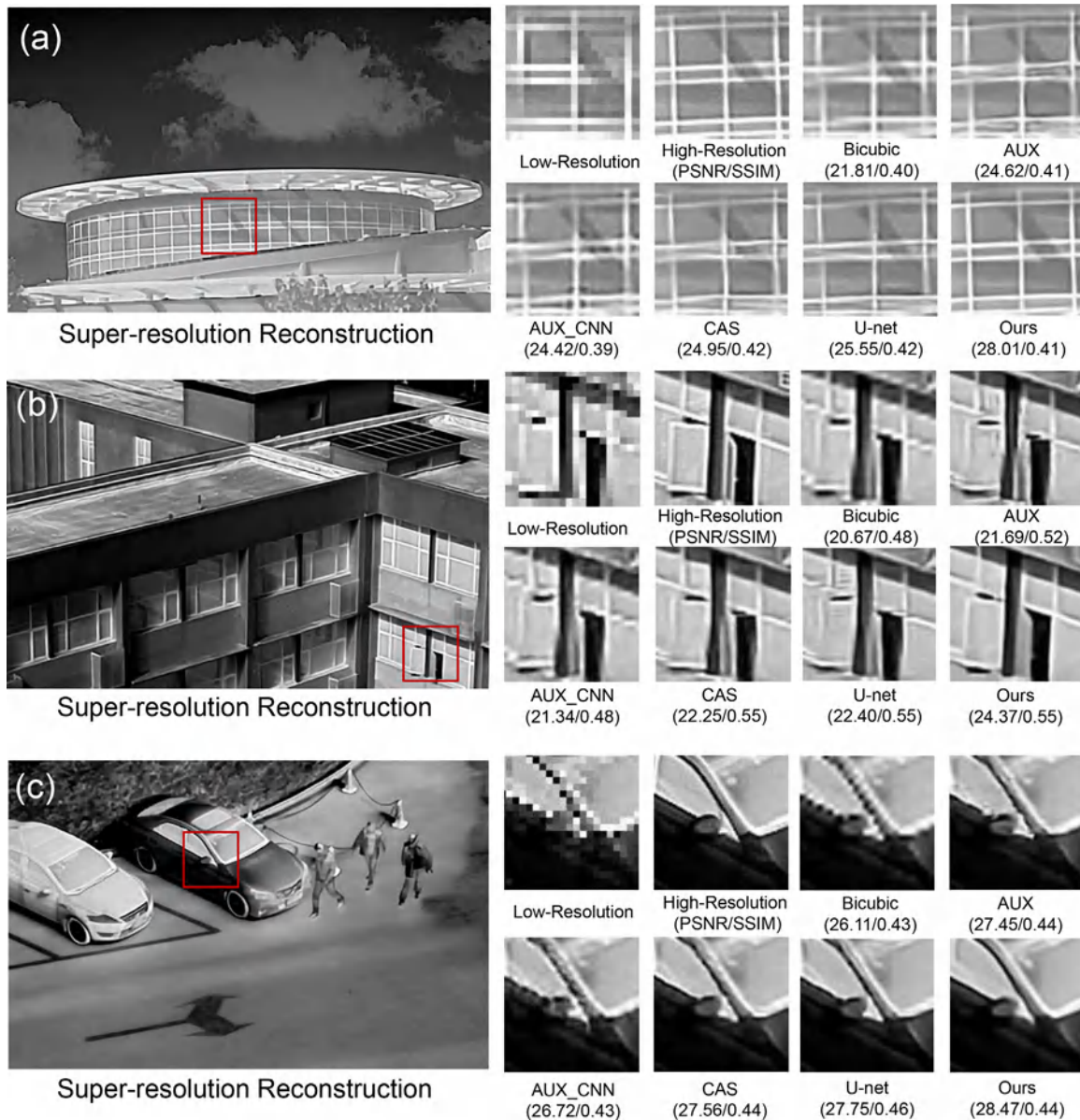


Fig. 4. The comparison of super-resolution imaging results with different scenes (Scale = 3).

the comparison of SR reconstruction results in three different scenes. It can be seen that our method has been sufficiently enhanced in the reconstructed image, whether in edge details or the recovery of spatial frequency components. Compared with bicubic interpolation, auxiliary neural network (AUX) [43], infrared image super-resolution imaging algorithm based on the auxiliary convolutional neural network (AUX-CNN) [44], cascade super-resolution (CAS) [45], and skip connected super-resolution (U-net) approach [40], our method improves the peak signal-to-noise ratio by 4.08dB, 2.36dB, 2.79dB, 2.03dB and 1.71dB, respectively. In addition, from the visual imaging performance, our results are consistent with the HR truth image and avoid the artifact phenomenon in the SR reconstruction result. Therefore, from a comprehensive point of view, the SR image obtained by the proposed method is more prominent. At the same time, it also verifies the feasibility of applying a dual-regression network to improve the SR reconstruction performance.

After verifying the feasibility of the network, we employed the network for heterogeneous image fusion processing and made comparisons with the anisotropic diffusion and Karhunen Loeve transform

(ADF) [46], fourth order partial differential equations (FPDE) [47], multi scale guided (MGFF) [48], multi singular value decomposition (MSVD) [49] and two scale image fusion using saliency detection (TIF) [50] methods, respectively, and the corresponding reconstruction results are shown in Fig. 5. Although it is difficult to accurately evaluate the visual quality of these methods, we can perceive apparent differences between them. As shown in Fig. 5, all the fusion methods have accomplished the task of merging the information of infrared and visible images to some extent. Overall, our method embraces more textual details while highlighting the important targets.

The reconstruction results are suitable for human eye perception due to the advantages of the high signal-to-noise ratio of the output image and complementary fusion information. From the objective data in Fig. 6, the evaluation indexes of the fused image in spatial frequency, edge intensity, and average gradient were improved over the existing imaging algorithm by 3.35, 8.97, and 0.94, respectively. The comparative data in Table 2 also verify the feasibility of introducing super-resolution networks to improve the reconstruction performance. The image fusion task is reformulated as a problem of maintaining the structure



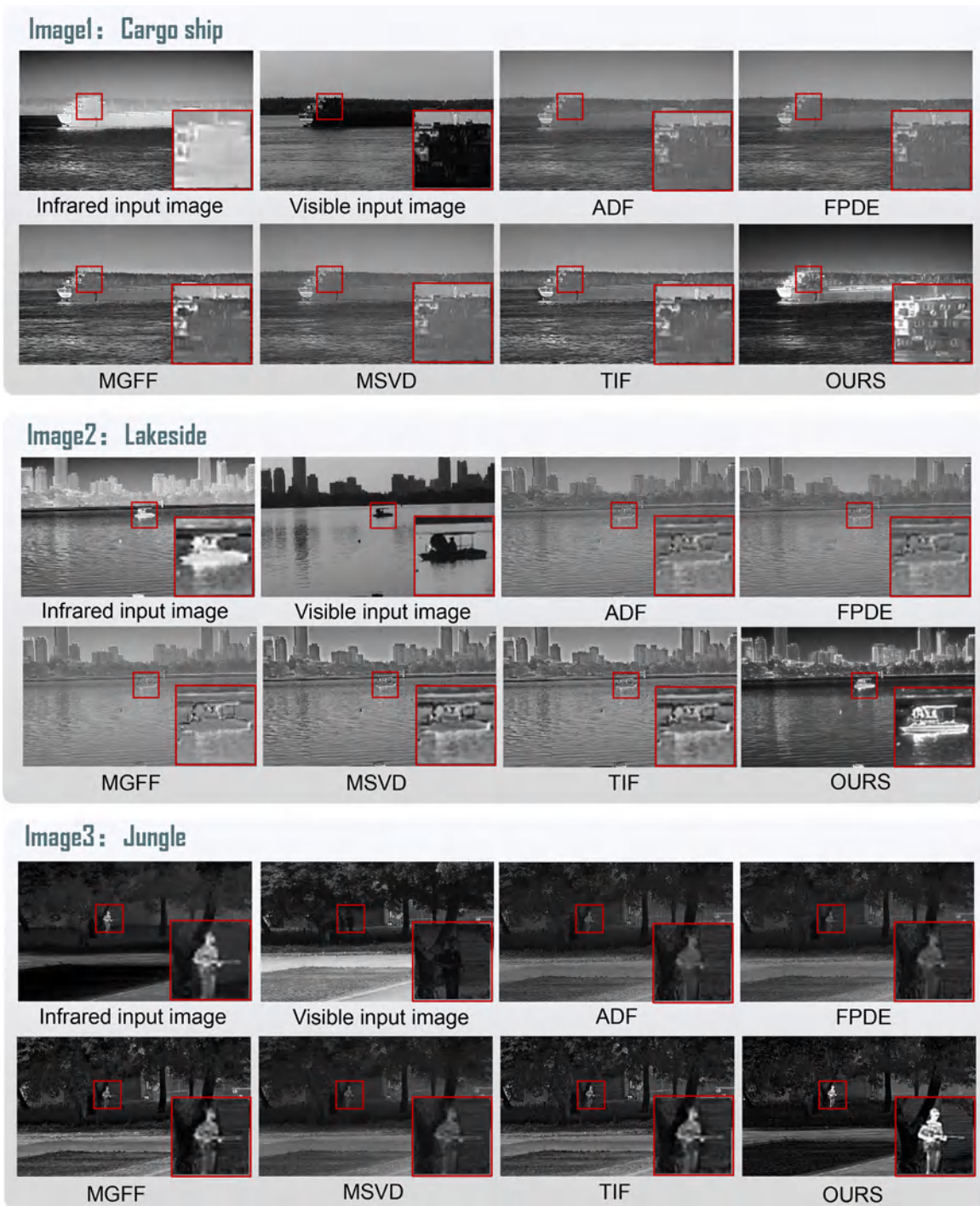


Fig. 5. The comparison of imaging fusion results with different scenes.

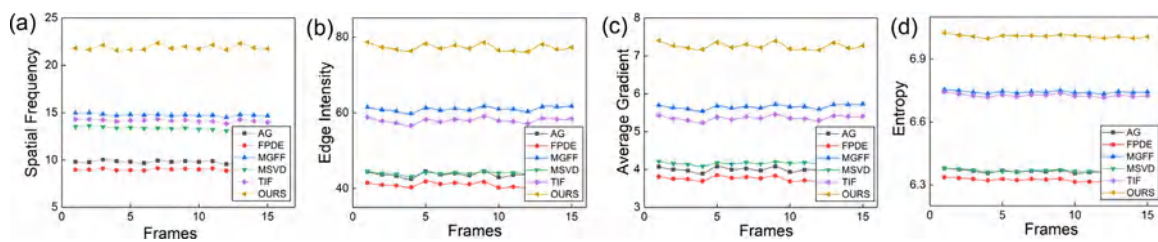
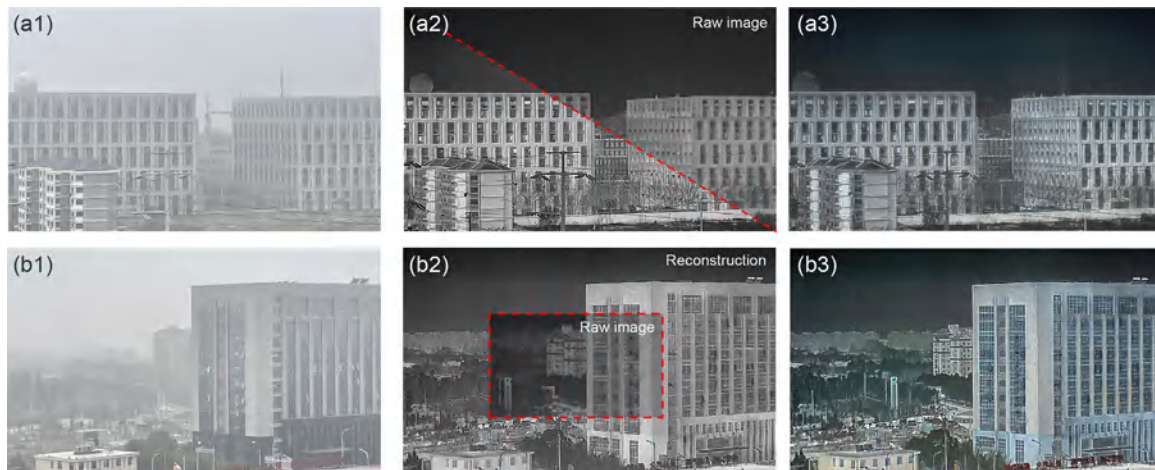


Fig. 6. Index evaluation curve under continuous frames of the same scene.

**Table 2**

The comparison of imaging fusion evaluation index with different scenes. The bold text indicates the best result.

Number	Methods	AG	Edge intensity	Entropy	Mutinf	Qcv	Rmse	SF
Image 1	ADF	6.2074	59.8872	6.8278	1.8177	0.1008e+03	0.0625	15.6246
Image 1	FPDE	5.7650	55.6909	6.7977	1.8299	0.1142e+03	0.0622	14.0359
Image 1	MGFF	6.9055	68.4516	<b>7.0109</b>	1.6050	0.2067e+03	0.0644	17.3653
Image 1	MSVD	5.3488	50.8112	6.7623	<b>1.8511</b>	0.1082e+03	0.0622	14.5423
Image 1	TIF	6.0770	60.4062	6.9641	1.6900	0.0927e+03	0.0634	15.6507
Image 1	Ours	<b>7.3762</b>	<b>71.4749</b>	6.6295	1.3591	<b>0.6739e+03</b>	<b>0.0698</b>	<b>18.3819</b>
Image 2	ADF	4.4591	45.9656	6.8876	2.1563	0.4278e+03	0.0705	12.7356
Image 2	FPDE	4.4130	45.5108	6.8797	2.1499	0.4231e+03	0.0704	12.1605
Image 2	MGFF	5.3440	60.8598	<b>7.2126</b>	1.9414	0.4549e+03	0.0725	17.3130
Image 2	MSVD	4.5975	46.9185	6.8986	2.1679	0.4229e+03	0.0705	13.9639
Image 2	TIF	5.2618	55.5220	7.1010	2.0178	0.2774e+03	0.0717	14.5801
Image 2	Ours	<b>5.9050</b>	<b>66.7448</b>	7.0556	<b>2.2175</b>	<b>0.9902e+03</b>	<b>0.1070</b>	<b>18.6244</b>
Image 3	ADF	4.0698	44.4168	6.2849	0.9410	1.1340e+03	0.0650	9.8957
Image 3	FPDE	3.8069	41.4784	6.2352	0.9521	1.1079e+03	0.0647	9.0497
Image 3	MGFF	5.6396	60.8485	6.6629	0.9208	1.0847e+03	0.0661	14.7737
Image 3	MSVD	4.1813	44.2260	6.2795	0.9723	1.1131e+03	0.0648	13.9639
Image 3	TIF	5.4249	58.7020	6.6533	0.8981	1.0184e+03	0.0665	14.1778
Image 3	Ours	<b>7.4351</b>	<b>78.8509</b>	<b>7.0185</b>	<b>1.2953</b>	<b>2.3615e+03</b>	<b>0.1062</b>	<b>22.4927</b>



**Fig. 7.** The imaging results of the proposed algorithm in severe weather (foggy days). (a1, b1) Visible image. (a2, b2) Infrared image. (a3, b3) Fusion image.

and intensity ratio of the infrared-visible image, solving the problem of poor quality fusion performance and thermal information blurring due to the low resolution of the infrared image in conventional fusion imaging.

For the fusion imaging problem under severe weather (foggy days), we have also explored it accordingly. As shown in Fig. 7, under a foggy sky, the scene captured by the visible detector is muddy and contains an amount of interference information. On the contrary, long-wave infrared detectors capture unique signals by virtue of the characteristics of penetrating smoke imaging and thermal radiation sensing. The multi-scale feature extraction network effectively realizes the high-frequency information fusion of different detectors in the proposed method. An excellent color fusion image can be achieved with the help of color information from the visible detector, as depicted in Fig. 7(a3, b3). However, the infrared image also has the imaging problem of poor contrast due to less thermal radiation information on foggy days. By regressing the output of the super-resolution network, the corresponding high-frequency detail information is basically restored, as shown in Fig. 7(a2, b2).

In addition, the recovery of image color information is also an uncertainty problem. Deep learning-based color image reconstruction is mainly established on specific scenes and cannot recover color information that does not appear in the training set. Therefore, this im-

poses strict requirements on the training set, which should contain as much color information as possible for various scenes. Fig. 8 portrays the multi-modal imaging results of heterologous images based on the regression network. Various modes of reconstruction such as pseudo-color, SR reconstruction, and edge extraction are realized. See supplementary visualization materials 1, 2, and 3 for specific imaging videos. The experimental results indicate that the network is able to perform fused images containing thermal information of infrared images and high-frequency information of visible images, which comprehensively enhances the resolution of detailed textures of infrared images. At the same time, the obtained colored image is consistent with the visual perception effect of the human eyes. With the guidance of thermal radiation signals, the contour markings of moving objects can be unambiguous marked to further facilitate the information perception ability. For instance, the image resolution and thermal information of toy guns held by pedestrians are significantly improved in infrared images. From the reconstruction results in Fig. 8(c), we can clearly observe that the fused target images are effectively highlighted in the low light environment, which will be conducive to the subsequent target recognition and tracking. In general, our method can offer robust adaptability in different imaging environments, and it will also provide a promising way to improve the quality of infrared-visible fusion.



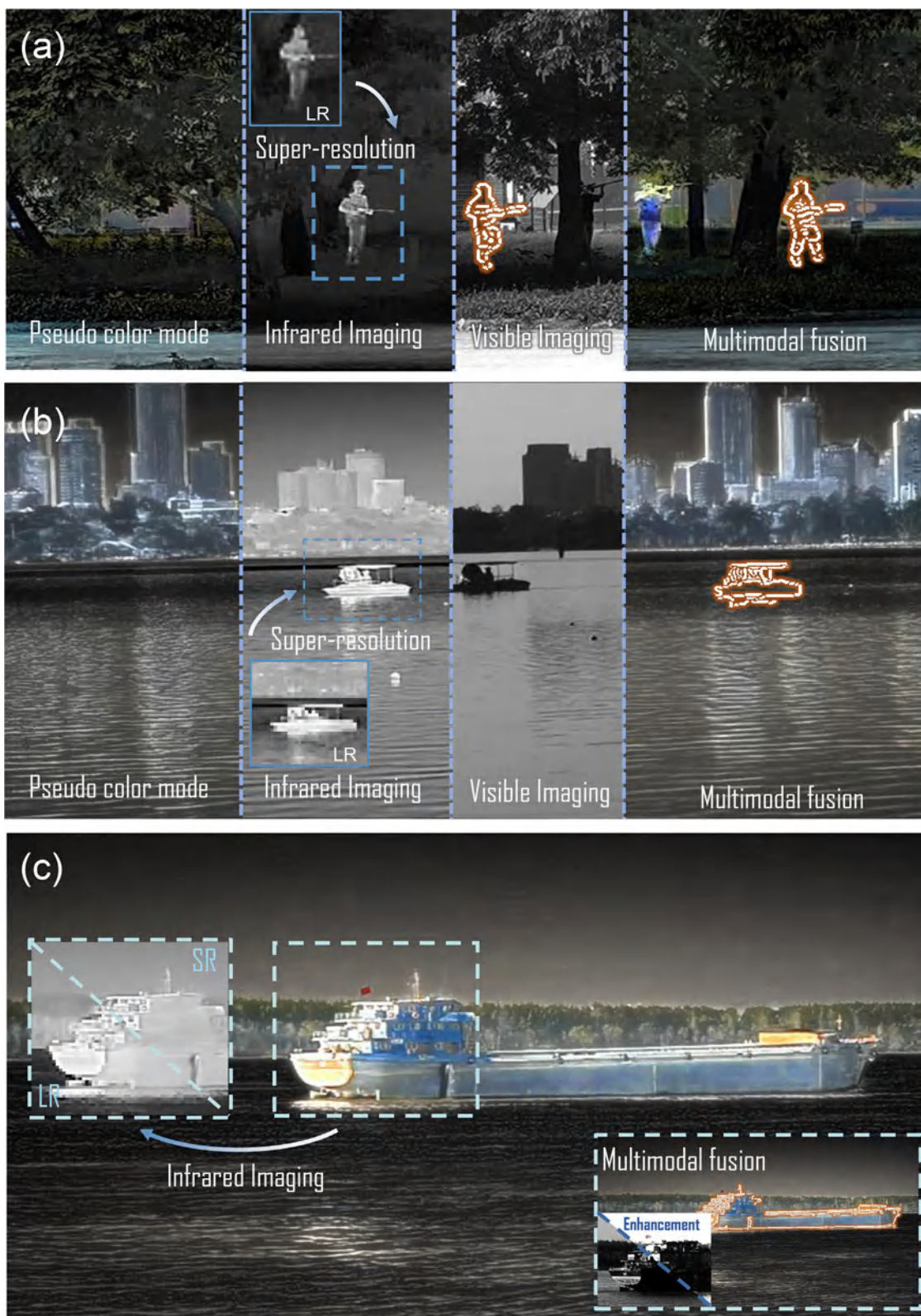


Fig. 8. Cross modal reconstruction results in different scenes.



## 6. Conclusion

To address the bottleneck of low-quality fusion imaging caused by different imaging mechanisms and mismatched spatial resolution of heterogeneous detectors, an infrared-visible cross-modal color fusion network based on DL is proposed. Affording the conception of semantic segmentation and style transfer, the encoding-decoding fusion network is adopted for end-to-end learning to improve the feature expression ability and suppress the interference of useless information. The corresponding dual-loss function is designed to expand the weight difference between thermal target and background. Experimental results prove the superiority in terms of visual quality and quantitative criteria compared to five representative methods. The evaluation indexes of spatial frequency, edge intensity, and average gradient were improved by 3.35, 8.97, and 0.94, respectively, which significantly improved the imaging quality of the fused images and verified the application potential of the network. On this basis, the imaging output of HR infrared reconstruction, heterologous image pseudo color fusion, edge feature extraction and other modes are realized, which opens new avenues for subsequent HR reconnaissance and identification tasks.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## CRedit authorship contribution statement

**Bowen Wang:** Conceptualization, Methodology, Visualization, Writing – original draft. **Yan Zou:** Visualization, Investigation. **Linfei Zhang:** Investigation. **Yuhai Li:** Formal analysis. **Qian Chen:** Supervision, Funding acquisition. **Chao Zuo:** Funding acquisition, Supervision, Writing – review & editing, Project administration.

## Acknowledgement

This work was supported by the [National Natural Science Foundation of China \(61905115, 62105151, 62175109, U21B2033\)](#), [Leading Technology of Jiangsu Basic Research Plan \(BK20192003\)](#), [Youth Foundation of Jiangsu Province \(BK20190445, BK20210338\)](#), [Fundamental Research Funds for the Central Universities \(30920032101\)](#), and [Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense \(JSGP202105\)](#).

## Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:[10.1016/j.optlaseng.2022.107078](https://doi.org/10.1016/j.optlaseng.2022.107078).

## References

- [1] Stathaki T. Image fusion: algorithms and applications. Elsevier; 2011.
- [2] Sahu DK, Parsai M. Different image fusion techniques—a critical review. *Int J Mod Eng Res (IJMER)* 2012;2(5):4298–301.
- [3] Chen Y, Cheng L, Wu H, Mo F, Chen Z. Infrared and visible image fusion based on iterative differential thermal information filter. *Opt Lasers Eng* 2022;148:106776.
- [4] James AP, Dasarthy BV. Medical image fusion: a survey of the state of the art. *Inform Fus* 2014;19:4–19.
- [5] Nematidine SM, Gupta D. Nonsampled contourlet domain visible and infrared image fusion framework for fire detection using pulse coupled neural network and spatial fuzzy clustering. *Fire Saf J* 2018;101:84–101.
- [6] Shen J, Zhao Y, Yan S, Li X, et al. Exposure fusion using boosting laplacian pyramid. *IEEE Trans Cybern* 2014;44(9):1579–90.
- [7] Mertens T, Kautz J, Van Reeth F. Exposure fusion: A simple and practical alternative to high dynamic range photography. In: *Computer graphics forum*, vol. 28. Wiley Online Library; 2009. p. 161–71.
- [8] Li ZG, Zheng JH, Rahardja S. Detail-enhanced exposure fusion. *IEEE Trans Image Process* 2012;21(11):4672–6.
- [9] Xiang T, Yan L, Gao R. A fusion algorithm for infrared and visible images based on adaptive dual-channel unit-linking pcnn in nsct domain. *Infrared Phys Technol* 2015;69:53–61.
- [10] Ma J, Chen C, Li C, Huang J. Infrared and visible image fusion via gradient transfer and total variation minimization. *Inform Fus* 2016;31:100–9.
- [11] Ma J, Ma Y, Li C. Infrared and visible image fusion methods and applications: a survey. *Inform Fus* 2019;45:153–78.
- [12] LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521(7553):436–44.
- [13] Deng L, Yu D. Deep learning: methods and applications. *Found Trend Signal Process* 2014;7(3–4):197–387.
- [14] Russell S, Norvig P. Artificial intelligence: a modern approach; 2002.
- [15] Ertel W. Introduction to artificial intelligence. Springer; 2018.
- [16] Feng S, Chen Q, Gu G, Tao T, Zhang L, Hu Y, Yin W, Zuo C. Fringe pattern analysis using deep learning. *Adv Photon* 2019;1(2):025001.
- [17] Feng S, Zuo C, Hu Y, Li Y, Chen Q. Deep-learning-based fringe-pattern analysis with uncertainty estimation. *Optica* 2021;8(12):1507–10.
- [18] Vouloimos A, Doulamis N, Doulamis A, Protopapadakis E. Deep learning for computer vision: a brief review. *Comput Intell Neurosci* 2018;2018.
- [19] O'Mahony N, Campbell S, Carvalho A, Harapanahalli S, Hernandez GV, Krpalkova L, Riordan D, Walsh J. Deep learning vs. traditional computer vision. In: *Science and Information Conference*. Springer; 2019. p. 128–44.
- [20] Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:171204621* 2017.
- [21] Zhao Z-Q, Zheng P, Xu S-t, Wu X. Object detection with deep learning: a review. *IEEE Trans Neural Netw Learn Syst* 2019;30(11):3212–32.
- [22] Liu L, Ouyang W, Wang X, Fieguth P, Chen J, Liu X, Pietikainen M. Deep learning for generic object detection: a survey. *Int J Comput Vis* 2020;128(2):261–318.
- [23] Uçar A, Demir Y, Güzelış C. Object recognition and detection with deep learning for autonomous driving applications. *Simulation* 2017;93(9):759–69.
- [24] Eitel A, Springenberg JT, Spinello L, Riedmiller M, Burgard W. Multimodal deep learning for robust rgb-d object recognition. In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE; 2015. p. 681–7.
- [25] Van Ouwkerk J. Image super-resolution survey. *Image Vis Comput* 2006;24(10):1039–52.
- [26] Wang L, Zhao S. Super resolution ghost imaging based on fourier spectrum acquisition. *Opt Lasers Eng* 2021;139:106473.
- [27] Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, Liu T, Wang X, Wang G, Cai J, et al. Recent advances in convolutional neural networks. *Pattern Recognit* 2018;77:354–77.
- [28] Li H, Wu X-J, Kittler J. Infrared and visible image fusion using a deep learning framework. In: *2018 24th international conference on pattern recognition (ICPR)*. IEEE; 2018. p. 2705–10.
- [29] Wang B, Zou Y, Zhang L, Hu Y, Yan H, Zuo C, Chen Q. Low-light-level image super-resolution reconstruction based on a multi-scale features extraction network. In: *Photonics*, vol. 8. Multidisciplinary Digital Publishing Institute; 2021. p. 321.
- [30] Gurrola-Ramos J, Dalmau O, Alarcón T. U-Net based neural network for fringe pattern denoising. *Opt Laser Eng* 2022;149:106829.
- [31] Qian J, Feng S, Tao T, Hu Y, Li Y, Chen Q, Zuo C. Deep-learning-enabled geometric constraints and phase unwrapping for single-shot absolute 3d shape measurement. *APL Photonics* 2020;5(4):046105.
- [32] Ma J, Yu W, Chen C, Liang P, Guo X, Jiang J. Pan-gan: an unsupervised pan-sharpening method for remote sensing image fusion. *Inform Fus* 2020a;62:110–20.
- [33] Bell-Kligler S, Shocher A, Irani M. Blind super-resolution kernel estimation using an internal-gan. *arXiv preprint arXiv:190906581* 2019.
- [34] Yang X, Jiang P, Jiang M, Xu L, Wu L, Yang C, Zhang W, Zhang J, Zhang Y. High imaging quality of fourier single pixel imaging based on generative adversarial networks at low sampling rate. *Opt Lasers Eng* 2021;140:106533.
- [35] Ram Prabhakar K, Sai Srikr V, Venkatesh Babu R. Deepfuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: *Proceedings of the IEEE international conference on computer vision*; 2017. p. 4714–22.
- [36] Li H, Wu X-J. Densefuse: a fusion approach to infrared and visible images. *IEEE Trans Image Process* 2018;28(5):2614–23.
- [37] Zhang C, Hu H, Tai Y, Yun L, Zhang J. Trustworthy image fusion with deep learning for wireless applications. *Wirel Commun Mob Comput* 2021;2021.
- [38] Ma J, Liang P, Yu W, Chen C, Guo X, Wu J, Jiang J. Infrared and visible image fusion via detail preserving adversarial learning. *Inform Fus* 2020b;54:85–98.
- [39] Li J, Huo H, Li C, Wang R, Feng Q. Attentionfgan: infrared and visible image fusion using attention-based generative adversarial networks. *IEEE Trans Multimedia* 2020;23:1383–96.
- [40] Zou Y, Zhang L, Liu C, Wang B, Hu Y, Chen Q. Super-resolution reconstruction of infrared images based on a convolutional neural network with skip connections. *Opt Lasers Eng* 2021;146:106717.
- [41] Gatys LA, Ecker AS, Bethge M, Hertzmann A, Shechtman E. Controlling perceptual factors in neural style transfer. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2017. p. 3985–93.
- [42] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer; 2015. p. 234–41.
- [43] Han TY, Kim DH, Lee SH, Song BC. Infrared image super-resolution using auxiliary convolutional neural network and visible image under low-light conditions. *J Vis Commun Image Represent* 2018;51:191–200.
- [44] Zou Y, Zhang L, Chen Q, Wang B, Hu Y, Zhang Y. An infrared image super-resolution imaging algorithm based on auxiliary convolution neural network. In: *Optics Frontier Online 2020: Optics Imaging and Display*, vol. 11571. International Society for Optics and Photonics; 2020. 115711B.

- [45] He Z, Tang S, Yang J, Cao Y, Yang MY, Cao Y. Cascaded deep networks with multiple receptive fields for infrared image super-resolution. *IEEE Trans Circuits Syst Video Technol* 2018;29(8):2310–22.
- [46] Bavirisetti DP, Dhuli R. Fusion of infrared and visible sensor images based on anisotropic diffusion and karhunen-loeve transform. *IEEE Sens J* 2015;16(1):203–9.
- [47] Bavirisetti DP, Xiao G, Liu G. Multi-sensor image fusion based on fourth order partial differential equations. In: 2017 20th International conference on information fusion (Fusion). IEEE; 2017. p. 1–9.
- [48] Bavirisetti DP, Xiao G, Zhao J, Dhuli R, Liu G. Multi-scale guided image and video fusion: a fast and efficient approach. *Circuits, Systems, and Signal Processing* 2019;38(12):5576–605.
- [49] Naidu V. Image fusion technique using multi-resolution singular value decomposition. *Def Sci J* 2011;61(5):479.
- [50] Bavirisetti DP, Dhuli R. Two-scale image fusion of visible and infrared images using saliency detection. *Infrared Phys Technol* 2016;76:52–64.

Behind the Paper

## Deep learning in optical metrology: a review

Deep learning is currently leading to a paradigm shift from physics-based modeling to data-driven learning in optical metrology. In such a context, we present an overview of the current status and the latest progress of applying deep learning technologies in the field of optical metrology.

Published in Electrical & Electronic Engineering  
Mar 03, 2022



**Chao Zuo**

Professor, Nanjing University of Science and Technology

+ Follow

In 2016, the Google-owned artificial intelligence (AI) company DeepMind shocked the world by defeating Lee Se-dol four matches to one with its AlphaGo AI system, alerting the world to deep learning, a new breed of machine learning that promised to be smarter and more creative than before<sup>1</sup>. Since then, we have witnessed its rapid progress and extensive applications in solving many tasks in computer vision, computational imaging, and computer-aided diagnosis with unprecedented performance. Meanwhile, tech giants Google, Facebook, Microsoft, Apple, and Amazon have ignited the “art” of data manipulation and developed easy-to-use, open-source deep learning frameworks. These deep learning frameworks allow us to build complex and large-scale deep learning models using a collection of pre-built and optimized components in a more clear, concise, and user-friendly way, without getting into too many details of underlying algorithms. Deep learning has left the halls of academia very quickly and is ready to reshape an array of companies across multiple industries.

On the other hand, optical metrology is the science and technology of making measurements with use of light as standards or information carriers. Although optical metrology is a rapidly growing area, it is not a new discipline. The development of physical sciences has been driven from the very beginning by optical metrology techniques. In return, optical metrology has been revolutionized by the invention of laser, charged coupled device (CCD), and computer, developing into a broad and interdisciplinary field relating to diverse disciplines such as photomechanics, optical engineering, computer vision, and computational imaging. In light of the great success of deep learning in these related fields, researchers in optical metrology were unable to hold back their curiosities with regards to adopting this technology to further push the limits of optical metrology and provide new solutions in order to meet the upcoming challenges in the perpetual pursuit of higher accuracy, sensitivity, repeatability, efficiency, speed, and robustness. In this context, we incidentally become the “*first to eat crab*” research group (SCILab: [www.scilabora.com](http://www.scilabora.com))



[tory.com](http://www.tory.com)) of introducing deep learning to optical metrology.



**Fig. 1 | The rise of deep learning.** In 2016, the AlphaGo defeated Lee Se-dol four matches to one, alerting the world to deep learning. Meanwhile, tech giants ignited the “art” of data manipulation and developed easy-to-use, open-source deep learning frameworks.

### ***“The first to eat crab”***

For many phase measuring optical metrology techniques, including optical interferometry, digital holography, electronic speckle pattern interferometry, Moiré profilometry, and fringe projection profilometry, the physical quantities to be measured (such as the surface shape, displacement, strain, roughness, defects, etc.) are directly or indirectly encoded in the phase information of the fringes formed by means of interference or projection. Consequently, phase demodulation, which analyzes the quasi-periodic fringe pattern for the wrapped phase extraction, is the most critical step because the measurement accuracy of these optical metrology techniques depends directly on the phase demodulation accuracy of recorded fringe patterns. How to extract the phase information with the highest accuracy, fastest speed, and full automation remains a research hotspot in the field of optical metrology.

Traditional fringe pattern analysis, or phase demodulation techniques can be broadly classified into two categories: spatial and temporal techniques. Temporal phase demodulation techniques (representatively, the phase-shifting technique<sup>2</sup>) detect

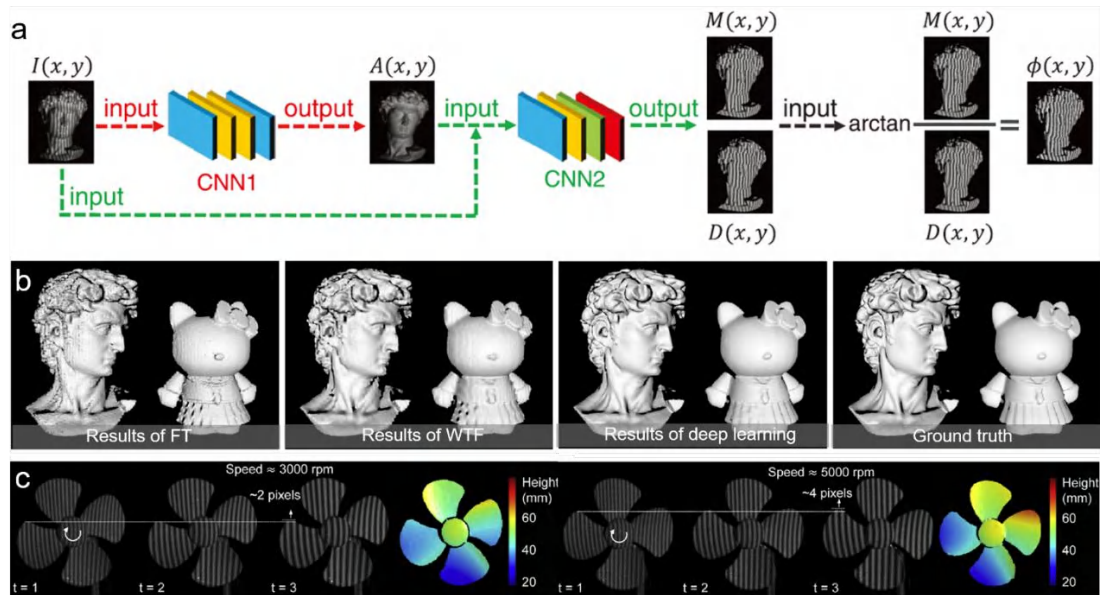
high-resolution pixel-wise phase distribution from the temporal variation of fringe signals at the cost of time-sequential data acquisitions. Spatial phase demodulation methods, such as Fourier transform (FT)<sup>3</sup> and windowed Fourier transform (WFT)<sup>4</sup>, are capable of estimating the phase distribution from a single fringe pattern, but fringe discontinuities and rich details of testing surfaces prevent them from high-accuracy phase measurement of complex surfaces. In this context, the goal of our first attempt is to develop a deep-learning-based fringe pattern analysis technique that is capable of combining the single-frame strength of spatial phase demodulation techniques with the high measurement accuracy of temporal phase demodulation techniques.

Initially, we tried to accomplish this goal by designing a deep convolutional neural network (CNN) with an end-to-end architecture, which directly links the input fringe image to the output phase map. However, it has been found that such an end-to-end learning scheme had difficulties in reproducing abrupt  $2\pi$  phase jumps in the wrapped phase map, making the training process fails to converge. After many twists and turns, we have finally devised a pragmatic and practical solution. Inspired by the physical model of conventional fringe analysis techniques, where the wrapped phase is calculated from the arctangent function, we attempted to predict the *sine* and *cosine* components of the fringe pattern from one input fringe image [Fig. 2a]. Encouragingly, such a strategy worked extremely well after appropriate network training, and can provide high-accuracy phase predictions close to those of the 12-step phase-shifting approach [Fig. 2b]. We further applied this learning-based fringe analysis technique to a high-speed fringe projection profilometry system, achieving an unprecedented 3D imaging frame rate up to 20,000 Hz [Fig. 2c]. This work was published in [Advanced Photonics](#) as a cover paper in 2019<sup>5</sup>, which now has become the most cited paper of [Advanced Photonics](#) since its inception.

### **“A logical hierarchy”**

At the beginning of 2020, the sudden COVID-19 raged and spread around the world. The New Year, an originally lively traditional Chinese festival, became unusually silent. During the long hours of leisure, I have noticed that researchers in optical metrology started actively participating in the explosively growing field of deep learning, as evidenced by the ever-increasing number of publications and exponentially growing citations to our earlier research. Within a few short years, deep-learning-based techniques have been gaining increasing attention and demonstrating promising performance in various optical metrology tasks, such as phase demodulation, phase unwrapping, system calibration, and error compensation. However, those research works are scattered rather than systematic. “Whether machine learning will be the driving force in optical metrology not only provides superior solutions to the growing new challenges but also tolerates imperfect measurement conditions with least efforts?” The answer to this key question deserves deeper thought and exploration. “Under these circumstances, a

comprehensive review that covers principles, implementations, advantages, applications, and challenges in utilizing deep learning for optical metrology tasks will be extremely useful.” The idea for a review article entitled “**Deep learning in optical metrology: a review**” was therefore born.



**Fig. 2 | Flowchart of fringe analysis using deep learning and the 3D reconstruction results of different approaches.** a Flowchart of fringe analysis using deep learning <sup>5</sup>. b Comparison of the 3D reconstructions of different fringe analysis approaches (FT <sup>3</sup>, WFT <sup>4</sup>, our deep learning-based method, and 12-step phase-shifting profilometry). c Measurement results of a desk fan rotating at different speeds using our method <sup>6</sup>.

On Mar 22, 2020, I said to my Ph.D. student Jiaming Qian, the primary co-author of this review article: “Let us together take an extensive and thorough examination of previously published relevant literature and create a compelling synthesis of gathered references.” Hundreds of deep learning papers on classical interferometry, digital holography, fringe projection profilometry, *etc.*, were then surveyed and categorized [Fig. 3]. In May, under the shadow of the epidemic, some senior Ph.D. students were allowed to return to the university on the premise of ensuring safety, including Jiaming. I enjoy using figures and tables to summarize research progress and suggest future research trajectories. So I drew a very sketchy optical metrology vein diagram for Jiaming: “Optical metrology covers a wide range of methods and applications today. It would be impractical for this review to discuss all the relevant technologies and trends.” My advice to Jiaming is to accept the fact that a review is different from a textbook: it should be more focused, and it’s OK to skip some topics so that it does not distract the readers.



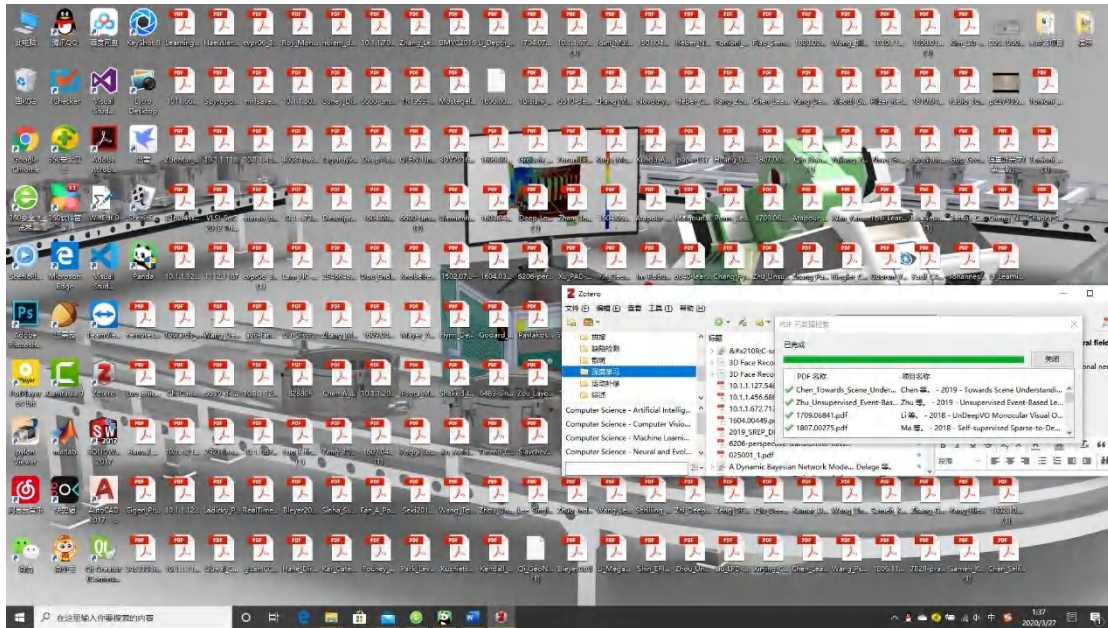


Fig. 3 | In the early hours of March 24, 2020, unorganized literatures on optical metrology using deep learning littered jiaming's computer screen. (zetero)

With this sketch in hand, Jiaming began to meticulously “carve” and “polish” it [Fig. 4]. Little did anyone predict that 19 months later a sketch would expand into a 54-page paper with double-column layout. Indeed, when it comes to optical metrology, ones subconsciously associate them with ‘fringe’ (phase measuring metrology) and ‘speckle’ (speckle metrology). We therefore restrict our focus to phase/correlation measurement techniques, such as interferometry, holography, fringe projection, and digital image correlation (DIC). Image processing plays an essential role in optical metrology, which is very similar to those of computer vision and computational imaging for the purpose of converting the observed measurements (intensity image in most cases) into the object quantities taking into account the physical model describing the image formation process. In most cases, image processing in optical metrology is not a one-step procedure, but *a logical hierarchy* consisting of three main steps, pre-processing, analysis and post-processing. Such a logical hierarchy provides a systematic framework throughout this review to classify and summarize the various optical metrology techniques and tasks that are otherwise fragmented. It also helps to decide what should and should not be included in the review.

### “A panoramic comparative picture”

After this, we proceed with the writing. We need to fill each layer of the hierarchical steps, which is essentially writing a mini-review of various types of algorithms distributed in different layers. I remind my students to imagine themselves as ‘artists of science’ and encourage them to develop how they write and present information. “Adding more words isn’t always the best way, we have to get something concisely

from a broad reading.”

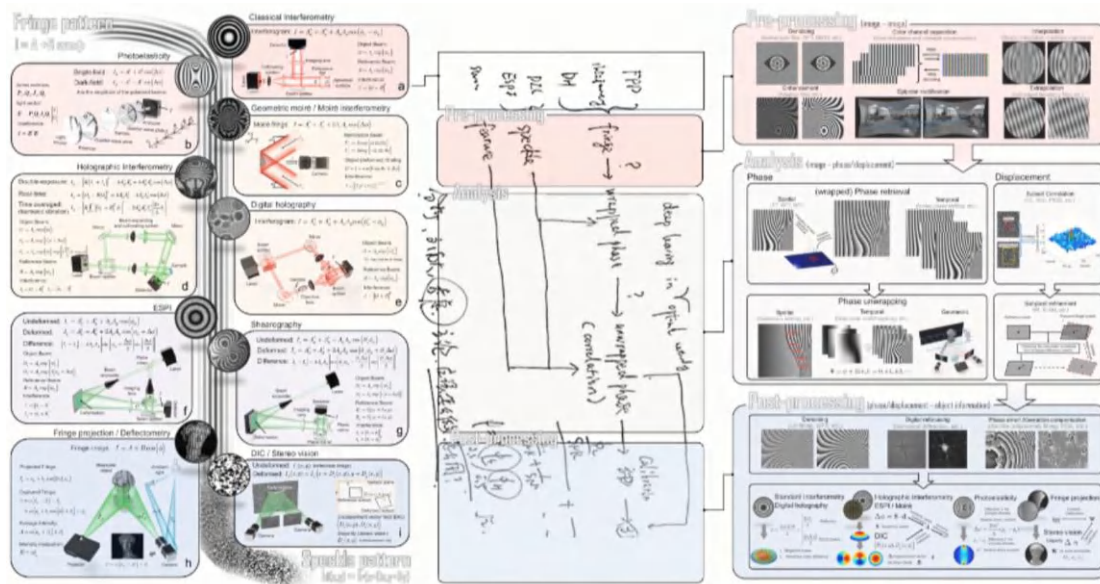


Fig. 4 | A summary of conventional optical metrology derived from a vein sketch.

Because of the significant changes that deep learning brings to the concept of optical metrology technology, almost all elementary tasks of digital image processing in optical metrology have been reformed by deep learning. This encouraged us to further summarize these existing researches leveraging deep learning in optical metrology according to a similar logical architecture [Fig. 5]. We went through multiple iterations to make sure that we had scanned the literature sufficiently and provided a clear, concise, appropriate, and informative review of conventional and new “learning-based” optical metrology techniques. Gradually, a clear and beautiful view unfolded before our eyes: the new deep learning-enabled optical metrology algorithms [Fig. 5] and their traditional counterparts [Fig. 4] echo each other, providing us with a *panoramic comparative picture*.

As a smooth transition between the old and the new, we gave a brief introduction to deep learning and summarized its threefold advantages from extensive literature: from “physics-model-driven” to “data-driven”, from “divide-and-conquer” to “end-to-end learning”, and from “solving ill-posed inverse problems” to “learning pseudo-inverse mapping”. In general, deep learning is particularly advantageous for many problems in optical metrology whose physical models are complicated and acquired information is limited. Strong empirical and experimental evidence suggests that using problem-specific deep learning models outperforms conventional knowledge or physical model-based approaches.

It is not enough for a review to be a summary of historical growth in the literature; it is also expected to provide a discussion about controversial issues in this field. In spite of the promising — in some cases impressive — results that have been

documented in the literature, *on the other side of the coin*, significant challenges do remain in this area. In collaboration with my colleagues, Assoc. Prof. Shijie Feng and Jing Han from Nanjing University of Science and Technology, and exchange student Pengfei Fan from Queen Mary University of London, the critical challenging issues of applying deep learning to optical metrology were gathered and discussed:

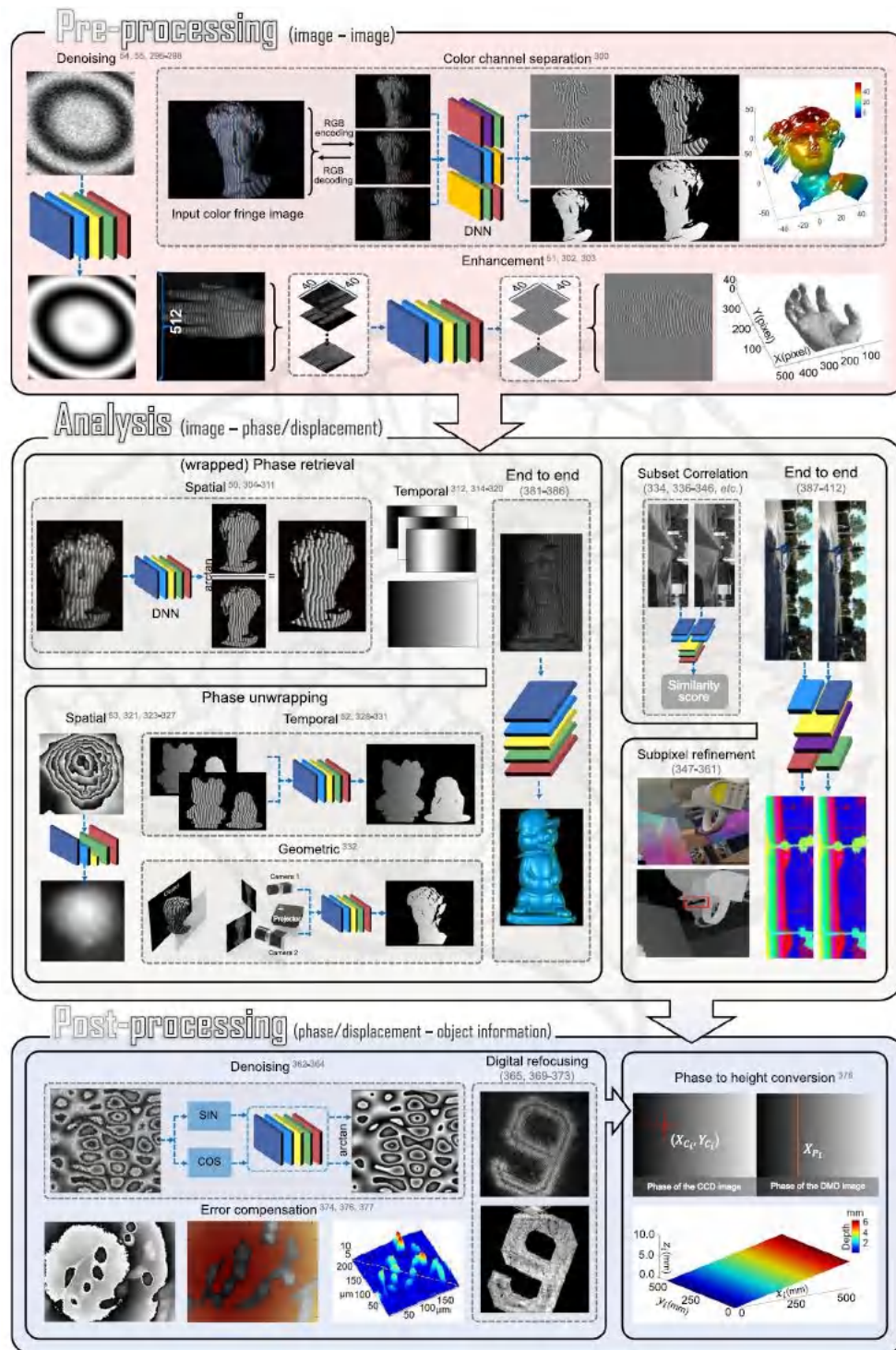
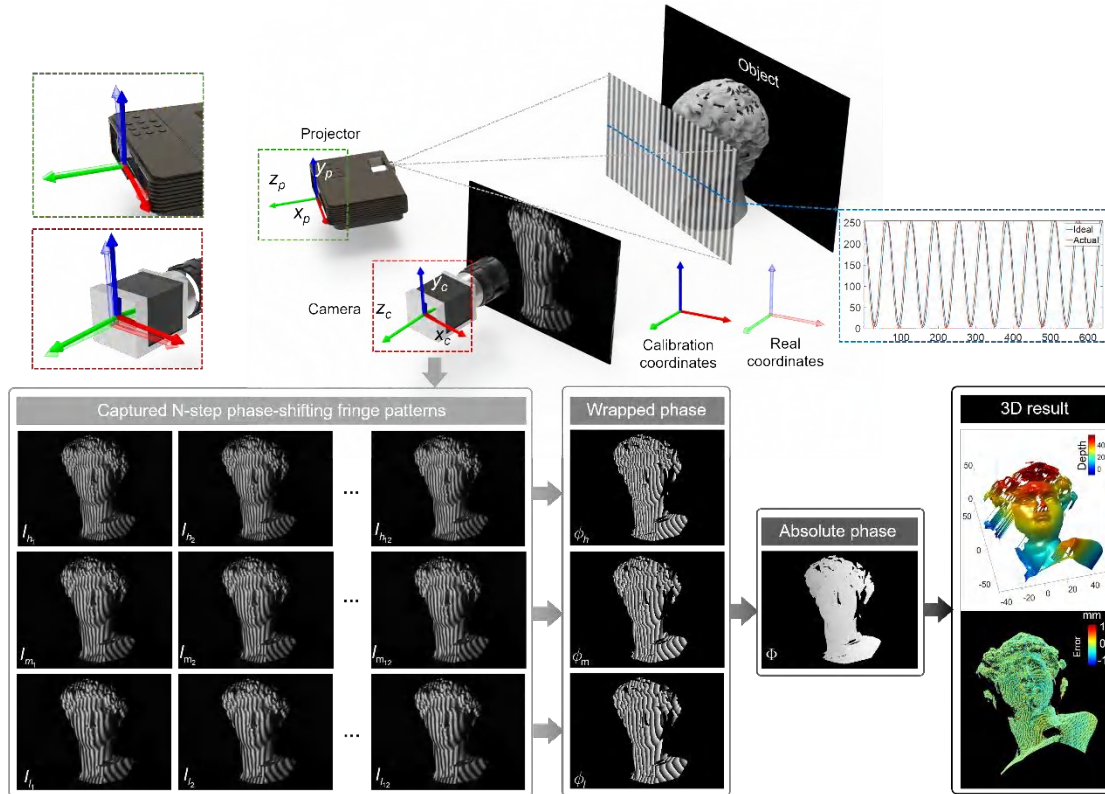


Fig. 5 | Repeatedly modified overview graphs of deep learning in optical metrology.



## “The other side of the coin”

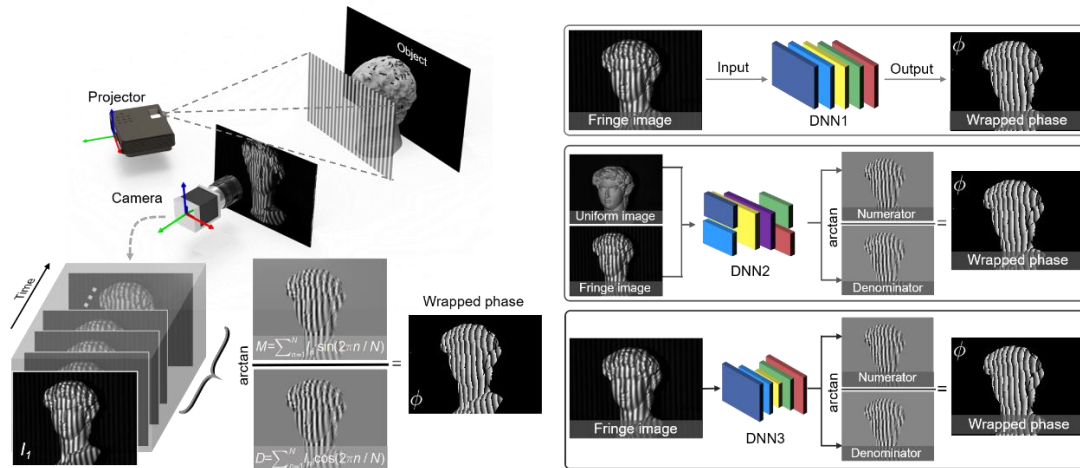
- For model training, we need to acquire large amounts of experimental data with labels, which, even if available, is laborious and requires professional experts [Fig. 6].



**Fig. 6 | The challenge of deep learning in optical metrology—the high cost of obtaining and labeling training data.** Taking fringe projection profilometry as an example, to collect high-quality training data, the traditional multi-frequency temporal method is used, which causes a large number of images to be projected for each set of training data. However, hardware errors, ambient light interference, calibration errors, etc. in actual operation make it difficult to obtain ideal ground truth through traditional algorithms

- So far, there has been no theoretical groundwork that could clearly explain the mechanisms to optimize network structure for a specific task or profoundly comprehend why a particular network structure is effective in a given task or not [Fig. 7].
- Generally, deep learning architectures used in optical metrology are highly specialized to a specific domain, and they should be implemented with extreme care and caution when solving issues that do not pertain to the same domain.
- Deep learning approaches have often been regarded as ‘black boxes’, and in optical metrology, accountability is essential and can cause severe consequences.
- Since the information cannot be “born out of nothing”, deep learning cannot

always produce a provably correct solution. The success of deep learning depends on the “common” features learned and extracted from the training samples, which may lead to unsatisfactory results when facing “rare samples”.



**Fig. 7 | The challenge of deep learning in optical metrology—empiricism in model design and algorithms selection.** Taking the phase extraction in fringe projection profilometry as an example, the same task can be implemented by different neural network models with different strategies: the fringe image can be mapped directly to the corresponding phase map by DNN1; we can also output the numerator and denominator of the arctangent function used to calculate the phase information from a fringe image and a uniform by DNN2; with more powerful DNN, we can predict from a fringe image the numerator and denominator

Listed above are among the most critical issues for optical metrology applications where the accuracy, reliability, repeatability, and traceability of measurement results are primary considerations. After identifying the research gaps, we hope the review paper should leave the reader with ***explicit opinions on its future trajectory***. After another round of brainstorming, we made the following suggestions for potential new research areas to explore in the future.

- Leveraging more emerging technologies of deep learning methods to optical metrology could promote and accelerate the recognition and acceptance of deep learning in more application areas.
- Combining Bayesian statistics with deep neuron networks to obtain quantitative uncertainty estimates allows us to assess when the network yields unreliable predictions (see our recent *Optica* paper on this point) <sup>7</sup>.
- A synergy of the physics-based models that describe the a priori knowledge of the image formation and data-driven models that learn a regularization from the experimental data can bring our domain expertise into deep learning to provide more physically plausible solutions to specific optical metrology problems.

We believe the above-mentioned aspects can provide inspiration for future scopes and continue to attract the interest of deep learning research in the optical metrology community in the years to come. Finally, we would also like to remind readers that the selection between deep learning and traditional algorithms should be considered rationally, given the “no free lunch theorem”. For several problems where traditional methods based on physics models, if implemented properly, can deliver straightforward and more than satisfying solutions, there is no need to use deep learning.

### ***“Revise, revise, revise”***

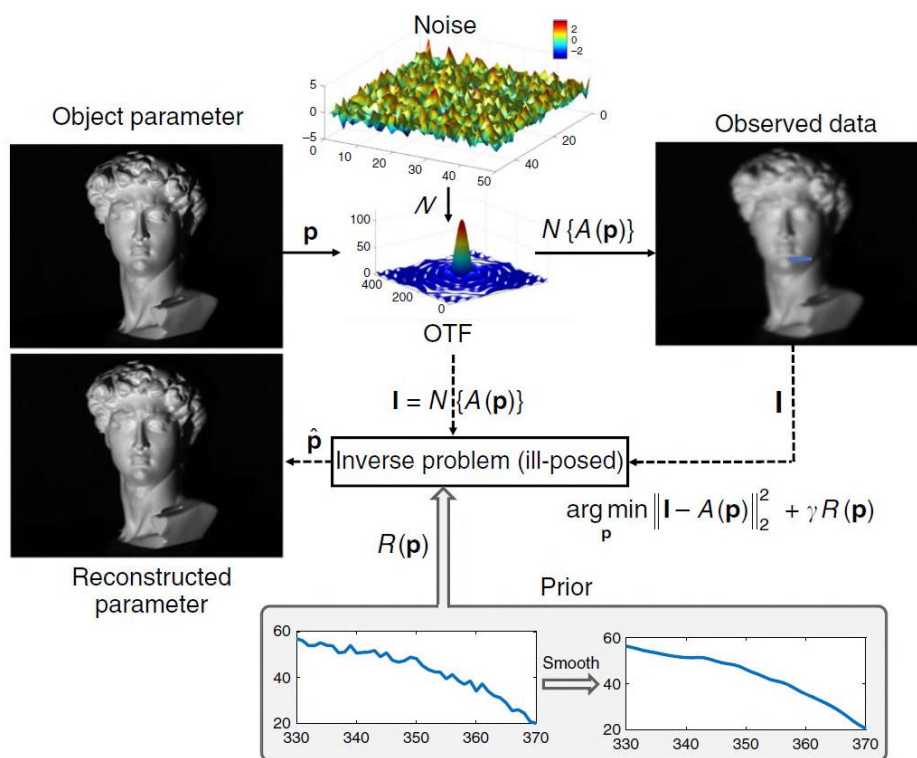
In Jan 2021, we finished the first draft and decided to submit it to *Light: Science & Applications*, which publishes original articles and reviews of high quality, high interest, and far-reaching impact. My Ph.D. student, Yixuan Li, double-checked the typesetting, grammar, and references according to the journal’s stylistic and formatting guidelines. After that, I consulted Prof. Kemao Qian at Nanyang Technological University (NTU), Singapore, an expert with more than 20 years of experience in optical metrology, to review our draft. Prof. Qian was my (unofficial) Ph.D. advisor when I was a visiting student at NTU from Sep 2012 and Feb 2014. My intuition told me that getting his perspective would be very helpful in enhancing the quality of this review.

After assessing the first draft, Prof. Qian offered three constructive criticisms. (1) The section “brief introduction to deep learning” only introduced the history of deep learning and artificial neural network, and had little relevance to optical metrology. Instead, readers should be interested in learning more about how deep learning can be used in optical metrology from this section. (2) The transition between sections of traditional optical metrology algorithms and their deep learning-enabled counterparts was sudden and there was insufficient evidence as to why deep learning should be used in optical metrology. (3) As a review, its main purpose is to help other researchers enter this field more easily by collecting and summarizing, synthesizing, and analyzing existing research. ***“Will deep learning be the future of optical metrology?” It is very difficult to draw a conclusion at the current stage, as the place of deep learning in optical metrology is not yet clear.*** So instead of giving a clear answer to this question, we try our best to paint a full and informative picture for our readers.

*“When you draft, you write for yourself. When you revise, you clarify for your readers.”* Prof. Qian’s advice epitomized this motto. With these criticisms in hand, the only thing we could do was ***revise, revise, revise!*** ***“We must strive for excellence.”*** I encouraged Jiaming. It had now become apparent that spending nearly half a year revising is a wise choice, because this made this review not only more in-depth in content but also more logically developed, from beginning to end.



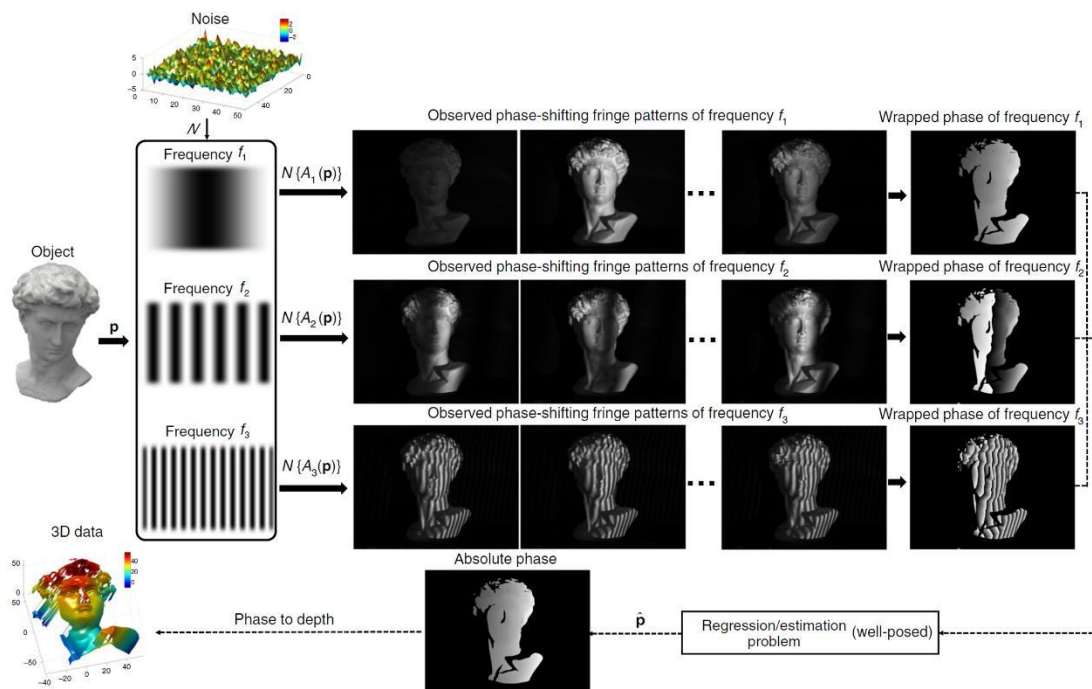
- For the first comment, we added more introduction to the fundamentals of deep learning, including the basic principles of neural networks, network structures, and training algorithms. We focused on the dominant network architecture for image- and vision-related tasks—convolutional neural networks (CNNs), and then discussed in detail the variants of classical CNN architecture—DNNs with a fully convolutional architecture, that shares characteristics with image processing algorithms in optical metrology (transforming the content of arbitrary-sized inputs into pixel-level outputs). By applying different types of training datasets, they can be trained for accomplishing different types of image processing tasks that we encountered in optical metrology. This provides an alternative approach to process images such that the produced results resemble or even outperform conventional image processing operators or their combinations. There are also many other potential desirable factors for such a substitution, *e.g.*, accuracy, speed, generality, and simplicity. All these factors were crucial to enable the fast rise of deep learning in the field of optical metrology.



**Fig. 8 | Inverse problems in computer vision.** In computer vision, such as image deblurring, the resulting inverse problem is ill-posed since the forward measurement operator mapping from the parameter space to the image space is usually poorly conditioned. The classical approach is to impose certain prior assumptions (smoothing) about the solution that helps in regularizing its retrieval

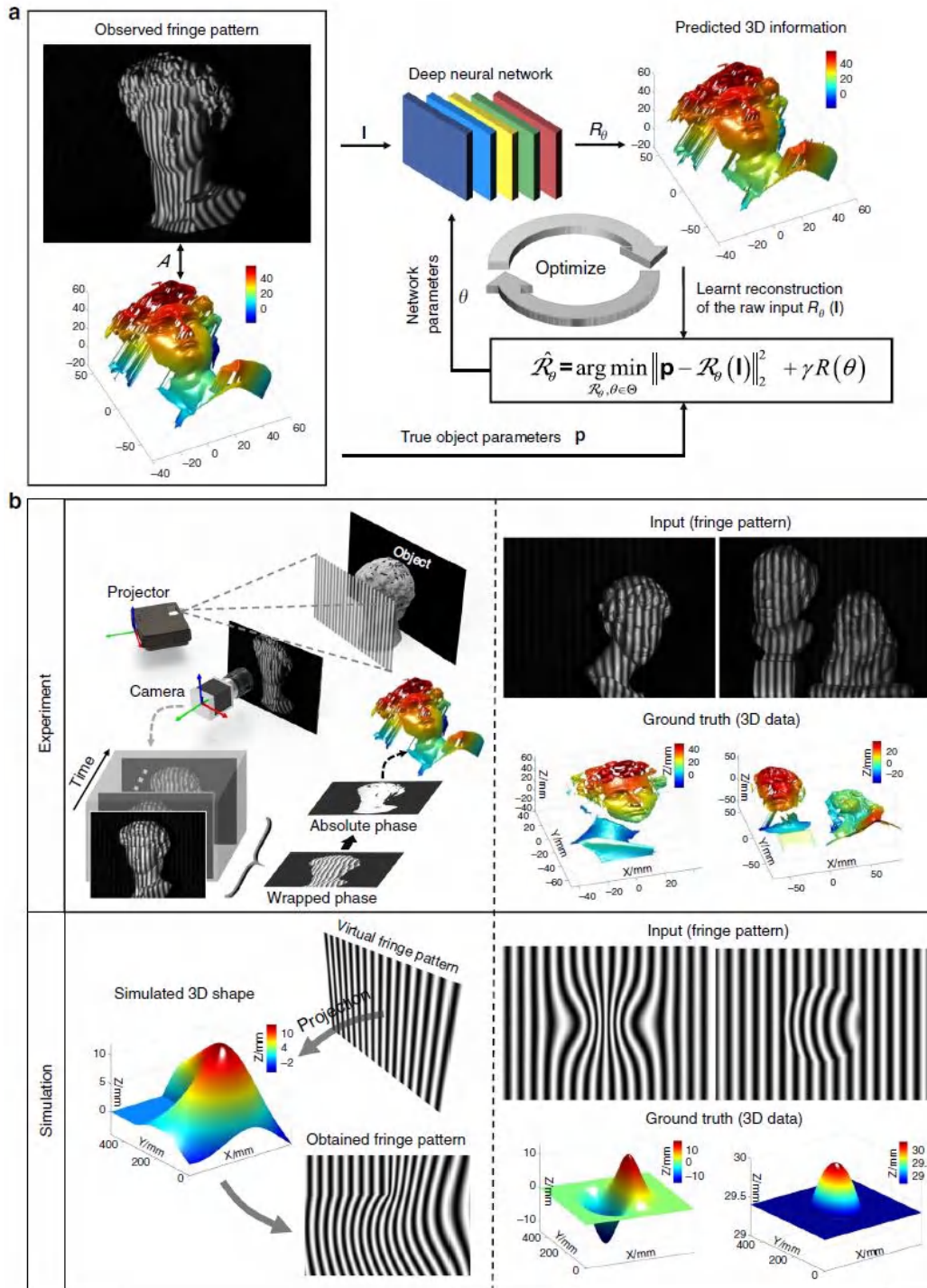
- For the second comment, we tried to explain the reason for the transition from the perspective of solving inverse problems. In optical metrology, we

have to conclude in general from the effect (i.e., the intensity at the pixel) to its cause (i.e., shape, displacement, deformation, or stress of the surface). Such information recovery process is similar to those of computer vision and computational imaging, presenting as an inverse problem that is often ill-posed. Tremendous progress has been achieved in terms of accurate mathematical modeling, regularization techniques, numerical methods, and their efficient implementations [Fig. 8]. For optical metrology, the situation becomes quite different due to the fact that the optical measurements are frequently carried out in a highly controlled environment. Instead of explicitly interpreting optical metrology tasks from the perspective of solving inverse problems, we prefer to reformulate the original ill-posed problem into a well-posed and adequately stable one by actively controlling the image acquisition process [Fig. 9]. However, for many challenging applications, harsh operating conditions may make such active strategies a luxurious or even unreasonable request. Under such conditions, deep learning is particularly advantageous for solving those optical metrology problems because the active strategies are shifted from the actual measurement stage to the preparation (network training) stage, and the “reconstruction algorithm” can be directly learned from the experimental data [Fig. 10]. If the training data is collected under the environment that reproduces the real experimental conditions, and the amount of data is sufficient, the trained model should reflect the reality more precisely and comprehensively, and is expected to produce better reconstruction results than conventional physics-model-based approaches.



**Fig. 9 | Inverse problems in optical metrology.** Optical metrology uses an “active” approach to transform the ill-posed inverse problem into a well-posed estimation or regression problem: by

acquiring additional phase-shifted patterns of different frequencies, absolute phase can be easily determined by multi-frequency phase-shifting and temporal phase unwrapping methods



**Fig. 10 | Deep learning-based optical metrology as a constraint optimization problem.** a In deep learning-based optical metrology, a set of true object parameters and the corresponding raw measured data are created at the training stage, and their mapping relation (learn a reconstruction algorithm) is established by training a deep neural network with all network parameters (neural network weights) learned from the dataset. b The principle of obtaining the dataset by real experiments or simulations with the knowledge of the forward model (left) and the obtained dataset (right)



Many great writers have commented on the importance of revision. William Zinsser said, “*Revision is the essence of good writing: it’s where the game is won or lost.*” Stephen King said, “*You need to revise for length. The formula. Second draft = first draft – 10%.*” I would have thought revision does not necessarily mean rewriting the whole paper, and a 10% revision was a lot. But after going through countless iterations, more than 1/3 content of the draft had been refreshed. This considerable improvement was mainly attributed to the careful and insightful guidance of Prof. Qian, who was far away in Singapore, devoting significant efforts to revise the manuscript with us round by round, through countless hours of the video conference [Fig. 11].

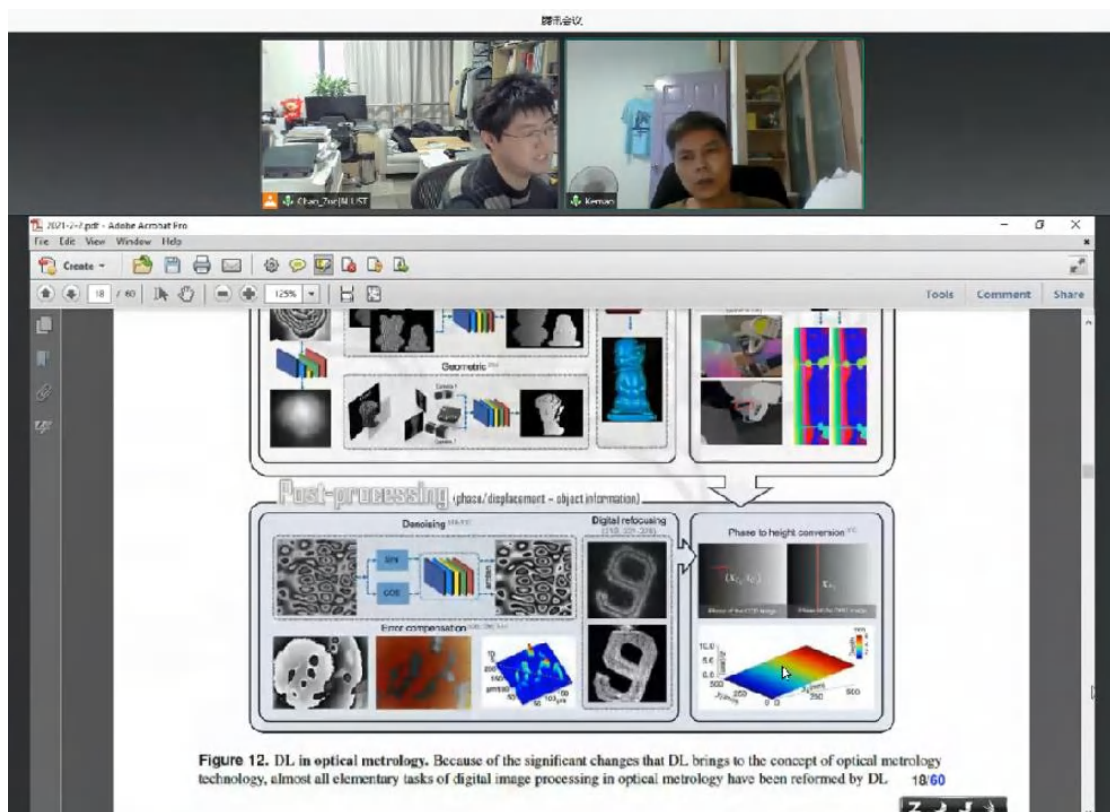


Fig. 11 | Professor Qian was revising the manuscript with Professor Zuo through the video conference.

### “Be the ‘go to’-reference”

In July 2021, we finalized the manuscript, which was carefully read, checked and approved by all co-authors. We submitted it to *Light: Science & Applications* as planned, along with a cover letter to the editor. The peer-review report came back after two months, and, encouragingly, all three reviewers gave very positive comments on our manuscript. The overall assessments of the three reviewers are provided as follows:

- Reviewer 1: “*This is a pretty comprehensive review paper for deep learning in*

*optical metrology. They start from introducing conventional methods in optical metrology and measurement processing, and then go into tutorial of deep learning and how deep learning is applied. I find that this paper is timely and can be considered for publication pending that the following comments are addressed.....”*

- Reviewer 2: *“The review article on Deep Learning in Optical Metrology is an excellent manuscript which is well written.....The authors have given a very good introduction to various optical metrological methods such as interferometry, holography, fringe projection, and DIC. The development of these methods through the past, their basics have been well explained. The figures presented have been well organized giving the reader a good comparative picture. ....The manuscript gives a comprehensive review of optical metrology techniques and how deep learning can be tailored. The presentation details are well handled. The manuscript deserves publication as such.”*
- Reviewer 3: *“The authors present us with a review paper in the field of deep learning applied to optical metrology. It is a very long manuscript with more than 40 pages of text. However, it is well written and a pleasure to read. As someone who is working in optical metrology for more than 20 years, it was a very good introduction into the basic concepts of deep learning in this field, and I learned a lot.....”*

The main constructive revision comments came from Reviewer 3, who suggested to give a simple but very detailed example on how to apply deep learning to optical metrology:

Reviewer 3: *“The manuscript is clearly a beautiful review paper. BUT, with some small changes it can even become so much more. I would really encourage the authors to give a simple but very detailed example on how to apply deep learning to optical metrology (including the math, the algorithmic implementation, etc.). This could be for example denoising or signal-reconstruction from corrupted data. This could evolve the paper into one of the fundamental “go to”-references if it comes to optical metrology and deep learning, because it is not only a good overview but also a basic tutorial. You could do this at the expense of shortening the first part (see above), because nobody really wants to read 5 to 10 pages of textbook knowledge about phase shifting, phase unwrapping, etc., but everybody wants to LEARN how to apply these new data-driven approaches.”*

Finally, the reviewer concluded:

*“I believe that this manuscript has big potential to become **one of the “go to”-references in deep learning for optical metrology**. It is well written, comes at the right time and includes numerous examples. However, I would strongly encourage the authors to consider the above comments and I would like to re-review it.”*

I deeply appreciated that the reviewers were here to help our paper succeed and by following their advice, finally we would emerge with a stronger version that will

hopefully end up becoming the definitive “go-to” guide on this topic. We were inspired to further include *a tutorial of applying deep learning to optical metrology* in the Supplementary Information, taking phase demodulation from a single fringe pattern as an example. In addition, we published the source codes and the corresponding datasets for this example. We demonstrate that a well-trained deep neural network can accomplish the phase demodulation task in an accurate and efficient manner, using only a single fringe pattern. Thus, it is capable of combining the single-frame strength of the spatial phase demodulation methods with the high measurement accuracy of the temporal phase demodulation methods. This turned out to be a considerable addition because it made the review more comprehensive and instructive, potentially increasing its readership. It should be noted that it took approximately four months to complete the processes of peer review, revision, and publication. During this period, many new papers and even competing reviews were published. To provide the most up-to-date review, we had to stay abreast of the literature by using Google Scholar, which alerted me daily updates of relevant literature based on keywords.

### ***“Predict Engage the future”***

Finally, let us return to the third point raised by Prof. Qian *“Is predicting the future futile or necessary?”* Undoubtedly, deep learning is currently prompting increasing interests and leading to a paradigm shift from physics- and knowledge-based modeling to data-driven learning in optical metrology. Strong empirical and experimental evidence suggests that using problem-specific deep learning models outperforms conventional knowledge or physical model-based approaches, especially for many optical metrology tasks whose physical models are complicated and acquired information is limited.

It has to be admitted that deep learning is still at the early stage of development for its applications in optical metrology. Many researchers are still skeptical and maintain a wait-and-see attitude towards its applications involving industrial inspection and medical care, etc. Shall we accept deep learning as the key problem-solving tool? Or should we reject such a black-box solution? These are controversial issues in the optical metrology community today. Looking on the bright side, it has promoted an exciting trend and fostered expectations of the transformative potential it may bring to the optical metrology society. However, we should not overestimate the power of deep learning by considering it as a silver bullet for every challenge encountered in the future development of optical metrology. In practice, we should assess whether the large amount of data and computational resources required to use deep learning for a particular task is worthwhile, especially when other conventional algorithms may yield comparable performance with lower complexity and higher interpretability. “Will deep learning replace the role of traditional technologies within the field of optical metrology for the years to come?”



“It is clear no one can predict the future, but we can engage it. If you are still an ‘outsider’ or new to this field. I encourage you to try it out! *It is easy, and often works!*”

---

For more information, please refer to this recent publication: Zuo, C., Qian, J., Feng, S. et al. Deep learning in optical metrology: a review. *Light Sci Appl* **11**, 39 (2022).

DOI: <https://doi.org/10.1038/s41377-022-00714-x>

## Reference

1. Chen, J. X. The evolution of computing: AlphaGo. *Comput. Sci. Eng.* **18**, 4–7 (2016).
2. Zuo, C., Feng, S., Huang, L., Tao, T., Yin, W., & Chen, Q., A. Phase shifting algorithms for fringe projection profilometry: A review. *Opt. Lasers Eng.* **109**, 23–59 (2018).
3. Takeda, M., Ina, H. & Kobayashi, S. Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry. *J Opt Soc Am A* **72**, 156–160 (1982).
4. Kemao, Q. Windowed Fourier transform for fringe pattern analysis. *Appl. Opt.* **43**, 2695–2702 (2004).
5. Feng, S., Chen, Q., Gu, G., Tao, T., Zhang, L., Hu, Y., Yin, W., Zuo, C., Fringe pattern analysis using deep learning. *Adv. Photon.* **1**, 025001 (2019).
6. Feng, S., Zuo, C., Yin, W., Gu, G. & Chen, Q. Micro deep learning profilometry for high-speed 3D surface imaging. *Opt. Lasers Eng.* **121**, 416–427 (2019).
7. Feng, S., Zuo, C., Hu, Y., Li, Y. & Chen, Q. Deep-learning-based fringe-pattern analysis with uncertainty estimation. *Optica* **8**, 1507–1510 (2021).

NEWS & VIEWS

Open Access

# Optical metrology embraces deep learning: keeping an open mind

Bing Pan<sup>1</sup>✉

## Abstract

Optical metrology practitioners ought to embrace deep learning with an open mind, while devote continuing efforts to look for its theoretical groundwork and maintain an awareness of its limits.

Optical metrology is the science and technology concerning measurements using light. The development of physical sciences has been driven from the very beginning by optical metrology techniques. In return, optical metrology has been revolutionized by several major inventions of physical sciences, such as the laser, charged coupled device (CCD), and computer technology. Although optical metrology technologies have developed into problem-solving backbones in many science and engineering applications, they have already implemented the transition to their digital avatars for nearly half a century, entering an era of diminishing returns. After the three previous revolutions brought about by the laser, imaging sensor, and digital computing, which technology will reinvigorate optical metrology?

Deep learning is a type of machine learning that uses artificial neural networks to learn a mapping between input and output data<sup>1</sup>. Once trained, these models can predict outputs from supplied input data. In 2016, AlphaGo beating Lee Sedol, the best human player at Go, four matches to one was a truly seminal event in the history of machine learning and deep learning. Since then, we have witnessed its explosive growing and extensive applications in solving many tasks in computer vision, computational imaging, and computer-aided diagnosis<sup>2</sup>. In light of the tremendous success of deep learning in these related fields, researchers in optical metrology were unable to hold back their curiosities with regards to

adopting this technology to further push the limits of optical metrology and provide new solutions in order to meet the upcoming challenges in the perpetual pursuit of higher accuracy, sensitivity, repeatability, efficiency, speed, and robustness.

The research group led by Prof. C. Zuo at Nanjing University of Science and Technology is a pioneer in introducing deep learning to optical metrology with a particular focus on fringe pattern analysis and fringe projection profilometry. In 2019, they developed a deep-learning-based fringe pattern analysis technique capable of combining the single-frame strength of spatial phase demodulation techniques with the high measurement accuracy of multi-frame phase-shifting techniques<sup>3</sup>. The network is trained on single fringe pattern matched with the label phase (ground-truth) reconstructed by the standard 12-step phase-shifting algorithm of the same sample. After training based on extensive dataset, the neural network can transform a single fringe pattern into an accurate phase map from that almost reproduces the result of the 12-step phase-shifting method, which is an astonishing feat for the field. Subsequently, researchers in optical metrology started actively tilling this fertile field, as evidenced by the ever-increasing number of publications. Within a few short years, deep learning has been applied to various tasks of optical metrology, such as fringe denoising<sup>4</sup>, phase unwrapping<sup>5,6</sup>, and single-shot profilometry<sup>7,8</sup>.

In a recent issue of *Light: Science & Applications*, Zuo et al.<sup>9</sup> presented a comprehensive review in the field of deep learning applied to optical metrology. They start from conventional methods and typical signal processing

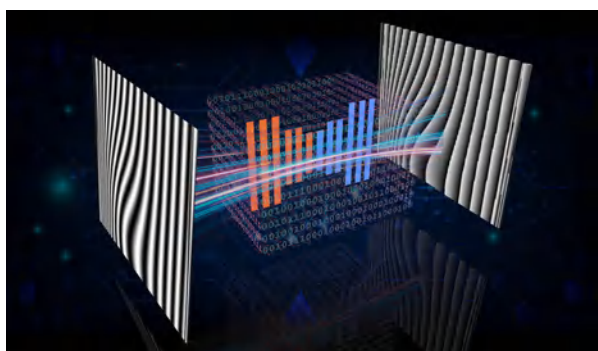
Correspondence: Bing Pan (panb@buaa.edu.cn)

<sup>1</sup>School of Aeronautic Science and Engineering, Beihang University, 100191 Beijing, China

© The Author(s) 2022



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.



**Fig. 1 The pros and cons of applying deep learning to optical metrology.** In deep-learning-based fringe analysis, a well-trained neural network can transform a single fringe pattern into an accurate phase map from that almost reproduces the result of the multi-step phase-shifting method, which is an astonishing feat for the field. But its internal mechanism tends to be very difficult to explain (“Black Box problem”)

tasks in optical metrology, and then introduce the idea of data-driven evaluation with deep learning. As a smooth transition between the old and the new, they provided a brief introduction to deep learning and summarized the threefold advantages of its application to optical metrology: from “physics-model-driven” to “data-driven”, from “divide-and-conquer” to “end-to-end learning”, and from “solving ill-posed inverse problems” to “learning pseudo-inverse mapping”. Then a comprehensive overview where deep learning has already infiltrated almost every aspect of image processing tasks is presented, suggesting a paradigm shift from physics-based modeling to data-driven learning in optical metrology. The panoramic comparative picture reveals that using problem-specific deep learning models outperforms conventional knowledge or physical model-based approaches in most cases, especially for many optical metrology tasks whose physical models are complicated and acquired information is limited.

While promising, in many cases pretty impressive, results have been documented in the literature, Zuo et al. admitted that these works still represent early days in the application of deep learning to optical metrology. It is sensible to maintain a clear head and recognize that deep learning is not magic: it is essentially the process of using computers to help us find patterns within data. Since the information cannot “born out of nothing”, deep neural networks are usually pretty brittle, i.e., if we do not feed in the RIGHT kind of data in the RIGHT kind of format using the RIGHT kind of network model and training algorithm, we will get poor results.

In many applications of computer vision, people are always happy when the result looks good and realistic, no matter whether it is interpretable and quantifiable or not. However, adhering to the famous creed by Galileo:

“Measure what is measurable, and make measurable when it is so”, practitioners in optical metrology is both open-minded and rigorous. In optical metrology, it is not only necessary to get a good-looking result, but also need to make sure that the result is accurate, reliable, repeatable, and traceable. Though we hope that such deep learning approaches always have a provably correct solution, no one can guarantee, at least not yet. Another well-known disadvantage of deep neural networks is their “black box” nature. Simply put, we do not know how or why the network came up with a certain output (Fig. 1). However, the interpretability is critical to optical metrology, as it allows us trust the methodology and understand the causes of mistakes. Shall we accept deep learning as the key problem-solving tool? Or should we reject such a black-box solution? These are controversial issues in the optical metrology community nowadays.

Conjuring more from less must pay a price. There are still significant challenges in deep learning-based optical metrology: First, for model training, we need to acquire large amounts of experimental data with labels, which, even if available, is laborious and requires professional experts. Second, we need to look for the theoretical groundwork that would clearly explain the ways to define the optimal network structure and comprehend the reasons for its success or failure. Third, we should recognize that the power of deep learning approaches often comes at the expense of generalization (the ability to deal with never-before-experienced situations), i.e., their performance can be system, environment, and even sample dependent. Nevertheless, the progress of science comes from the continuous exploration to solve the unknown. So, I encourage optical metrology practitioners to embrace deep learning with an open mind while maintain an awareness of its limits.

All in all, nothing in deep learning-based optical metrology is to be feared. It is only to be understood and quantified. Recent research into Bayesian deep learning promises to assess the reliability of the network by explicitly quantifying uncertainty, which provides us an additional choice between “trusting the network without doubts” and “denying it completely”, namely “trusting it conditionally”<sup>10</sup>. As the emerging field slowly matures, deep learning is expected to graduate from black-box empirical representations to full-blown theoretical foundations, with a more profound impact not only on optical metrology, but also on optics and photonics as a whole.

Published online: 17 May 2022

#### References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).




2. Falk, T. et al. U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Methods* **16**, 67–70 (2019).
3. Feng, S. J. et al. Fringe pattern analysis using deep learning. *Adv. Photonics* **1**, 025001 (2019).
4. Yan, K. T. et al. Fringe pattern denoising based on deep learning. *Opt. Commun.* **437**, 148–152 (2019).
5. Wang, K. Q. et al. One-step robust deep learning phase unwrapping. *Opt. Express* **27**, 15100–15115 (2019).
6. Yin, W. et al. Temporal phase unwrapping using deep learning. *Sci. Rep.* **9**, 20175 (2019).
7. Qian, J. M. et al. Deep-learning-enabled geometric constraints and phase unwrapping for single-shot absolute 3D shape measurement. *APL Photonics* **5**, 046105 (2020).
8. Li, Y. X. et al. Deep-learning-enabled dual-frequency composite fringe projection profilometry for single-shot absolute 3D shape measurement. *Opto-Electron. Adv.* **5**, 210021 (2022).
9. Zuo, C. et al. Deep learning in optical metrology: a review. *Light Sci. Appl.* **11**, 39 (2022).
10. Feng, S. J. et al. Deep-learning-based fringe-pattern analysis with uncertainty estimation. *Optica* **8**, 1507–1510 (2021).

NEWS & VIEWS

Open Access

# Exploiting optical degrees of freedom for information multiplexing in diffractive neural networks

Chao Zuo <sup>1,2,3</sup>✉ and Qian Chen<sup>3</sup>

## Abstract

Exploiting internal degrees of freedom of light, such as polarization, provides efficient ways to scale the capacity of optical diffractive computing, which may ultimately lead to high-throughput, multifunctional all-optical diffractive processors that can execute a diverse range of tasks in parallel.

In the last decades, artificial intelligence (AI) technologies, especially artificial neural networks (ANNs), have led to a revolution in a range of applications, including autonomous driving, remote sensing, medical diagnosis, natural language processing, and the Internet of Things. However, the rapid progress of AI and the increasing scale of ANNs are actually accompanied with a tremendous amount of computational resources and energy costs<sup>1</sup>. The main reason behind this is that the dominant computational algorithm for ANNs consists of a large number of matrix-vector multiplications, which are typically the most computationally-intensive operations with the computing cost scales as the square of the input dimension<sup>2</sup>. Optical neural networks (ONNs) built using optical matrix-vector multipliers are promising candidates for next-generation neuromorphic computation, because they offer a potential solution to the energy consumption problem faced by their electrical counterparts<sup>3</sup>. In addition, the constituent scalar multiplication operations can be performed in parallel completely in the optical domain, at

the speed of light, and with zero energy consumption in principle<sup>4</sup>.

Optical computing, or more specifically ONNs, where people seek to perform neuromorphic computation with optics, is, in fact, not a new idea. In 1987, Mostafa and Psaltis<sup>5</sup>, for the first time, focused on the need and practical implementation of optical neural computers. Taking inspiration from the distributed topology of the brain, they created a physical implementation of neural networks by arranging optical components in the way neurons are arranged in the human brain. Since then, research in optical neuromorphic computing has flourished, spanning decades of development efforts on various novel optical implementations of neural networks<sup>6</sup>. But until recently experimental implementations of large-scale, highly parallel, high-speed, and trainable ONNs have been made with the breakthroughs in deep learning, optoelectronics, and photonic material engineering, leading to a resurgence of interest in this area.

ONNs are usually built based on an optical architecture that is mathematically described as an input-output function, i.e., a scattering matrix relating the input to the output electric field. And this naturally implements a matrix-vector multiplication, which can be realized by a diverse set of optical architectures, including integrated silicon photonic neuromorphic circuits<sup>7</sup>, fiber-optic sensor arrays<sup>8</sup>, and convolutional networks through diffractive optics<sup>9–11</sup>. Introduced by Ozcan Research Group

Correspondence: Chao Zuo ([zuochao@njjust.edu.cn](mailto:zuochao@njjust.edu.cn))

<sup>1</sup>Smart Computational Imaging Laboratory (SCILab), School of Electronic and Optical Engineering, Nanjing University of Science and Technology, 210094 Nanjing, Jiangsu Province, China

<sup>2</sup>Smart Computational Imaging Research Institute (SCIRI) of Nanjing University of Science and Technology, 210019 Nanjing, Jiangsu Province, China  
Full list of author information is available at the end of the article

© The Author(s) 2022



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

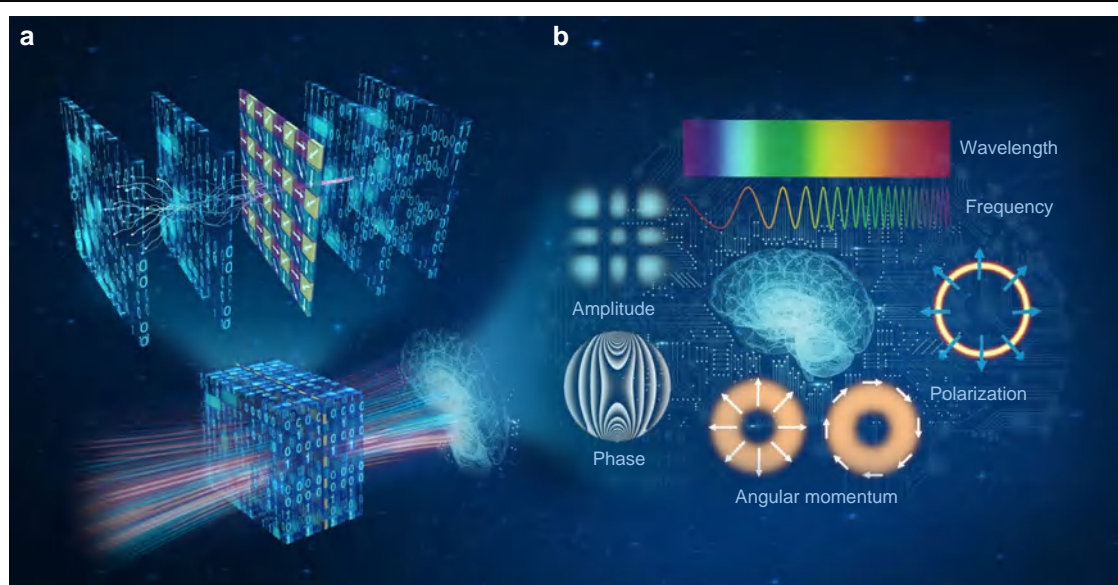
at the University of California, Los Angeles (UCLA), ONNs formed through the integration of successive spatially engineered transmissive diffractive layers, i.e., diffractive neural networks, have been demonstrated to enable both statistical inference and optical information processing, such as image classification<sup>9</sup>, single-pixel image reconstruction<sup>12</sup>, quantitative phase imaging<sup>13</sup>, and imaging through random diffusers<sup>14</sup>.

The diffractive neural network has its roots in Fourier optics, wherein a simple positive lens applies a physical two-dimensional Fourier transform to the wave field, and the prevalent wave propagation is described by Kirchhoff's diffraction integral that amounts to a convolution of the field with the impulse response of free space. These operations provide basic building blocks of convolutional neural networks (CNNs), making diffractive neural networks well-suited for most vision computing applications. By leveraging the light-matter interaction as an implementation of element-wise multiplication, the “pixels” on the diffractive surfaces embody the “neurons” on the network layers, which are interconnected by the physics of optical diffraction. As an analogy to standard neural networks, the complex-valued transmission coefficient (including amplitude and phase) of each pixel is a learnable network parameter, which is iteratively optimized based on error back-propagation algorithms, using standard deep learning tools implemented in a computer. After this training stage, the resulting transmissive layers are fabricated with 3D printing or lithography to

construct a task-specific physical network that computes based on the diffraction of the light passing through these trained diffractive layers.

Though most of the current diffractive neural networks are constructed based on linear optical materials, “deep” diffractive neural networks show evident “depth” advantages: an increase in the number of diffractive layers and neurons improves its statistical inference accuracy and information processing capability<sup>9,15</sup>. More specifically, adding more trainable diffractive layers into a given network increases the dimensionality of the solution space that can be all-optically processed by the network. It has been recently demonstrated that a diffractive neural network can be trained to perform an arbitrary complex-valued linear transformation between its input and output fields with negligible error, provided that the total number of engineered pixels in the network is sufficient<sup>16</sup>. In a more general sense, a diffractive neural network can be regarded as a special, task-specific optical system, which performs specific computational tasks with the use of light information carriers. The object field can be viewed as a source of information flow characterized by various fundamental properties, which can all be ingeniously manipulated to extend the information processing capacity of diffractive networks.

In a recent issue of *Light: Science & Applications*, the UCLA group introduced polarization division multiplexing (PDM), a long-established technique of enhancing the transmission capacity in telecommunications, to all-optically perform multiple, arbitrarily-selected linear



**Fig. 1 Information multiplexing in diffractive neural networks.** (a) Polarization-multiplexed diffractive neural networks utilizing a series of structured diffractive surfaces and a simple polarizer array. By enabling the trainable diffractive layers to communicate with the polarization elements embedded in the diffractive volume, a single network can create multiple computing channels that can be accessed using specific combinations of input and output polarization states. (b) Exploiting the internal degrees of freedom of light provide new possibilities for information multiplexing to enhance the performance and capacity of optical diffractive networks



transformations through a single diffractive network<sup>17</sup> (Fig. 1a). Instead of using birefringent, anisotropic, or polarization-sensitive materials for trainable diffractive layers, their polarization-multiplexed diffractive networks are still built based on standard (isotropic) diffractive surfaces where the trainable coefficients are independent of the polarization state of the input light. To gain additional sensitivity to different polarization states and polarization mode diversity, a non-trainable, pre-selected linear polarizer array (at 0°, 45°, 90°, and 135°) is inserted within the trainable diffractive surfaces, and different target linear transformations are uniquely assigned to different combinations of input and output polarization states. They demonstrated that a single well-trained polarization-multiplexed diffractive network could successfully perform multiple (2 or 4) arbitrarily-selected linear transformations, which had not yet been implemented by using metasurfaces or metamaterials-based designs<sup>11,17</sup>. Such a polarization-multiplexed diffractive computing framework is poised to be used to build all-optical, passive processors that can execute multiple inference and optical information processing tasks in parallel.

Harnessing the intrinsic high-dimensionality of light brings new insights into the diffractive neural network design by providing additional *degrees of freedom* to both optical signals and systems. The concept of degrees of freedom was first introduced in optics by Laue<sup>18</sup> in 1914 as the decisive property in determining the information capacity of optical signals and systems, even before Shannon's information theory for communication systems was established<sup>19</sup>. According to Laue<sup>18</sup>, and later Gabor<sup>20</sup>, Francia<sup>21</sup>, and Lukosz<sup>22,23</sup>, the number of degrees of freedom in optics is most often understood to be the number of independent parameters needed to represent an optical signal or system, which is closely related to the number of independent communication channels available for the information transfer in the field of electrical communication. However, unlike communication systems, an optical system transmits many kinds of information, which can be divided into two groups: (1) "*dimensional*" information which is related to spatial intervals by coordinates  $x$ ,  $y$ , and  $z$ , as well as to temporal intervals  $t$ ; (2) "*internal*" information which is related to physical properties of light, including amplitude, phase, wavelength, polarization, coherence, and angular momentum. The total number of degrees of freedom can be expressed through the product of freedom degree numbers related to all these different kinds of information. Though optical systems are often expected to transmit as much information as possible, the number of degrees of freedom is a fundamental invariant of an optical system, as noted by Gabor<sup>20</sup> and Lukosz<sup>22</sup>. Within this limit, it is possible to increase the degrees of freedom for one kind of information at the expense of that of another kind<sup>22,23</sup>.

The concept of degrees of freedom can be straightforwardly extended to the diffractive neural network, as a special kind of optical signal processing system. For example, the information content of the input or output signal, which is often an image formed through a pupil of finite size, can be quantified by the definite number of resolvable regions in which the signals can be independent (defined as  $N_i$  and  $N_j$  for the input or output signal in ref. <sup>15</sup>), taking both the diffraction limit and sampling theorem into account<sup>21,22</sup>. Diffractive neural networks manipulate light by reshaping the spatial profile of an input beam into a desired output beam. If the diffractive neural network is designed to perform arbitrary linear transformations from the input beam to the output beam, as demonstrated in refs. <sup>15,16</sup>, the entire optical system can be described by a single  $N_i \times N_j$  matrix, mapping  $N_i$  input degrees of freedom to  $N_j$  output degrees of freedom. In optical communications, the concept of "modes" or "eigenfunctions", is commonly used to provide an "economical" description of degrees of freedom of the optical signal, reducing complicated wave functions to a small number of mode amplitudes, as in propagating fiber modes and ideal laser beams<sup>24</sup>. In such a sense, the linear transformation function realized by the diffraction neural network is similar to that of an optical mode converter<sup>25</sup>. In contrast, diffractive neural networks can be built in a "deep" manner, consisting of several diffractive surfaces containing a large number of trainable neurons. Such a multi-layer design presents additional spatial optical degrees of freedom, significantly enhancing the information capacities and processing capability of the network compared with a single diffractive layer, as demonstrated by the UCLA group<sup>15-17</sup>. In particular, any linear transformation matrix from  $N_i$  to spatial degrees of freedom has  $N_i \times N_j$  free parameters. When the degrees of freedom of the diffractive neural network, i.e., the number of controllable parameters, is no less than  $N_i \times N_j$ , the network has in theory the capability to perform arbitrary linear transformations between the input and output signals perfectly. In their recent work<sup>17</sup>, two additional degrees of freedom of polarization are introduced to the input signal for simultaneously carrying the different information through the network. The two orthogonal polarization states carried by the beam present an attractive avenue to enhance the maximum information capacity of the diffractive neural network by a factor of  $N_p$  (from  $N_i \times N_j$  to  $N_p \times N_i \times N_j$ ), where  $N_p$  is the number of unique linear transformations assigned to different input/output states of polarization combinations. It should be mentioned that the use of polarization freedom as high-dimensional information carriers has been reported by Lohmann et al.<sup>26</sup> for optical super-resolution imaging and Chen et al.<sup>27</sup> in optical data communications.

The study of the UCLA group published in *Light Science & Application* is part of a larger movement to scale the capacity of optical diffractive computing by exploiting the *internal* degrees of freedom of light, such as polarization, spectrum, coherence, and orbital angular momentum, in addition to the *spatial* degrees of freedom (Fig. 1b). With such a multidimensional multi-link upgrade, diffractive neural networks can transmit optical signals over more independent channels, which could lead to all-optical multiplexed diffractive processors that can execute multiple tasks in parallel. Another benefit of polarization multiplexing is that the effective bandwidth can be reduced to the half of that of single-polarization transmission. That makes a high-dimensional diffractive neural network possible by using lower numerical-aperture optics, which has been proved to be extremely important for reducing the physical size of diffractive neural networks and relaxing the stringent requirements on the interlayer distances<sup>15,16</sup>. Finally, in most current diffractive network designs, the input field is assumed to be monochromatic, spatially coherent, and forward-propagating. A variety of computational imaging techniques that exploit partial coherence and evanescent waves for improving imaging performance (especially spatial resolution) prompted us to consider the possibility of their adaptation to diffraction neural networks. We believe that significant progress in developing high-performance optical diffractive computing schemes could be made if it became common practice to consider explicitly the internal degrees of freedom of light as the physical source of information gain.

#### Author details

<sup>1</sup>Smart Computational Imaging Laboratory (SCILab), School of Electronic and Optical Engineering, Nanjing University of Science and Technology, 210094 Nanjing, Jiangsu Province, China. <sup>2</sup>Smart Computational Imaging Research Institute (SCIRI) of Nanjing University of Science and Technology, 210019 Nanjing, Jiangsu Province, China. <sup>3</sup>Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, 210094 Nanjing, Jiangsu Province, China

#### Conflict of interest

The authors declare no competing interests.

Published online: 06 July 2022

#### References

- Sze, V. et al. Hardware for machine learning: challenges and opportunities. In: *2017 IEEE Custom Integrated Circuits Conference (CICC)* 1–8 (IEEE, 2017), <https://doi.org/10.1109/CICC.2017.7993626>.
- LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
- Wetzstein, G. et al. Inference in artificial intelligence with deep optics and photonics. *Nature* **588**, 39–47 (2020).
- Brunner, D. et al. Parallel photonic information processing at gigabyte per second data rates using transient states. *Nat. Commun.* **4**, 1364 (2013).
- Abu-Mostafa, Y. S. & Psaltis, D. Optical neural computers. *Sci. Am.* **256**, 88–95 (1987).
- Denz, C. *Optical Neural Networks*. (Springer Science & Business Media, Wiesbaden, 2013).
- Shen, Y. C. et al. Deep learning with coherent nanophotonic circuits. *Nat. Photonics* **11**, 441–446 (2017).
- Teğin, U. et al. Scalable optical learning operator. *Nat. Comput. Sci.* **1**, 542–549 (2021).
- Lin, X. et al. All-optical machine learning using diffractive deep neural networks. *Science* **361**, 1004–1008 (2018).
- Chang, J. L. et al. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Sci. Rep.* **8**, 12324 (2018).
- Liu, C. et al. A programmable diffractive deep neural network based on a digital-coding metasurface array. *Nat. Electron.* **5**, 113–122 (2022).
- Li, J. X. et al. Spectrally encoded single-pixel machine vision using diffractive networks. *Sci. Adv.* **7**, eabd7690 (2021).
- Mengu, D. & Ozcan, A. All-optical phase recovery: diffractive computing for quantitative phase imaging. *Adv. Opt. Mater.* <https://doi.org/10.1002/adom.202200281> (2022).
- Luo, Y. et al. Computational imaging without a computer: seeing through random diffusers at the speed of light. *eLight* **2**, 4 (2022).
- Kulce, O. et al. All-optical information-processing capacity of diffractive surfaces. *Light: Sci. Appl.* **10**, 25 (2021).
- Kulce, O. et al. All-optical synthesis of an arbitrary linear transformation using diffractive surfaces. *Light: Sci. Appl.* **10**, 196 (2021).
- Li, J. X. et al. Polarization multiplexed diffractive computing: all-optical implementation of a group of linear transformations through a polarization-encoded diffractive network. *Light: Sci. Appl.* **11**, 153 (2022).
- Laue, M. V. Die Freiheitsgrade von strahlenbündeln. *Ann. Phys.* **349**, 1197–1212 (1914).
- Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
- Gabor, D. in *Progress in Optics* (ed Wolf, E.) 109–153 (Elsevier, 1961).
- Di Francia, G. T. Degrees of freedom of an image. *J. Opt. Soc. Am.* **59**, 799–804 (1969).
- Lukosz, W. Optical systems with resolving powers exceeding the classical limit. *J. Opt. Soc. Am.* **56**, 1463–1471 (1966).
- Lukosz, W. Optical systems with resolving powers exceeding the classical limit. II. *J. Opt. Soc. Am.* **57**, 932–941 (1967).
- Miller, D. A. B. Waves, modes, communications, and optics: a tutorial. *Adv. Opt. Photonics* **11**, 679–825 (2019).
- Miller, D. A. B. All linear optical devices are mode converters. *Opt. Express* **20**, 23985–23993 (2012).
- Lohmann, A. W. & Paris, D. P. Superresolution for nonbirefringent objects. *Appl. Opt.* **3**, 1037–1043 (1964).
- Chen, Z. Y. et al. Use of polarization freedom beyond polarization-division multiplexing to support high-speed and spectral-efficient data transmission. *Light: Sci. Appl.* **6**, e16207 (2017).

# 2024

# THE NOBEL PRIZE



## THE NOBEL PRIZE IN PHYSICS 2024

Illustrations: Niklas Elmehed



John J. Hopfield

Geoffrey E. Hinton

"for foundational discoveries and inventions  
that enable machine learning  
with artificial neural networks"

THE ROYAL SWEDISH ACADEMY OF SCIENCES

2024年诺贝尔物理学奖授予了两位人工智能（AI）科学家  
Geoffrey Hinton 和 John Hopfield，  
以表彰他们在人工神经网络方面的开创性工作。



# 2024

# THE NOBEL PRIZE



Illustrations: Niklas Elmehed

## THE NOBEL PRIZE IN CHEMISTRY 2024



David  
Baker

“for computational  
protein design”

Demis  
Hassabis

“for protein structure prediction”

John M.  
Jumper

THE ROYAL SWEDISH ACADEMY OF SCIENCES

2024年诺贝尔化学奖授予了David Baker、Demis Hassabis和John M. Jumper三位科学家。David Baker因其在计算蛋白质设计方面的卓越贡献而获奖，而Demis Hassabis和John M. Jumper则因他们结合人工智能（AI）模型在蛋白质结构预测领域的开创性工作而共同获奖。





# 南京理工大学智能计算成像实验室

## 实验室简介

南京理工大学智能计算成像实验室 (SCILab: www.scilaboratory.com) 隶属于南理工国家一级重点学科“光学工程”学科带头人陈钱教授领衔的“光谱成像与信息处理”教育部长江学者创新团队、首批“全国高校黄大年式教师团队”。实验室学术带头人左超教授为国际光学工程学会/美国光学学会/英国物理学会会士 (SPIE/Optica/IOP Fellow)、教育部长江学者特聘教授、连续入选科睿唯安全球高被引科学家。依托于教育部先进光电成像技术与仪器工程研究中心、江苏省光谱成像与智能感知重点实验室、教育部高维信息智能感知与系统重点实验室等7个高层次研究平台，以及科技部图像测量技术研究示范型国家国际科技合作基地，教育部先进光电成像理论与技术学科创新引智基地2个高层次国合平台。实验室致力于研发新一代计算成像与传感技术，在国家重大需求牵引及重点项目支持下开展新型光学成像的机理探索、工程实践以及先进仪器的研制工作，并开拓其在生物医药、智能制造、国防安全等领域的前沿应用。



## 科学研究

实验室在科技部重点研发计划、基金委重大仪器与重点项目、江苏省基础前沿引领专项等重大项目支持下，在非干涉定量相位显微成像、高速结构光投影三维成像、远场被动红外超分辨成像、先进生物医学光学成像等方面取得了系列重要研究成果。研究成果已在SCI源刊上发表论文 250 余篇，其中 40 余篇论文被选作Light、Optica 等期刊封面论，20余篇论文入选ESI高被引/热点论文，论文被引超过17000次。获中国光学工程学会技术发明奖一等奖、江苏省科学技术奖一等奖、日内瓦国际发明展“特别嘉许金奖”等。科研成果得到了 30 余名院士、上百余名国际学会 Fellow 的正面评价，被 Nature、MIT Technology Review、Phys.org、《人民日报》等报道百余次，引起了国内外同行的极大关注。



## 人才培养

实验室指导博士生6人获全国光学工程优秀博士论文/提名奖、2人获国际光学工程学会(SPIE)光学与光子学教育奖学金、5人获中国光学学会王大珩光学奖、10人获Light全国光学博士生学术联赛全国百强。研究生获国家奖学金30余人次，国际会议最佳报告/海报奖 40 余人次。学生团队获全国“挑战杯”、“互联网+”、“创青春”、“研电赛”金奖/特等奖十余次，2023年，获“互联网+”全球总冠军。师生双创事迹得到央视《焦点访谈》、人民网、新华网、光明日报、中国教育电视台等百余家媒体报道，社会辐射影响广泛。

招聘老师和博士后、欢迎报考研究生  
脚踏实地,仰望星空; 划过历史星河,镌刻人生印记!



# AI for Imaging & Metrology



Contact us today.

zuochao@njust.edu.cn +86-25-84315587

[http:// www.scilaboratory.com](http://www.scilaboratory.com)

©2024 Chao Zuo @ SCILab and its logo, as well as all other trademarks used herein are trademarks of their respective owners and used under license.