Research Article Vol. 14, No. 5 / May 2026 / Photonics Research 1893

PHOTONICS Research

Spatiotemporal fringe pattern analysis using physics-informed deep learning

Xinran Zhou,[1,2,3,†] Yiheng Liu,[1,2,3,†] Jie Tang,[1,2,3] Yutong Xiao,[1,2,3] Yifan Liu,[1,2,3] Shijie Feng,[1,2,3,4,5] Qian Chen,[3,4,6] and Chao Zuo[1,2,3,4,7]

[1]Smart Computational Imaging Laboratory (SCILab), School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[2]Smart Computational Imaging Research Institute (SCIRI) of Nanjing University of Science and Technology, Nanjing 210019, China
[3]Jiangsu Key Laboratory of Visual Sensing & Intelligent Perception, Nanjing 210094, China
[4]State Key Laboratory of Extreme Environment Optoelectronic Dynamic Measurement Technology and Instrument, Taiyuan 030051, China
[5]e-mail: shijiefeng@njust.edu.cn
[6]e-mail: chenqian@njust.edu.cn
[7]e-mail: zuochao@njust.edu.cn
[†]These authors contributed equally to this work.

Received 17 November 2025; revised 20 January 2026; accepted 3 February 2026; posted 6 February 2026 (Doc. ID 584978); published 24 April 2026

In optical metrology, particularly in phase-measuring profilometry techniques, achieving both high measurement accuracy and high temporal resolution remains a long-standing challenge. For high-speed dynamic measurements, many existing methods still adopt a frame-by-frame processing strategy, tending to overlook the rich temporal information embedded across adjacent frames. This constraint makes it difficult to maintain robust performance under ultra-short exposure conditions. Taking fringe projection profilometry (FPP) as a representative example, we propose a physics-informed spatiotemporal fringe pattern analysis (PIST-FPA) framework that unifies physical modeling with deep learning for high-speed, high-fidelity three-dimensional (3D) reconstruction. Rather than processing each frame independently, PIST-FPA jointly analyzes a short sequence of fringe images within a temporal window, exploiting both spatiotemporal correlations and physics-based priors. PIST-FPA incorporates a background-estimation network for illumination correction, a physics-guided optical-flow alignment for motion compensation, and an attention-enhanced U-Net for spatiotemporal phase demodulation, forming a unified and interpretable learning-based processing pipeline. This hybrid paradigm effectively decouples object motion from surface morphology, enabling accurate and temporally consistent phase retrieval even under ultra-short exposure and low-signal-to-noise conditions. Both synthetic and real experiments at 10,000 frames per second demonstrate that PIST-FPA substantially outperforms both the conventional Fourier transform method and single-frame deep-learning method in terms of accuracy, robustness, and spatiotemporal consistency. By bridging physical modeling and data-driven learning in a unified spatiotemporal framework, PIST-FPA establishes a generalizable paradigm for physics-informed dynamic optical metrology, paving the way for next-generation high-speed phase-measuring profilometry and intelligent optical instrumentation.

https://doi.org/10.1364/PRJ.584978

1. INTRODUCTION

Optical metrology, which uses light as an information carrier to enable non-contact measurement [1], is essential in fields such as manufacturing, scientific research, and engineering. Non-contact optical techniques like holographic interferometry [2], electronic speckle pattern interferometry [3], and fringe projection profilometry [4] are widely applied in the fields of industrial inspection [5], cultural heritage digitization [6], and biomedical imaging [7]. Among those methods, the 3D information of the object is encoded in the phase of a two-dimensional fringe pattern, so the fidelity of measurement is fundamentally constrained by the precision of phase recovery from the recorded fringes.

Conventionally, fringe analysis strategies can be broadly categorized into two main approaches: the Fourier transform (FT) method based on spatial transformations [8] and the phase-shifting (PS) method based on temporal variations [9]. The FT approach stands as a representative single-frame analysis method that involves applying a Fourier transform to

a single distorted fringe pattern, subsequently filtering out the first-order harmonic component, which carries the essential phase information within the frequency domain, and ultimately recovering the phase through an inverse transform operation. Its core advantage lies in requiring only one image for measurement, which enables extremely high acquisition speed and makes it inherently suitable for capturing dynamic scenes and transient events. Nevertheless, the effectiveness of frequency-domain filtering plays a crucial role in determining the accuracy of this method. When the measured object exhibits complex surface geometry with sharp edges or steep slopes, spectral aliasing is likely to occur, preventing the fundamental frequency component from being perfectly separated. As a result, the reconstruction accuracy is significantly degraded, making it difficult to obtain highly precise measurements in complex regions [10–13].

In contrast, the PS method performs phase demodulation in the temporal domain [14]. By capturing multiple (typically three or more) fringe patterns with precise phase-shifting increments across a time series, it enables point-by-point, high-precision phase value calculation, significantly enhancing the measurement's signal-to-noise ratio (SNR). In static scenes, the PS method effectively suppresses interference from ambient light and uneven surface reflections, achieving sub-millimeter scale phase measurement accuracy to reconstruct detailed 3D shape. However, this high precision comes at the cost of reduced temporal resolution. During dynamic measurements, object motion between frames violates the static-scene assumption underlying phase-shifting algorithms, leading to motion blur and phase errors [15].

Recent advances in artificial intelligence (AI) have led to the widespread application of deep learning across various fields, thanks to its powerful ability to extract features and recognize complex patterns [16]. It has achieved notable progress in medical imaging [17,18], natural language processing [19,20], computer vision [21,22], and other domains, while also bringing new opportunities for the development of single-frame structured light 3D measurement technology [23]. In fringe analysis applications, Feng *et al.* have successfully utilized deep learning techniques to achieve high-precision phase prediction from a single frame image [24], even handling phase extraction under non-sinusoidal fringe conditions [25]. Building upon this foundation, Yin *et al.* proposed integrating frequency-domain filtering as an embeddable physical module to collaborate with the network for phase learning, enhancing both accuracy and model universality [26]. To further improve neural network generalization, Li *et al.* adopted a mixture-of-experts (MoE) architecture [27]. By integrating multiple expert models specializing in different fringe features and dynamically weighting their outputs, the model achieves adaptive phase recovery across varying system parameters and environments, significantly enhancing its generalization capability.

Furthermore, recent research indicates that deep learning technologies can enable 3D imaging beyond the inherent limitations of imaging systems. Chen *et al.* proposed a deep-learning-based time-super-resolved high-speed 3D imaging technique [28,29]. Leveraging frequency-domain multiplexing, their method records 3D information of objects at multiple time points within a single camera exposure via temporal integration. Using conventional low-speed cameras, this allows 3D imaging at speeds nearly an order of magnitude higher than the camera's frame rate. Wang *et al.* developed an ultrafast 3D imaging technique based on single-shot super-resolved fringe projection profilometry (SSSR-FPP), achieving dynamic 3D reconstruction at a frame rate of up to 100,000 Hz. This technique requires only a couple of fringe images as input, which are characterized by low resolution, low SNR, and pixelation. A specially trained deep neural network then directly decodes high-resolution unwrapped phase and fringe order information [30].

Although intelligent algorithms represented by deep learning have successfully enabled high-precision phase demodulation of single-frame fringe images [31–33], does this imply that simplifying a continuous image sequence into a series of independent single-frame measurements is the optimal strategy when processing dynamic scenes? We contend that this frame-by-frame processing paradigm may fail to fully leverage the valuable temporal information inherent in dynamic sequences. Due to the camera's high-speed continuous acquisition, adjacent frames contain rich redundant information about object morphology and motion. Effectively utilizing this information would undoubtedly enhance the quality and robustness of the final 3D reconstruction. Particularly under extreme measurement conditions, such as high-speed dynamics, exposure times are often reduced to the microsecond range, which leads to a sharp deterioration in the SNR of the acquired fringe patterns [34]. Specifically, ultra-short exposure cuts down the photon budget and weakens fringe modulation contrast, making the relative contributions of read noise and shot noise far more pronounced, thus turning the input fringe data into a noise-dominated signal [35]. Since deep neural networks rely heavily on distinguishable structural information to learn effective representations for phase demodulation, the obscuration of fringe structures by random noise makes it difficult for the network to distinguish valid features from noise. Such noise interference impairs the reliability of feature extraction, ultimately degrading the accuracy of phase reconstruction and subsequent 3D reconstruction. Under such strong noise interference, the effective spatial information carried by a single frame becomes extremely limited, making it difficult to support reliable feature learning and phase prediction by neural networks. This severely limits the reliability and accuracy of the reconstruction results.

Current deep-learning-based fringe analysis methods largely rely on single-frame processing, leaving temporal redundancy unexploited. While feasible for dynamic measurements, this paradigm is inherently prone to noise interference, which can compromise accuracy in low-SNR conditions. To overcome the information bottleneck of single-frame measurements, we introduce a physics-informed spatiotemporal fringe pattern analysis (PIST-FPA) framework, which, to the best of our knowledge, is the first to integrate spatiotemporal feature decoupling with physics-guided modeling for phase-measuring profilometry. The workflow of PIST-FPA is shown in Fig. 1. This approach aims to overcome the information bottleneck of

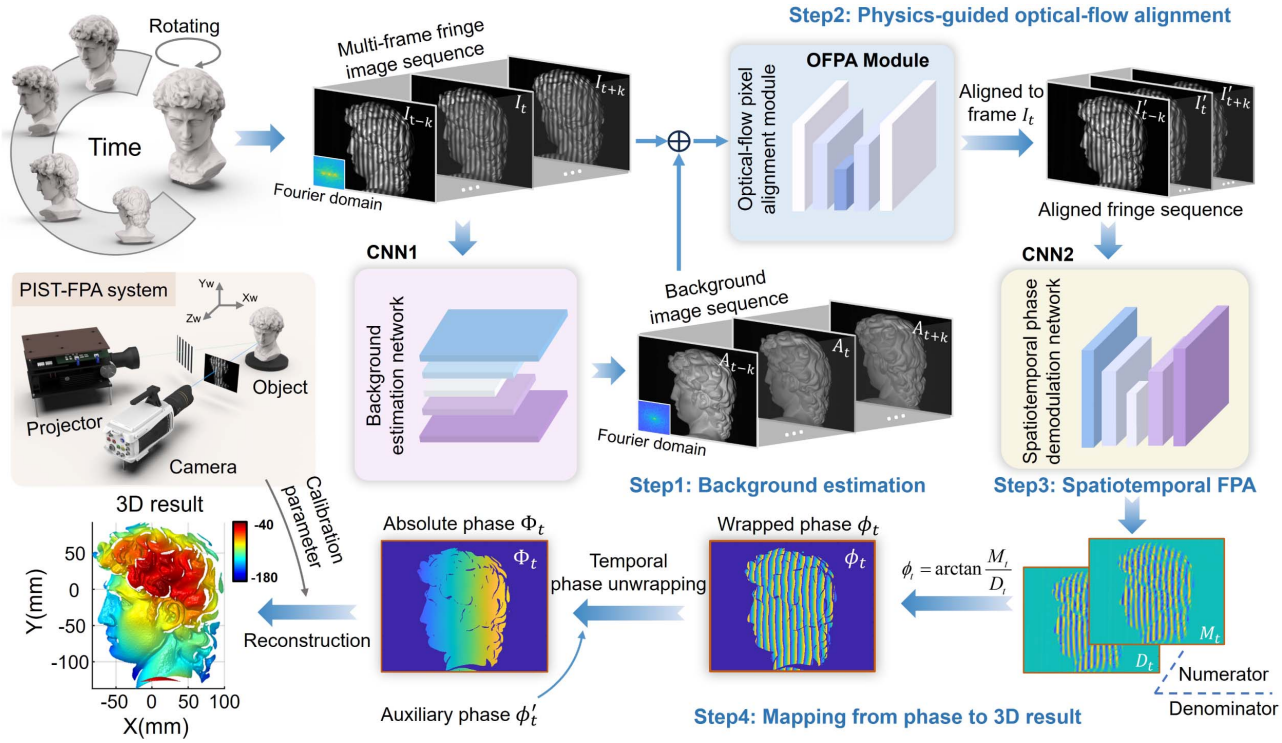

Fig. 1. Workflow of the proposed PIST-FPA method. This method integrates optical-flow physical priors with spatiotemporal information from adjacent frames. Step 1: a sequence of fringe images is processed by CNN1 to separate the background intensity. Step 2: the background intensity sequence from CNN1, together with the fringe image sequence, is aligned to the central frame by using the OFPA module. Step 3: CNN2 predicts the phase of the central frame from the aligned fringe sequence. Step 4: mapping the phase to 3D coordinates to obtain the final 3D reconstruction.

single-frame images by synergistically leveraging temporal redundancy and physical prior knowledge. First, a background-estimation network (CNN1) predicts and separates background intensity from the input raw multi-frame sequence to obtain pure object intensity information. Subsequently, an optical-flow pixel alignment (OFPA) module guided by a physical prior processes this intensity sequence. It computes the optical-flow field between adjacent frames to compensate for object motion, precisely mapping all frames to a unified reference frame and achieving decoupling of motion and shape features. Finally, the aligned temporal fringe sequence is passed into a spatiotemporal phase demodulation network (CNN2), which adaptively integrates multi-scale spatiotemporal features using attention mechanisms to precisely recover the phase of the central frame. By effectively integrating spatiotemporal information with physical prior knowledge, this method establishes a novel paradigm for achieving high-precision measurements in the domain of high-speed dynamic structured light 3D imaging.

2. PRINCIPLE

A. Single-Frame Fringe Analysis

Phase retrieval from fringe images is a foundational task and a canonical example of deep-learning applications in optical metrology. The acquired intensity distribution of fringe image $I(x,y)$ can be represented as

$$I(x,y) = A(x,y) + B(x,y)\cos[\phi(x,y) + 2\pi f x] + n(x,y), \tag{1}$$

where $A(x,y)$ is the background intensity, $B(x,y)$ is the amplitude, f is the frequency of the illumination grating, $\phi(x,y)$ is the object's phase, and $n(x,y) \sim \mathcal{N}(0,\sigma_n^2)$ is the noise. According to the phase retrieval algorithm, the phase retrieval is as follows:

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

where c represents constants related to phase demodulation algorithms. In single-frame fringe analysis algorithms based on deep learning, the neural network is trained to output $M(x,y)$ and $D(x,y)$ as shown in Eq. (2) [24]:

$$[M(x,y), D(x,y)] = f_{\mathrm{DNN}}[I(x,y)], \tag{3}$$

where f_{DNN} represents the constructed single-frame fringe deep neural network, which takes $I(x,y)$ as input and is trained to output $M(x,y)$ and $D(x,y)$. Then, the final phase distribution is calculated by applying the inverse tangent. For static fringe patterns with sufficient SNR, single-frame demodulation algorithms combined with deep learning can already achieve high accuracy. However, in high-speed measurement tasks, the microsecond-level short exposure strategy used to minimize motion blur compromises the image's SNR, presenting

a significant challenge to the robustness and accuracy of single-frame phase demodulation.

From the perspective of information theory and inverse problem solving, single-frame fringe analysis faces an inherent information bottleneck. For any pixel captured by the camera, its fringe intensity I is determined by the background light intensity A, modulation amplitude B, and the phase ϕ to be retrieved. Since we only have one observation (I) but three unknowns (A, B, ϕ), this constitutes a typical mathematically ill-posed underdetermined problem. To solve this equation, existing single-frame methods should introduce spatial priors, assuming that pixels within a local window or the receptive field of a neural network have the same parameters or vary smoothly. This reliance on local spatial neighborhood information is the core of the "information bottleneck" in single-frame measurements.

To evaluate this limitation, we analyze the theoretical accuracy limit of phase estimation based on the Cramér–Rao lower bound (CRLB). The variance lower bound of single-frame phase estimation is approximately

$$\mathrm{Var}(\hat{\phi}_{\mathrm{single}}) \propto \frac{\sigma_n^2}{B^2}. \tag{4}$$

This implies that in high-speed scenarios with ultra-short exposure, the noise variance σ_n^2 increases sharply due to photon shot noise, while the modulation amplitude B is constrained by limited light intensity. Consequently, this results in a dramatic rise in the CRLB, imposing a fundamental limit on the achievable measurement accuracy. The detailed derivation is provided in Appendix A.

In contrast, the proposed PIST-FPA framework breaks this bottleneck by exploiting redundant information within a time window of length T. After aligning the sequence via the OFPA module, the Fisher information is effectively accumulated in the temporal dimension. Our derivation (see Appendix B) demonstrates that the theoretical variance lower bound of PIST-FPA is reduced to

$$\mathrm{Var}(\hat{\phi}_{\mathrm{PIST}}) = \frac{1}{T}\mathrm{Var}(\hat{\phi}_{\mathrm{single}}). \tag{5}$$

This theoretical result indicates that PIST-FPA reduces uncertainty by a factor of T through spatiotemporal joint analysis, thereby breaking the inherent information bottleneck of single-frame measurements from a physical perspective.

B. Physics-Informed Spatiotemporal Fringe Pattern Analysis

Unlike traditional single-frame fringe analysis methods, this paper focuses on extracting spatiotemporal information from the entire grating sequence. For this purpose, we design and propose a novel framework—PIST-FPA. This method aims to accurately reconstruct an object's 3D shape using a sequence of fringe images. The entire workflow comprises three core components: background intensity separation, optical-flow-based pixel alignment, and phase information extraction. This framework is implemented through two convolutional neural networks (CNN1 and CNN2), with an OFPA module based on physical priors introduced between them to address pixel misalignment caused by object motion. The detailed architectural design of the PIST-FPA module is illustrated in Fig. 2.

Specifically, the first stage of our proposed method is background intensity separation. Assuming continuous illumination from a single fixed grating, the input raw multi-frame fringe image sequence is initially processed by a specially designed convolutional neural network module, CNN1. We utilize the classic U-Net as the core architecture of CNN1 [36]. In our implementation, CNN1 is a four-level U-Net for frame-wise regression from a single normalized fringe frame $I(x, y, t)$ to the background component $A(x, y, t)$. The encoder uses 2×2 max-pooling with channel numbers $\{C, 2C, 4C\ 8C\}$ and a $16C$-channel bottleneck; the decoder mirrors the encoder using $\times 2$ upsampling followed by a 2×2 convolution and a skip concatenation. The design objective of CNN1 is to precisely regress the corresponding background intensity component from the image containing complex fringe modulation on a frame-by-frame basis:

$$[A(x, y, t)] = f_{\mathrm{CNN1}}[I(x, y, t)]. \tag{6}$$

After processing with CNN1, the background intensity information is extracted from the fringe images, resulting in a sequence of background intensity images. This preprocessing step effectively addresses complex and uneven illumination and environmental light interference encountered in actual measurements.

Next, as illustrated in Fig. 2(b), when an object moves, its motion causes the pixel coordinates to shift. To address this, we employ optical-flow methods to calculate pixel displacements across different time points. Assuming the background intensity and surface reflectance of scene pixels remain constant between consecutive frames, the relationship between image intensities at different time points for the frame can be expressed as

$$A(x, y, t) = A(x + \mathrm{d}x, y + \mathrm{d}y, t + \mathrm{d}t), \tag{7}$$

where $d = (\mathrm{d}x, \mathrm{d}y)$ represents the offset of the pixels at different times. To align the image, Eq. (7) is expanded using a Taylor series, and after organization, it can be expressed as

$$\frac{\partial A}{\partial x}u + \frac{\partial A}{\partial y}v + \frac{\partial A}{\partial t} = 0, \tag{8}$$

where u and v denote the object's velocities along the x and y directions, respectively, and can be represented by the vector $\mathbf{v} = (u, v)$. We introduce the Lucas–Kanade (LK) optical-flow algorithm to calculate the object's motion velocity [37]. Assuming the object's motion within the small window $m \times m$ ($m > 2$) exhibits similarity, according to Eq. (8), we have

$$\mathbf{v} = (C^T C)^{-1} C^T b, \tag{9}$$

where $C = [\nabla A(x_1, y_1), \ldots, \nabla A(x_m, y_m)]^T$ and $b = -[\frac{\partial A(x_1, y_1)}{\partial t}, \ldots, \frac{\partial A(x_m, y_m)}{\partial t}]^T$. Based on the estimated motion parameters, pixel-to-pixel alignment of the grating sequence images can be performed to obtain $I'(x, y, t)$.

Then, to analyze the aligned fringe sequences, we constructed CNN2, which is responsible for regressing the sine and cosine terms required to compute the phase at the calculation time from the aligned fringe pattern sequence:

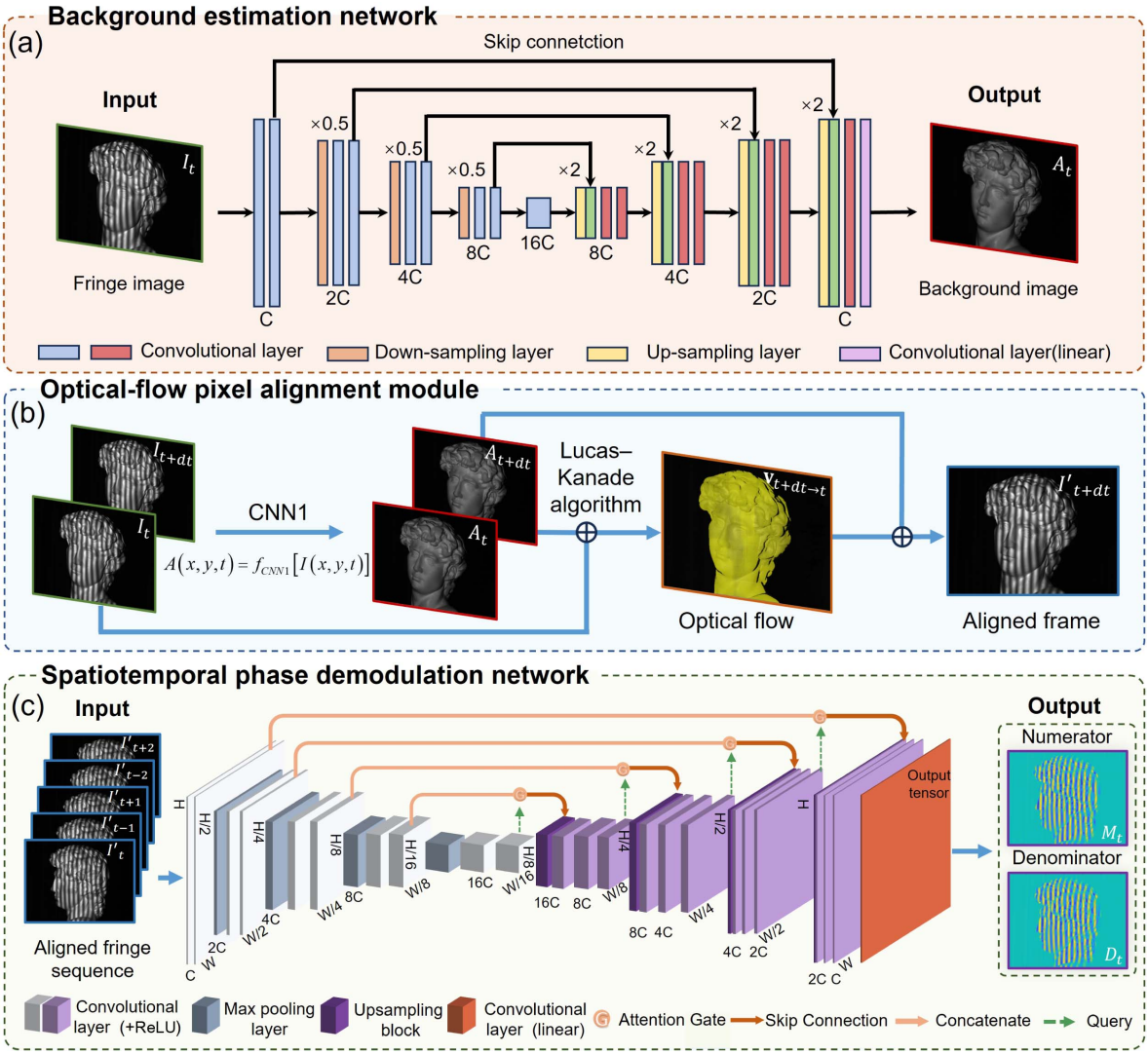

Fig. 2. Detailed structure of PIST-FPA modules. (a) Detailed network architecture of the background estimation network (CNN1). The network extracts the background intensity from the input fringe image to assist the subsequent optical-flow calculation. (b) Workflow of the OFPA module, which aligns fringe images in dynamic scenes using optical-flow information. The module first processes the fringe image sequence with CNN1 to separate background intensity, and then applies the optical-flow algorithm to calculate the pixel-wise motion between adjacent frames. (c) Detailed network architecture of CNN2. The network takes the aligned fringe sequence as input and predicts the numerator and denominator of the phase using multi-scale spatiotemporal features, enhanced by attention mechanisms and skip connections.

$$[M(x,y,t), D(x,y,t)] = f_{\mathrm{CNN2}}(S), \tag{10}$$

where $S = \{I'(x,y,t-\Delta t), \ldots, I'(x,y,t), \ldots, I'(x,y,t+\Delta t)\}$. Finally, using Eq. (2), the wrapped phase with high precision is calculated, which then allows for the reconstruction of the object's 3D surface profile through phase unwrapping and system calibration parameters [38]. Figure 2(c) illustrates that the CNN2 used in this method follows the Attention U-Net architecture [39]. This network employs a classic encoder-decoder structure. In the encoder (downsampling path), deep features are progressively extracted from the image using multiple convolutional and max pooling layers, which simultaneously reduce the spatial resolution of the feature maps. In the decoder (upsampling path), low-resolution deep feature maps are progressively restored to the original image resolution using multiple upsampling blocks and convolutional layers. To fuse features from different levels and prevent gradient vanishing, the network employs skip connections and attention gates. Feature maps from corresponding encoder layers are concatenated with those from the decoder. Specifically, an attention gate is introduced within the skip connection. This gate automatically learns to focus on important regional features within the image while suppressing interference from irrelevant background areas, enabling the network to reconstruct detailed information with greater precision.

To further enhance generalization capabilities, we designed a loss function that maintains consistency across temporal, spatial, and frequency domains:

$$L_{\mathrm{TSF}} = \lambda_1 L_{\mathrm{TS}} + \lambda_2 L_{\mathrm{TF}}, \tag{11}$$

where L_{TS} denotes the spatiotemporal consistency loss function, L_{TF} denotes the time-frequency consistency loss function, and λ_1 and λ_2 represent their respective weights. We used the mean squared error (MSE) to define the spatiotemporal consistency loss function:

$$L_{TS} = \frac{1}{TN_xN_y}\sum_{t=1}^{T}\sum_{x=1}^{N_x}\sum_{y=1}^{N_y}[\mathbf{y}_t(x,y) - \mathbf{y}'_t(x,y)]^2, \tag{12}$$

where $\mathbf{y}_t$ denotes the model's prediction of M and D at the (x, y) pixel position in frame t, while $\mathbf{y}'_t$ represents their ground truth. Next, we establish the time-frequency consistency loss function L_{TF}:

$$L_{TF} = \frac{1}{TF_xF_y}\sum_{t=1}^{T}\sum_{f_x=1}^{F_x}\sum_{f_y=1}^{F_y}||\mathcal{F}(\mathbf{y}_t)(f_x,f_y)| - |\mathcal{F}(\mathbf{y}'_t)(f_x,f_y)||, \tag{13}$$

where $\mathcal{F}(\cdot)(f_x, f_y)$ denotes the complex value in the frequency domain coordinates (f_x, f_y) for the frame image t, and $|\cdot|$ represents the modulus of the complex number. The spatiotemporal-frequency-domain consistency loss function L_{TSF} established in this study imposes multiple complementary constraints on phase prediction. On one hand, spatiotemporal loss ensures the physical accuracy of phase demodulation and temporal continuity. On the other hand, the frequency-domain loss serves as an effective regularization method, optimizing phase smoothness and SNR by suppressing spectral artifacts. These multiple constraints work synergistically to achieve high-precision dynamic phase demodulation.

3. EXPERIMENT

To validate the effectiveness of the method presented under the scenario of fringe projection profilometry, we assembled a high-speed fringe projection platform using a Vision Research Phantom V611 camera in combination with a Texas Instruments DLP 4100 projector. To obtain a dataset for training neural networks while reducing the cost of acquiring real-world data, we developed a digital-twin-based fringe projection system. This virtual simulation system was developed using the 3D rendering software Blender [40], with a primary focus on accurately replicating the real physical system based on its calibration parameters. This encompasses the intrinsic and extrinsic parameters of both the camera and projector, which can be obtained through the calibration of the real system, including resolution, focal length, principal point location, distortion coefficient, and the relative pose between the two. Through rigorous virtual calibration, we ensured high consistency between the geometric projection models of both the real system and the digital-twin system [41].

During the dataset generation stage, we utilized a large number of open-source 3D models with varying levels of complexity and surface characteristics (e.g., models from Thingi10K) [42] as measurement subjects. We synthesized diverse dynamic scenarios to evaluate these models, incorporating translation, rotation, and complex trajectories. To establish a reliable ground truth, we employed a step-and-hold acquisition strategy. At each stationary pose, both a single-shot fringe image and a reference 12-step phase-shifting sequence were captured. Since the object remained static during this process, the resulting phase-shifting reconstruction is free of motion artifacts, serving as a high-fidelity ground truth. Correspondingly, by sequentially aggregating the single-shot fringe images from each pose, we obtained a fringe sequence representing the equivalent continuous motion of the object. We generated a dataset comprising 1000 dynamic sequences, where 70% was used for training, 20% for validation, and 10% for testing. To compute training labels, we employed a 12-step phase-shifting method. The input fringe sequence intensities are normalized to [0, 1] during training. Built in TensorFlow, the network is trained for 200 epochs on an NVIDIA GTX 4090 GPU using the Adam optimizer with an initial learning rate of 0.0001 and a batch size of two.

Finally, to address the gap between simulated and real-world data (sim-to-real gap), we collected 240 sets of real-world dynamic scene data using a physical system, with the camera capturing at a frame rate of 10,000 frames per second (FPS). To obtain reliable ground truth, we also used the step-and-hold acquisition strategy. Subsequently, we fine-tuned the pre-trained model using this real-world data by employing a transfer learning strategy. To ultimately evaluate the model's performance, we further collected more diverse real-world dynamic scenes not encountered during the fine-tuning phase. This independent real-world test data serves as the final performance baseline, helping to assess the effectiveness and robustness of our proposed method.

A. Impact of Time Window Length on Neural Network Performance

The selection of the temporal window length T involves a fundamental trade-off between utilizing temporal redundancy and avoiding motion-induced misalignment. A larger T introduces richer temporal contextual information, facilitating noise suppression and feature learning. However, it also increases the cumulative displacement of the object at the window edges relative to the central reference frame. To ensure reliable motion compensation, this cumulative displacement should remain within the linear valid range of the optical-flow alignment algorithm (typically denoted as $\delta_{max} \approx 1$ to 2 pixels). Given the camera frame rate F, the object depth Z, the camera focal lengths (f_x, f_y), and the object's tangential velocity components (v_x, v_y), the maximum allowable window size is given by

$$T \le \frac{2\cdot\gamma\cdot\delta_{max}\cdot F\cdot Z}{\sqrt{(f_xv_x)^2+(f_yv_y)^2}} + 1. \tag{14}$$

Equation (14) reveals that the temporal window scales linearly with the sampling density (frame rate) but is inversely proportional to the projection of object velocity on the sensor plane. Further, given the subsequent processing of temporal fringe patterns via a convolutional neural network, the alignment constraints inherent to the original LK algorithm are relaxed. This implies that a wider range of alignment deviation, specifically $\gamma\cdot\delta_{max}$, is permissible within our framework. The detailed derivation is provided in Appendix C.

According to Eq. (C13), the maximum allowable time window size is physically constrained by the object motion speed to ensure the cumulative displacement remains within the linear range (extended by the neural network correction factor, taken as $\gamma \approx 4$ in our experiments) of our proposed method. To validate this theoretical model and determine the optimal T for our experimental setup, we conducted a quantitative analysis using the digital-twin simulation dataset. We set the camera frame rate at $F = 10{,}000$ Hz and the object depth at $Z = 1$ m, the camera focal length at $f_x = f_y = 1814$ pixels, and introduced Gaussian noise with a variance of $\sigma^2 = 2.5$ to the input fringe patterns. We then evaluated the phase reconstruction accuracy under three distinct object translational speeds of 10.0 m/s, 20.0 m/s, and 40.0 m/s, while varying window length across the set $T \in \{1, 3, 5, 7, 9\}$. For comparison, all models adopt the same Attention U-Net backbone architecture, loss function, and training hyperparameters, with only minor variations in the input layer due to differing channel counts. All the experiments were implemented on the Blender digital-twin simulation dataset, and Gaussian noise was added to the input fringe patterns to verify the noise robustness of the model under different dynamic conditions. The experimental results are reported in Table 1.

In the low-speed scenario (10.0 m/s), the theoretical constraint on the window size is relaxed. Consequently, expanding the time window from $T = 1$ to $T = 9$ yields a monotonic reduction in phase mean absolute error (MAE), decreasing from 0.0535 rad to 0.0363 rad with an improvement of 32.15%. This confirms that when the alignment condition is satisfied, a larger temporal window provides richer redundancy, facilitating more effective noise suppression and feature learning.

As the motion speed increases to 20.0 m/s, the physically allowable window size decreases. The experimental results show that the minimum MAE reaches 0.0424 rad at $T = 5$. Notably, when T further increases to 7, the error rebounds significantly to 0.0512 rad. This error curve indicates that for $T > 5$, the cumulative displacement at the window edges exceeds the valid range of our proposed method, leading to alignment artifacts that outweigh the benefits of temporal averaging.

Under the high-speed condition (40.0 m/s), the theoretical constraint becomes even more stringent. The optimal window shrinks to $T = 3$, achieving the lowest MAE of 0.0445 rad. A window of $T = 5$ leads to accuracy degradation, with the MAE increasing to 0.0501 rad, validating our theoretical prediction that excessive temporal integration in fast-motion scenes introduces severe misalignment errors.

B. Validation of Spatiotemporal Joint Mechanisms with Virtual Data

First, we tested this method using virtual data with an object motion speed of 18 m/s and the optimal temporal window length $T = 5$. We incorporated Gaussian noise with a mean of zero and a variance of σ^2 (where $\{\sigma^2 | 0 \le \sigma^2 \le 2.5\}$) to the synthetic grating. For comparison, we also introduced Fourier transform profilometry (FTP) [43,44] and single-frame fringe pattern analysis (SF-FPA) [24]. In our implementation, SF-FPA is a four-level U-Net for frame-wise regression from a single normalized fringe frame $I(x, y, t)$ to the two intermediate maps (M, D) for phase recovery. For fair comparison, SF-FPA is trained and evaluated on the same dataset split with the same training protocol as our method, but without multi-frame input, optical-flow alignment, or spatiotemporal fusion. We evaluated each method using the MAE across the entire field of view as a quantitative metric.

In addition to reconstruction accuracy, we also compared the inference time of the involved methods to evaluate the computational overhead of the proposed framework. We measured the end-to-end runtime of SF-FPA and our PIST-FPA under the same input resolution and testing setting, and we further report the runtime breakdown of PIST-FPA into its main components, CNN1, OFPA, and CNN2. As shown in Table 2, PIST-FPA introduces additional cost mainly due to multi-frame background regression and optical-flow alignment, which is the computational trade-off for the accuracy and robustness improvements reported above.

The test object maintained uniform translational motion throughout the test. To assess the performance of this method, we set the noise variance to zero. As shown in the first rows of Figs. 3(a) and 3(c), the phase error distributions of the three methods differ under noise-free fringe conditions. In regions with smoother morphology, the phase errors of all three solutions approach zero. However, in areas with significant depth variations, the proposed framework demonstrates lower errors compared to traditional FTP and SF-FPA. Subsequently, to evaluate the proposed method's noise robustness, we artificially introduced Gaussian noise with variance $\sigma^2 = 2.5$ to the input fringe patterns. Given the low light intensity in the experimental setting, this noise level adequately simulates real-world interference under weak signal conditions, making it sufficient for evaluating the robustness of the method. The second rows of Figs. 3(a) and 3(c) show the phase demodulation results under noisy conditions for the three methods. Experimental results demonstrate that the FTP method is extremely sensitive to noise interference, exhibiting significant demodulation errors.

Table 1. Phase MAE under Different Time Window Lengths T and Object Motion Speeds at 10,000 FPS[a]

		MAE (rad)				
Motion Speed	**Optimal Window**	$T = 1$	$T = 3$	$T = 5$	$T = 7$	$T = 9$
10.0 m/s	**9.82**	0.0535	0.0527	0.0419	0.0410	**0.0363**
20.0 m/s	**5.41**	0.0542	0.0468	**0.0424**	0.0512	0.0539
40.0 m/s	**3.20**	0.0538	**0.0445**	0.0501	0.0560	0.0597

[a]The "Optimal Window" represents the theoretical upper bound calculated by Eq. (14).

Table 2. Comparison of Inference Time for Different Methods (CNN1, OFPA, and CNN2 are Modules of PIST-FPA) on One NVIDIA GeForce RTX 4090 Card

Method	Inference Time Breakdown (ms)	Total Time (ms)
SF-FPA	43.820 (U-Net)	43.820
PIST-FPA	25.978 (CNN1) + 22.293 (OFPA) + 30.908 (CNN2)	79.179

In contrast, both deep learning methods demonstrate superior robustness. Furthermore, unlike the SF-FPA method that relies solely on spatial features, our proposed framework achieves high phase demodulation accuracy in both smooth regions and areas with abrupt depth variations by coupling spatiotemporal information. To enable more detailed local comparisons, we selected two regions of interest (ROIs) where curvature changes dramatically. Notably, even under high-noise conditions, the proposed method successfully restores high-precision phase within the ROIs, robustly validating the effectiveness and superiority of the spatiotemporal joint mechanism in handling complex morphologies.

To further assess each method's robustness, we systematically investigated the performance of the three approaches as the noise variance varied within the range [0, 2.5]. The results are shown in Fig. 3(b). This waterfall plot clearly reveals that although the MAE of all methods increases with rising noise intensity, our proposed method consistently maintains a significant performance advantage throughout the entire test interval. Specifically, the traditional FTP method exhibits the highest error due to its inherent sensitivity to noise, with its MAE increasing from 0.133 rad to 0.141 rad. While the deep-learning-based SF-FPA method outperforms FTP, its performance is constrained by relying solely on spatial information. As the image SNR decreases, spatial features become overwhelmed by noise, causing its MAE to increase by approximately 47% from 0.034 rad. In contrast, our proposed method effectively suppresses error accumulation by introducing temporal features as critical supplementary information, limiting the error increase to 38%. Crucially, even under the highest noise conditions, its error remains lower than the initial error of SF-FPA in near-noise-free environments. This result not only validates the noise suppression effect of multi-frame temporal constraints but also fully demonstrates the exceptional robustness of our

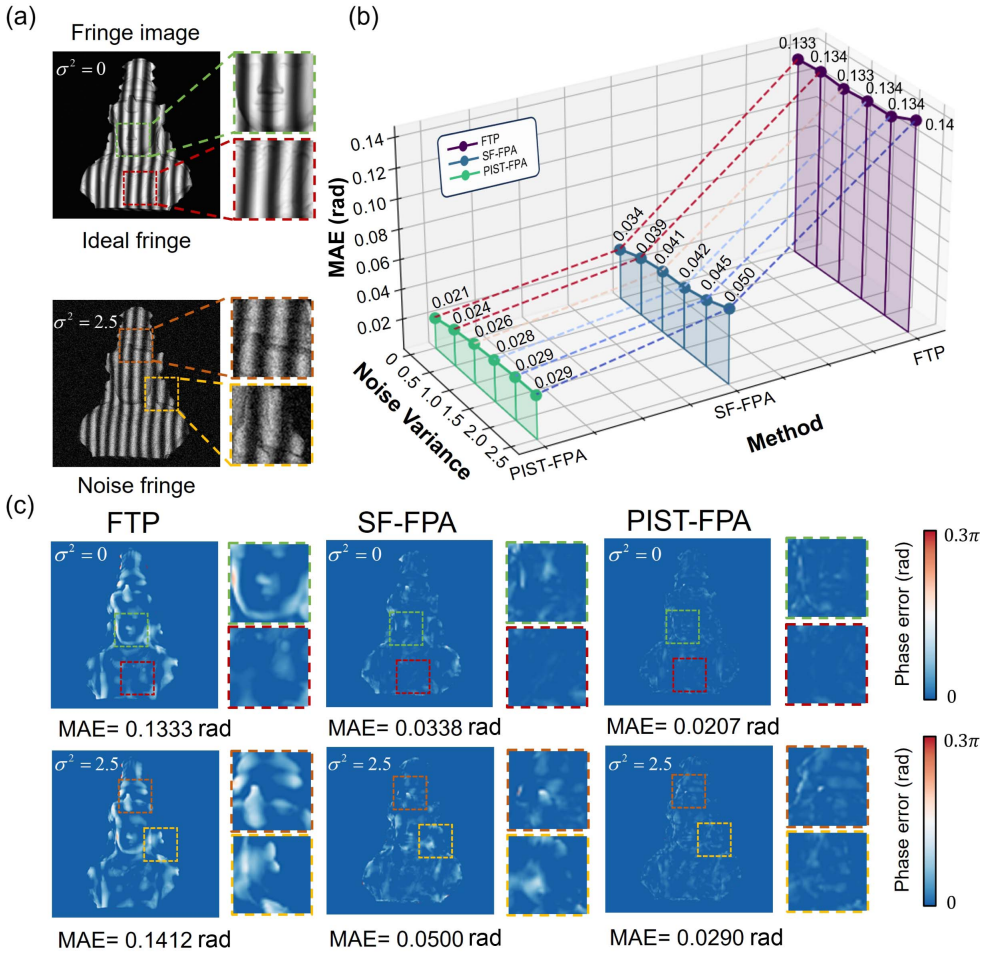

Fig. 3. Quantitative comparison of phase demodulation accuracy and noise robustness among different methods on simulated datasets. (a) Input fringe images under noise-free and noisy conditions. (b) Waterfall plot of MAE versus noise variance for the three methods—FTP, SF-FPA, and the proposed PIST-FPA. (c) Phase error distribution for FTP, SF-FPA, and PIST-FPA at $\sigma^2 = 0$ and $\sigma^2 = 2.5$.

Table 3. Phase Error Comparison between Our Method and the Single-Frame Deep Learning Method under Different Noise Variances[a]

Noise Variance	0	0.5	1	1.5	2	2.5
SF-FPA	0.0417	0.0509	0.0534	0.0552	0.0569	0.0578
PIST-FPA	0.0297	0.0350	0.0366	0.0378	0.0386	0.0394

[a]Phase error values are expressed in radians.

proposed spatiotemporal joint analysis framework in complex environments.

To further evaluate the generalization capability and universality of the proposed framework, we extended the test baseline to a large-scale dataset comprising 100 diverse and mutually non-overlapping complex dynamic scenes. We conducted a comprehensive quantitative assessment of the performance of the proposed method and the SF-FPA model using the MAE across the entire test set as the evaluation metric. As shown in Table 3, under this test baseline, our proposed method (PIST-FPA) demonstrates comprehensive superiority. Under ideal noise-free conditions, the MAE of the proposed method (0.0297 rad) is approximately 28.8% lower than that of SF-FPA (0.0417 rad), exhibiting higher intrinsic accuracy. As noise levels increase, the robustness advantage of the proposed method becomes increasingly evident. Even under the highest noise level ($\sigma^2 = 2.5$), its MAE (0.0394 rad) remains significantly lower than SF-FPA's 0.0578 rad. Notably, even in the most severe noise conditions, our method's accuracy still outperforms SF-FPA's performance under ideal noise-free conditions.

C. Validation of Spatiotemporal Joint Mechanisms with Real Data

Next, we constructed a real-world experimental platform to evaluate the practical performance of our proposed method. The experiment employed a high-speed imaging system with a frame rate set to 10,000 FPS and an extremely short exposure time of 95 μs. The images were captured and stored in an 8-bit grayscale format. According to the statistical analysis, the average grayscale of the captured data is approximately 12, with the median grayscale around 9, representing only 3.53% of the full-scale range. The overall SNR is only about 8 dB, indicating that the image was captured under low-light conditions, with a limited dynamic range and a low SNR. These extreme conditions presented a significant challenge for the 3D reconstruction algorithm.

We evaluated the method on real data by imaging a rigidly moving model at a short exposure and comparing against three representative baselines: traditional Fourier transform profilometry (FTP), data-driven single-frame deep learning (SF-FPA), and the state-of-the-art physics-informed single-frame approach PI-FPA [26]. Specifically, data was collected in a "step-and-hold" manner to simulate continuous motion. During the static phase of the object, the system projected a single-frame measurement grating as well as 12-step phase-shifted gratings for ground truth acquisition. The object was then translated to the next position, and this measurement procedure was repeated iteratively. In our implementation, we set the temporal window size to $T = 5$, constructing the network input from single-frame fringe patterns acquired at five consecutive positions to effectively capture spatiotemporal motion features. Figure 4(a) shows a representative input fringe image, together with the wrapped and absolute phase maps recovered by our proposed method. Figure 4(b) reports reconstruction results for FTP, SF-FPA, PI-FPA, the proposed PIST-FPA, and the 12-step phase-shifting result used as ground truth, with enlarged views of two detail-rich regions (ROIs A and B). Visually, the traditional FTP method exhibits severe performance degradation when processing low-SNR data, resulting in significant loss of detail in areas with steep depth variations and even noticeable artifacts. Although the deep-learning-based SF-FPA demonstrates improvements, it still exhibits local blurring and morphological distortion in high-frequency detail regions. PI-FPA achieves excellent noise suppression and clearer contour restoration over SF-FPA in smooth regions, with only slight blurring and minor deviations at sharp depth transition edges relative to the ground truth. In contrast, our proposed PIST-FPA framework exhibits outstanding detail recovery capabilities, with reconstruction results that visually align closely with the ground truth.

For a quantitative view, Figs. 4(c) and 4(d) plot X–Z cross-sections through ROI A and ROI B, respectively, comparing each method to the ground truth. The shaded orange band marks a 0.1 mm tolerance. FTP shows large excursions and SF-FPA exhibits intermediate deviations. In contrast, PI-FPA's data points closely match the ground truth within the 0.1 mm tolerance band in smooth regions, yet deviate moderately with a maximum error of 0.2 mm at depth jumps. For our PIST-FPA method, its data points lie almost entirely within the 0.1 mm tolerance band in both ROI A and ROI B, including the critical depth transition positions, thus exhibiting outstanding detail recovery capabilities with reconstruction results that align closely with the ground truth.

To further verify the ability of the proposed method for dynamic 3D measurement under extreme transient conditions, we conducted a reconstruction experiment on a high-speed rotating fan. In our implementation, we set the temporal window size to $T = 5$. As shown in Fig. 5(a), the fringe pattern sequences (I_1 to I_5) captured by the high-speed camera exhibit extremely low SNR. Specifically, the average grayscale of the captured data is approximately 8.3, with a median grayscale of 8. Even under such challenging conditions, our proposed PIST-FPA method successfully achieved high-quality dynamic 3D reconstruction of the rotating blades. The corresponding dynamic video results are detailed in Visualization 1. The reconstruction sequence below Fig. 5(a) visually demonstrates the 3D shape at five consecutive time points, with key geometric features such as the blade's contour and curvature accurately restored. More importantly, the entire dynamic process exhibits

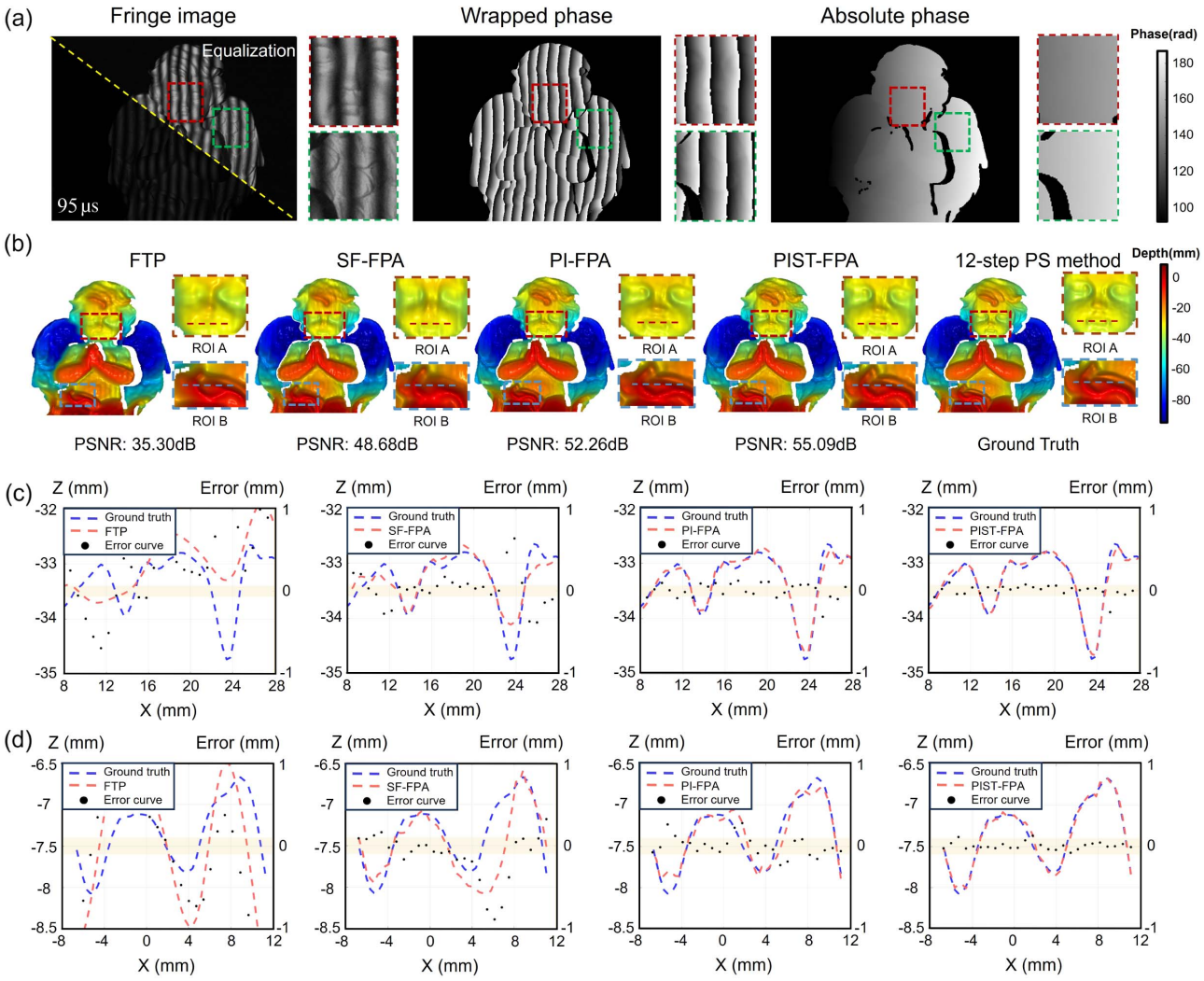

Fig. 4. 3D reconstruction accuracy and detail recovery capability were validated using real experimental data. (a) Fringe images captured with a 95 μs exposure and the wrapped and absolute phases recovered by our proposed method. (b) 3D reconstructions by the 12-step phase-shifting ground truth, FTP, SF-FPA, PI-FPA, and the proposed PIST-FPA, with enlarged views of ROI A and ROI B. (c), (d) Horizontal profiles (X–Z) of the reconstructions for ROI A and ROI B, respectively. The orange shaded band indicates a discrepancy ≤0.1 mm relative to the ground truth.

smooth transitions, demonstrating excellent spatiotemporal continuity. The cross-sectional profile analysis in Fig. 5(b) further corroborates this, with the profile curves at five time points showing high overlap and smoothness, which demonstrates the exceptional stability and consistency of the reconstruction results across the temporal dimension. Figure 5(c) provides a definitive comparison between our results and the baseline method SF-FPA. Due to the lack of temporal information constraints, the SF-FPA reconstruction surface exhibits significant rugged artifacts and irregular ripples, with severe geometric distortion even occurring at the blade edges (boxed region). In contrast, the surfaces reconstructed by our proposed method exhibit smooth continuity and effectively suppress noise. This pair of results further demonstrates that the proposed spatiotemporal joint mechanism is crucial for achieving high-fidelity 3D reconstruction under low-SNR dynamic scenes.

D. Ablation Study on the Optical-Flow Alignment Module

To evaluate the critical role of the optical-flow alignment module within the PIST-FPA framework, we designed an ablation experiment. This experiment aims to verify whether the physical prior knowledge provided by optical flow can effectively guide the network in learning spatiotemporal signal correlations, thereby enhancing reconstruction accuracy under extreme conditions. To this end, we constructed two models for comparison—PIST-FPA, the complete model proposed herein that integrates an optical-flow alignment module and represents a strategy jointly driven by physical models and data, and ST-FPA, an ablation comparison model that removes the optical-flow alignment module and serves as a purely data-driven spatiotemporal analysis network baseline. Both models were trained on the same low-SNR dataset and tested/evaluated on novel 3D dynamic scenes not present in the training set.

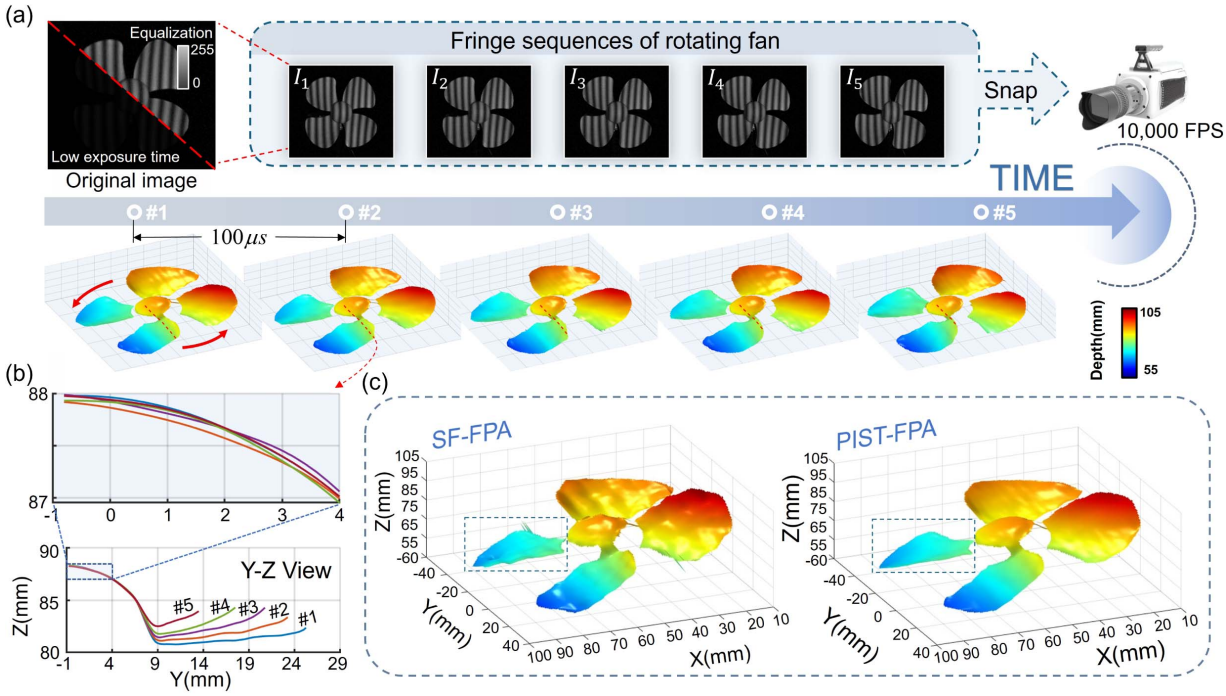

Fig. 5. 3D reconstruction results of the proposed method in dynamic scenarios. (a) From low-SNR fringe sequences (top) to their corresponding five frame continuous 3D reconstruction results (bottom). (b) Comparison of cross-sectional profiles of the reconstructed fan blade at five consecutive time points demonstrates the dynamic stability of the reconstruction results. (c) Comparison of 3D reconstruction quality from a single frame between the proposed method (PIST-FPA) and the baseline method (SF-FPA), emphasizing the significant advantages of the proposed method in noise suppression and surface detail recovery.

The experimental results are shown in Fig. 6. As depicted in the overall 3D reconstruction results of Fig. 6(a), both spatiotemporal methods effectively couple temporal information under low-SNR conditions, achieving high-precision 3D reconstruction in most smooth regions with performance superior to single-frame methods. However, in detail-rich areas, the performance gap between the two methods becomes pronounced. The PIST-FPA model, which integrates optical-flow motion constraints, yields reconstruction results that closely match the ground truth in geometrically detailed regions, demonstrating greater robustness. For a more detailed analysis, Fig. 6(b) presents magnified views of representative ROIs, and Fig. 6(c) provides quantitative cross-sectional profile comparisons along key contours. In the tangential profile of ROI A, the ST-FPA method exhibits obvious over-smoothing at the trough ($X = 5.5$ mm), with a reconstructed depth of −34.0 mm and an absolute error of approximately 0.7 mm compared to the ground truth. In contrast, our proposed PIST-FPA method reduces the error to 0.08 mm at this location, enabling accurate capture of high-frequency geometric details. In the radial profile of ROI B, as Y increases (especially in the region $Y > 12$ mm), ST-FPA exhibits significant error growth, with its Z-axis error reaching approximately 0.8 mm at $Y = 13$ mm. By comparison, the reconstruction curve of PIST-FPA closely follows the ground truth throughout, with a maximum deviation of only 0.09 mm, indicating that our method effectively improves reconstruction accuracy in edge regions. Both the local magnifications and profile curves clearly demonstrate that PIST-FPA's reconstruction results align better with the ground truth than ST-FPA. This ablation experiment demonstrates the critical role of the optical-flow alignment module within our proposed framework. By incorporating optical flow as a physical prior, we provide robust guidance for the network to learn spatiotemporal correlations between consecutive frames. This guidance enables the network to learn useful features from noise in a more targeted and efficient manner, particularly enhancing sensitivity to minute details in low-SNR images.

4. CONCLUSION

This paper addresses the fundamental trade-off between speed and accuracy for high-speed dynamic fringe analysis scenes, particularly the sharp performance degradation of existing single-frame deep learning methods under extreme conditions due to a single-frame information bottleneck. To address this challenge, this study proposes a novel spatiotemporal fringe analysis framework that integrates spatiotemporal feature decoupling with physical modeling. By jointly processing multiple fringe pattern sequences within a time window centered on the target frame, it deeply mines and utilizes the abundant temporal redundancy information contained within. Both synthetic and real experiments at 10,000 frames per second show consistent gains over Fourier-transform pipelines and state-of-the-art single-frame deep models in accuracy, noise robustness, and spatiotemporal consistency, especially under ultra-short

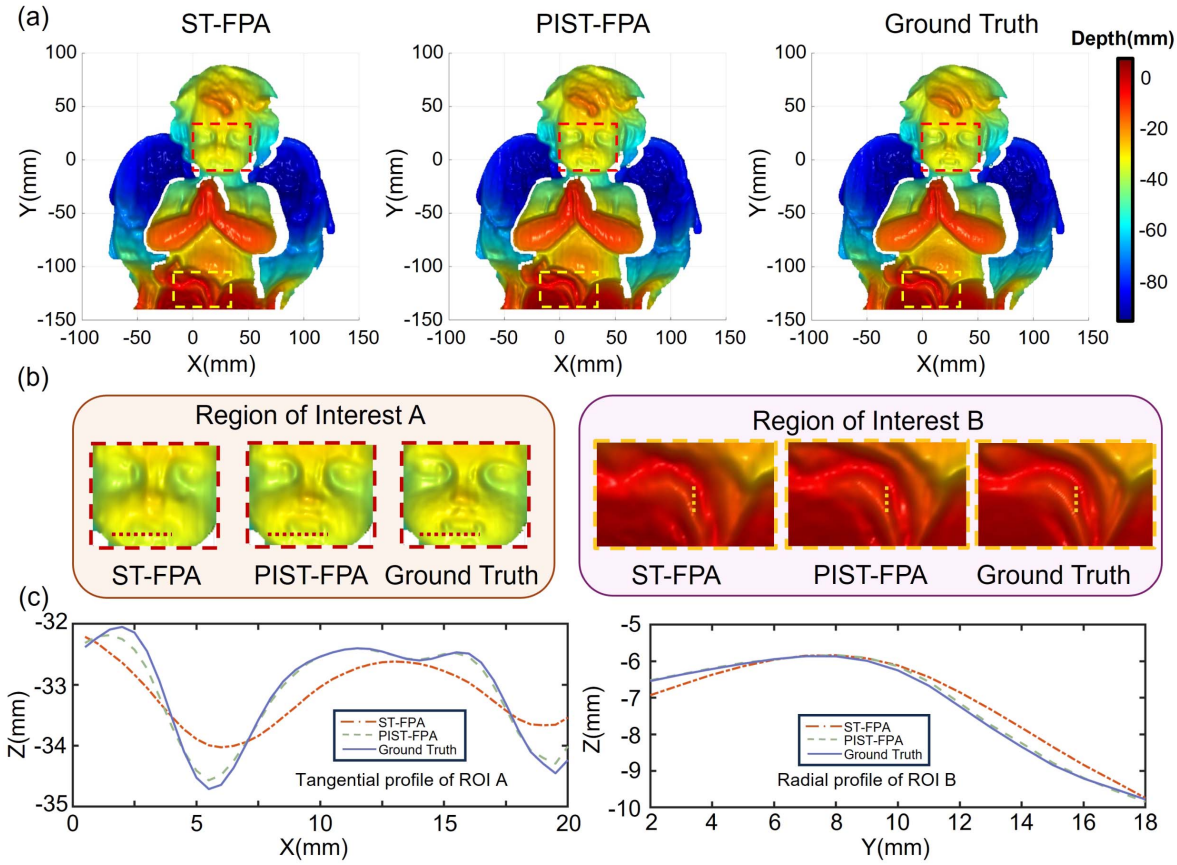

Fig. 6. Ablation study of the optical-flow alignment module. (a) Comparison of 3D reconstruction results between the proposed PIST-FPA and the reference ST-FPA methods. (b) Comparison of detail recovery capabilities between the proposed method and the reference method within the ROI region. (c) Quantitative cross-sectional profile analysis of ROIs A and B.

exposure. These merits endow it with significant potential value in emerging high-speed dynamic 3D measurement scenarios, including high-speed industrial quality inspection, biomedical dynamic tissue imaging, and robotic dynamic perception systems.

Despite these advantages, PIST-FPA involves temporal alignment. Benefiting from the powerful feature representation of the neural network, the strict requirement for precise pixel-level alignment is appropriately relaxed compared to traditional optical-flow methods, offering enhanced robustness. However, in the presence of strong non-rigid deformation, occlusion, or rapid view-dependent reflectance changes, the alignment accuracy may degrade and introduce residual artifacts. Notably, recent advances in optical flow have improved robustness to occlusion and appearance changes and have introduced strong learned-matching-based flow estimators, which could help alleviate these issues [45,46]. By bridging physical modeling and data-driven learning in a unified spatiotemporal framework, PIST-FPA establishes a generalizable paradigm for physics-informed dynamic optical metrology.

APPENDIX A: INFORMATION BOTTLENECK OF SINGLE-FRAME FRINGE MEASUREMENT

Single-frame fringe analysis is inherently limited by an information bottleneck, originating from the ill-posed nature of the inverse problem in phase retrieval. Below is the detailed mathematical modeling using the Cramér–Rao lower bound (CRLB).

For phase-measuring profilometry, the intensity of a fringe pattern captured at pixel (x, y) is defined as

$$I(x,y) = A(x,y) + B(x,y)\cos[2\pi f_0 x + \phi(x,y)] + n(x,y), \tag{A1}$$

where $A(x,y)$ is the background light intensity at pixel (x,y), $B(x,y)$ is the modulation amplitude of the fringe pattern, $\phi(x,y)$ is the target phase to be retrieved, $n(x,y) \sim \mathcal{N}(0,\sigma_n^2)$ is the additive Gaussian noise with variance σ_n^2, and f_0 is the spatial frequency of the illumination grating.

A critical observation is that for each pixel, there is only one observation $[I(x,y)]$ but three unknowns $[A(x,y), B(x,y), \phi(x,y)]$, forming an ill-posed underdetermined problem. To solve this, single-frame methods should introduce spatial priors and assuming A, B, ϕ are constant or smoothly varying within a local window (for a neural network, the receptive field). This reliance on spatial neighborhood information is the core of the single-frame information bottleneck.

To quantify the bottleneck, we derive the Fisher information matrix (FIM) and corresponding CRLB for phase estimation. Assume A, B, ϕ vary slowly within a local window containing N pixels (indexed as $k = 1, 2, \ldots, N$).

The unknown parameter vector is denoted as $\theta = [A, B, \phi]^T$, and the intensity model for the k-th pixel simplifies to

$$I_k = A + B\cos(2\pi f_0 x_k + \phi) + n_k. \tag{A2}$$

The FIM $\mathbf{J}_{\text{single}}$ for θ is calculated as the sum of the outer products of the intensity gradients with respect to θ (normalized by noise variance):

$$\mathbf{J}_{\text{single}} = \sum_{k=1}^{N} \frac{1}{\sigma_n^2} \nabla_\theta I_k \cdot \nabla_\theta I_k^T, \tag{A3}$$

where the gradient $\nabla_\theta I_k$ is

$$\nabla_\theta I_k = [1,\ \cos\theta_k,\ -B\sin\theta_k]^T, \tag{A4}$$

with $\theta_k = 2\pi f_0 x_k + \phi$, $\cos\theta_k = c_k$, and $\sin\theta_k = s_k$ for brevity.

Substituting Eq. (A4) into Eq. (A3), the FIM expands to

$$\mathbf{J}_{\text{single}} = \frac{1}{\sigma_n^2}\begin{bmatrix} \sum_{k=1}^N 1 & \sum_{k=1}^N c_k & -B\sum_{k=1}^N s_k \\ \sum_{k=1}^N c_k & \sum_{k=1}^N c_k^2 & -B\sum_{k=1}^N c_k s_k \\ -B\sum_{k=1}^N s_k & -B\sum_{k=1}^N c_k s_k & B^2\sum_{k=1}^N s_k^2 \end{bmatrix}. \tag{A5}$$

For a local window with uniformly distributed grating phases, the following approximations hold that $\sum_{k=1}^N c_k \approx 0$, $\sum_{k=1}^N s_k \approx 0$, $\sum_{k=1}^N c_k^2 \approx \sum_{k=1}^N s_k^2 \approx N/2$ (Parseval's identity), and $\sum_{k=1}^N c_k s_k \approx 0$.

Under these approximations, the FIM simplifies to a diagonal matrix:

$$\mathbf{J}_{\text{single}} \approx \frac{1}{\sigma_n^2}\begin{bmatrix} N & 0 & 0 \\ 0 & N/2 & 0 \\ 0 & 0 & B^2N/2 \end{bmatrix}. \tag{A6}$$

The CRLB states that the variance of any unbiased estimator $\hat{\phi}$ satisfies $\text{Var}(\hat{\phi}) \geq [\mathbf{J}_{\text{single}}^{-1}]_{\phi\phi}$. Thus, the CRLB for single-frame phase estimation is

$$\text{Var}_{\text{single}}(\hat{\phi}) \geq \frac{2\sigma_n^2}{B^2 N}. \tag{A7}$$

Equation (A7) quantifies the single-frame information bottleneck: phase estimation accuracy is inherently limited by the number of pixels N (constrained by the spatial window size), noise variance σ_n^2, and modulation amplitude B. In high-speed scenarios with ultra-short exposure, σ_n^2 increases sharply due to photon shot noise, while B is restricted by limited light intensity—leading to a dramatic rise in the CRLB and a fundamental upper bound on measurement accuracy.

APPENDIX B: PIST-FPA BREAKS THE INFORMATION BOTTLENECK VIA TEMPORAL REDUNDANCY

PIST-FPA overcomes the single-frame bottleneck by exploiting temporal redundancy within a time window of length T. Below is the derivation of its phase estimation CRLB and comparison with single-frame methods.

After aligning fringe sequences to the central frame using the OFPA (optical-flow pixel alignment) module, we introduce two key configurations for subsequent spatiotemporal analysis. We employ symmetric time indices defined as $t = -(T-1)/2, -(T-3)/2, \ldots, 0, \ldots, (T-3)/2, (T-1)/2$, and we assume a uniform phase variation induced by object motion, which is characterized by the relation $\phi(t) = \phi_0 + \Delta\phi \cdot t$. Here, ϕ_0 denotes the phase of the central frame and $\Delta\phi$ represents the phase increment per frame.

The aligned intensity sequence for a pixel is

$$I(t) = A + B\cos(2\pi f_0 x + \phi_0 + \Delta\phi \cdot t) + n(t), \tag{B1}$$

where $n(t) \sim \mathcal{N}(0, \sigma_n^2)$ is temporal noise, and the parameter vector is updated to $\theta = [A, B, \phi_0, \Delta\phi]^T$.

For the time window of PIST-FPA, the number of frames T is an odd integer, defined as $T = 2K + 1$. The spatiotemporal FIM $\mathbf{J}_{\text{ST}}$ is the sum of spatial and temporal intensity gradients, with the time summation range corresponding to $t = -K$ to $t = K$:

$$\mathbf{J}_{\text{ST}} = \sum_{k=1}^{N}\sum_{t=-K}^{K} \frac{1}{\sigma_n^2} \nabla_\theta I_{k,t} \cdot \nabla_\theta I_{k,t}^T, \tag{B2}$$

where the spatiotemporal gradient $\nabla_\theta I_{k,t}$ follows the physical model of aligned fringe sequences:

$$\nabla_\theta I_{k,t} = [1, c_{k,t},\ -Bs_{k,t},\ -Bts_{k,t}]^T. \tag{B3}$$

Here, $c_{k,t} = \cos(2\pi f_0 x_k + \phi_0 + \Delta\phi \cdot t)$, $s_{k,t} = \sin(2\pi f_0 x_k + \phi_0 + \Delta\phi \cdot t)$, ϕ_0 is the phase of the central frame, and $\Delta\phi$ represents the phase increment induced by object motion per frame.

We adopt the same spatial approximations as Appendix A and introduce temporal symmetry approximations adapted to the odd-frame window as $\sum_{t=-K}^{K} t s_{k,t}^2 \approx 0$ and $\sum_{t=-K}^{K} s_{k,t}^2 \approx \frac{T}{2}$.

With these approximations, $\mathbf{J}_{\text{ST}}$ simplifies to a block-diagonal matrix. The diagonal element corresponding to the central-frame phase ϕ_0 is derived as

$$[\mathbf{J}_{\text{ST}}]_{\phi_0\phi_0} = \frac{NB^2T}{2\sigma_n^2}. \tag{B4}$$

The CRLB for ϕ_0 in PIST-FPA is the inverse of Eq. (B4):

$$\text{Var}_{\text{ST}}(\hat{\phi}_0) \geq \frac{2\sigma_n^2}{B^2 NT}. \tag{B5}$$

Comparing Eq. (B5) with Eq. (A7), the phase estimation variance lower bound of PIST-FPA is

$$\text{Var}_{\text{ST}}(\hat{\phi}_0) = \frac{1}{T} \cdot \text{Var}_{\text{single}}(\hat{\phi}). \tag{B6}$$

This demonstrates that PIST-FPA reduces the phase estimation variance lower bound by a factor of T compared to single-frame methods. By mining temporal redundancy, PIST-FPA effectively increases the Fisher information, thus breaking the inherent information bottleneck of single-frame measurement from a physical perspective.

APPENDIX C: RELATIONSHIP AMONG OBJECT MOTION VELOCITY, TIME WINDOW LENGTH, AND FRAME RATE

This section establishes the relationship among object motion velocity, time window length, and camera frame rate for the PIST-FPA framework, based on optical imaging principles and Lucas–Kanade optical-flow validity conditions.

Let the camera frame rate be F, so the time interval between two consecutive frames is

$$\Delta t = \frac{1}{F}, \tag{C1}$$

where Δt has units of seconds. For an object with tangential linear velocity v_{obj} (in meters per second, m/s), the pixel displacement on the image plane between two consecutive frames Δd_{pix} is determined by the imaging system's magnification β (in pixels per meter, pixel/m):

$$\Delta d_{\text{pix}} = \beta \cdot v_{\text{obj}} \cdot \Delta t. \tag{C2}$$

Substituting Eq. (C1) into Eq. (C2) to explicitly incorporate the frame rate F, the inter-frame pixel displacement becomes

$$\Delta d_{\text{pix}} = \frac{\beta \cdot v_{\text{obj}}}{F}. \tag{C3}$$

PIST-FPA employs a center-symmetric temporal window to exploit spatiotemporal redundancy, where the window length T is an odd integer (e.g., $T = 3, 5, 7, 9$) to define a unique central reference frame. The number of frames between the window edge and the central frame is

$$K = \frac{T-1}{2}, \tag{C4}$$

where K is a non-negative integer representing the window half-width. The maximum cumulative pixel displacement of the object relative to the central frame (D_{max}, in pixels) is the product of K and Δd_{pix}:

$$D_{\text{max}} = K \cdot \Delta d_{\text{pix}}. \tag{C5}$$

Substituting Eqs. (C3) and (C4) into Eq. (C5), the cumulative displacement is derived as

$$D_{\text{max}} = \frac{T-1}{2} \cdot \frac{\beta \cdot v_{\text{obj}}}{F}. \tag{C6}$$

Frame alignment in PIST-FPA relies on the LK optical-flow algorithm, whose validity hinges on the first-order Taylor expansion of the brightness constancy constraint. The empirical threshold for robust alignment can be written as δ_{max}. Beyond this approximate range, the Taylor expansion approximation may no longer hold well, potentially leading to notable degradation in alignment accuracy. To maintain the physical consistency and reliability of the PIST-FPA framework,

$$D_{\text{max}} \le \delta_{\text{max}}. \tag{C7}$$

Substituting Eq. (C6) into Eq. (C7) yields the core constraint

$$\frac{T-1}{2} \cdot \frac{\beta \cdot v_{\text{obj}}}{F} \le \delta_{\text{max}}. \tag{C8}$$

Rearranging to solve for T, the theoretical upper bound is

$$T \le \frac{2 \cdot \delta_{\text{max}} \cdot F}{\beta \cdot v_{\text{obj}}} + 1. \tag{C9}$$

We integrate the pinhole camera model with the camera's calibrated intrinsic parameters. Specifically, f_x and f_y denote the camera's focal lengths along the x- and y-directions (in pixels, obtained via standard camera calibration), Z represents the object's depth relative to the image plane (in millimeters, mm), and v_x and v_y are the tangential velocity components of the object at depth Z (in millimeters per second, mm/s).

The pixel velocity components on the image plane s_x and s_y follow the pinhole camera model:

$$s_x = \frac{f_x}{Z} \cdot v_x, \quad s_y = \frac{f_y}{Z} \cdot v_y, \tag{C10}$$

with total pixel velocity magnitude:

$$s = \sqrt{s_x^2 + s_y^2} = \frac{1}{Z}\sqrt{(f_x v_x)^2 + (f_y v_y)^2}. \tag{C11}$$

Since $s = \beta \cdot v_{\text{obj}}$, substituting $v_{\text{obj}} = s/\beta$ into Eq. (C11) yields the relationship

$$T \le \frac{2 \cdot \delta_{\text{max}} \cdot F \cdot Z}{\sqrt{(f_x v_x)^2 + (f_y v_y)^2}} + 1. \tag{C12}$$

To ensure reliable motion compensation, this cumulative displacement should remain within the linear valid range of the optical-flow alignment algorithm (typically denoted as $\delta_{\text{max}} \approx 1$ to 2 pixels). Notably, this physical constraint can be extended by leveraging the error correction capability of the deep learning module in our framework: even if the cumulative displacement exceeds the linear range of optical flow, the neural network can compensate for alignment errors via learned phase priors. We thus introduce a neural network correction factor γ to characterize this extension:

$$T \le \frac{2 \cdot \gamma \cdot \delta_{\text{max}} \cdot F \cdot Z}{\sqrt{(f_x v_x)^2 + (f_y v_y)^2}} + 1. \tag{C13}$$

This expression explicitly quantifies the intrinsic relationship among object motion velocity, temporal window size, and camera frame rate.

Funding. National Natural Science Foundation of China (62522508, U21B2033, 62205147, 62571249); Fundamental Research Funds for the Central Universities (2023102001, 2024202002); National Key Laboratory of Shock Wave and Detonation Physics (JCKYS2024212111); China Postdoctoral Science Fund (2023T160318); Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense (JSGP202105, JSGP202201); Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX25_0695, SJCX25_0188).

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. K. J. Gåsvik, *Optical Metrology* (Wiley, 2003).
2. T. Kreis, *Handbook of Holographic Interferometry: Optical and Digital Methods* (Wiley, 2006).
3. P. K. Rastogi, *Digital Speckle Pattern Interferometry & Related Techniques* (Wiley, 2000).
4. S. S. Gorthi and P. Rastogi, "Fringe projection techniques: whither we are?" Opt. Lasers Eng. **48**, 133–140 (2010).

5. G. Sansoni, M. Trebeschi, and F. Docchio, "State-of-the-art and applications of 3D imaging sensors in industry, cultural heritage, medicine, and criminal investigation," Sensors **9**, 568–601 (2009).
6. F. Remondino, "Heritage recording and 3D modeling with photogrammetry and 3D scanning," Remote Sensing **3**, 1104–1138 (2011).
7. K. R. Ford, G. D. Myer, and T. E. Hewett, "Reliability of landing 3D motion analysis: implications for longitudinal analyses," Med. Sci. Sports Exercise **39**, 2021–2028 (2007).
8. X. Su and Q. Zhang, "Dynamic 3-D shape measurement method: a review," Opt. Lasers Eng. **48**, 191–204 (2010).
9. M. Kujawinska and W. Osten, "Fringe pattern analysis methods: up-to-date review," Proc. SPIE **3407**, 56–66 (1998).
10. K. Qian, "Two-dimensional windowed Fourier transform for fringe pattern analysis: principles, applications and implementations," Opt. Lasers Eng. **45**, 304–317 (2007).
11. J. Zhong and J. Weng, "Spatial carrier-fringe pattern analysis by means of wavelet transform: wavelet transform profilometry," Appl. Opt. **43**, 4993–4998 (2004).
12. L. Huang, K. Qian, B. Pan, *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**, 141–148 (2010).
13. Z. Zhang, Z. Jing, Z. Wang, *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**, 1152–1160 (2012).
14. C. Zuo, S. Feng, L. Huang, *et al.*, "Phase shifting algorithms for fringe projection profilometry: a review," Opt. Lasers Eng. **109**, 23–59 (2018).
15. S. Feng, C. Zuo, T. Tao, *et al.*, "Robust dynamic 3-D measurements with motion-compensated phase-shifting profilometry," Opt. Lasers Eng. **103**, 127–138 (2018).
16. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436–444 (2015).
17. D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," Annu. Rev. Biomed. Eng. **19**, 221–248 (2017).
18. M. Bakator and D. Radosav, "Deep learning and medical diagnosis: a review of literature," Multimodal Technol. Interaction **2**, 47 (2018).
19. T. Young, D. Hazarika, S. Poria, *et al.*, "Recent trends in deep learning based natural language processing," IEEE Comput. Intelligence Mag. **13**, 55–75 (2018).
20. D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning for natural language processing," IEEE Trans. Neural Netw. Learning Syst. **32**, 604–624 (2020).
21. A. Voulodimos, N. Doulamis, A. Doulamis, *et al.*, "Deep learning for computer vision: a brief review," Comput. Intelligence Neurosci. **2018**, 7068349 (2018).
22. G. Barbastathis, A. Ozcan, and G. Situ, "On the use of deep learning for computational imaging," Optica **6**, 921–943 (2019).
23. C. Zuo, J. Qian, S. Feng, *et al.*, "Deep learning in optical metrology: a review," Light Sci. Appl. **11**, 39 (2022).
24. S. Feng, Q. Chen, G. Gu, *et al.*, "Fringe pattern analysis using deep learning," Adv. Photonics **1**, 025001 (2019).
25. S. Feng, C. Zuo, L. Zhang, *et al.*, "Generalized framework for non-sinusoidal fringe analysis using deep learning," Photonics Res. **9**, 1084–1098 (2021).
26. W. Yin, Y. Che, X. Li, *et al.*, "Physics-informed deep learning for fringe pattern analysis," Opto-Electronic Adv. **7**, 230034 (2024).
27. X. Li, S. Feng, W. Chen, *et al.*, "Adaptive structured-light 3D surface imaging with cross-domain learning," Laser Photonics Rev. **19**, 2401609 (2025).
28. W. Chen, S. Feng, W. Yin, *et al.*, "Deep-learning-enabled temporally super-resolved multiplexed fringe projection profilometry: high-speed khz 3D imaging with low-speed camera," PhotoniX **5**, 25 (2024).
29. W. Chen, Y. Liu, S. Feng, *et al.*, "Dual-frequency angular-multiplexed fringe projection profilometry with deep learning: breaking hardware limits for ultra-high-speed 3D imaging," Opto-Electronic Adv. **8**, 250021 (2025).
30. B. Wang, W. Chen, J. Qian, *et al.*, "Single-shot super-resolved fringe projection profilometry (SSSR-FPP): 100,000 frames-per-second 3D imaging with deep learning," Light Sci. Appl. **14**, 70 (2025).
31. S. Feng, C. Zuo, W. Yin, *et al.*, "Micro deep learning profilometry for high-speed 3D surface imaging," Opt. Lasers Eng. **121**, 416–427 (2019).
32. W. Yin, Q. Chen, S. Feng, *et al.*, "Temporal phase unwrapping using deep learning," Sci. Rep. **9**, 20175 (2019).
33. S. Feng, Y. Xiao, W. Yin, *et al.*, "Fringe-pattern analysis with ensemble deep learning," Adv. Photonics Nexus **2**, 036010 (2023).
34. S. Zhang, "High-speed 3D shape measurement with structured light methods: a review," Opt. Lasers Eng. **106**, 119–131 (2018).
35. C. Chen, Q. Chen, J. Xu, *et al.*, "Learning to see in the dark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 3291–3300.
36. 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* (2015), pp. 234–241.
37. J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," Int. J. Comput. Vis. **12**, 43–77 (1994).
38. C. Zuo, L. Huang, M. Zhang, *et al.*, "Temporal phase unwrapping algorithms for fringe projection profilometry: a comparative review," Opt. Lasers Eng. **85**, 84–103 (2016).
39. O. Oktay, J. Schlemper, L. L. Folgoc, *et al.*, "Attention U-Net: learning where to look for the pancreas," arXiv, arXiv:1804.03999 (2018).
40. A. Puljčan, D. Zoraja, and T. Petković, "Simulation of structured light 3D scanning using Blender," in *International Symposium ELMAR* (2022), pp. 215–220.
41. Y. Liu, W. Chen, J. Jiang, *et al.*, "Digital-twin-driven unambiguous structured light 3D imaging with physics-aware learning," npj Nanophoton. **2**, 45 (2025).
42. Q. Zhou and A. Jacobson, "Thingi10K: a dataset of 10,000 3D-printing models," arXiv, arXiv:1605.04797 (2016).
43. M. Takeda and K. Mutoh, "Fourier transform profilometry for the automatic measurement of 3-D object shapes," Appl. Opt. **22**, 3977–3982 (1983).
44. C. Zuo, T. Tao, S. Feng, *et al.*, "Micro Fourier transform profilometry (μFTP): 3D shape measurement at 10,000 frames per second," Opt. Lasers Eng. **102**, 70–91 (2018).
45. S. Yuan, L. Luo, Z. Hui, *et al.*, "UnSAMFlow: unsupervised optical flow guided by segment anything model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 19027–19037.
46. W. Dai, H. Wu, X. Weng, *et al.*, "Multi-modal synergistic implicit image enhancement for efficient optical flow estimation," in *Proceedings of the Computer Vision and Pattern Recognition Conference* (2025), pp. 2173–2182.