OE Optical Engineering

RESEARCH PAPER

STU-Net: Swin Transformer U-Net for high-throughput live cell analysis with a lens-free on-chip digital holographic microscope

Wenhui Lin,[a,b,c,†] Yang Chen,[a,b,c,†] Xuejuan Wu,[a,b,c,†] Yufan Chen,[a,b,c]
Yanyan Gao,[a,b,c] and Chao Zuo[a,b,c,*]
[a]Nanjing University of Science and Technology, School of Electronic and Optical Engineering, Smart Computational Imaging Laboratory, Nanjing, China
[b]Nanjing University of Science and Technology, Smart Computational Imaging Research Institute, Nanjing, China
[c]Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense, Nanjing, China

ABSTRACT. A lens-free on-chip digital holographic microscope (LFOCDHM) is essential for a variety of biomedical applications such as cell cycle assays, drug development, digital pathology, and high-throughput biological screening. However, due to the unit magnification configuration of the lens-free system, the field-of-view (FOV) contains over a hundred times more cells than a conventional 10× microscope objective. Consequently, the segmentation process becomes labor-intensive and time-consuming due to the complex and variable morphology of cells within the large FOV. To address this issue, numerous deep learning-based cell segmentation methods have been proposed. Nevertheless, convolutional neural networks, limited by their localized receptive field, are unsuitable for segmenting and processing large FOV imaging results from LFOCDHM. Therefore, we propose a high-throughput live cell analysis processing method called Swin Transformer U-Net (STU-Net). Based on the reconstructed phase results, a shift window is utilized to compute the self-attention to extract its features at five scales, which can compute the normalized inner distance and pixel-level classification and achieve high-throughput accurate cell segmentation (accuracy >0.9743). We validated the robustness and generalizability of our STU-Net by the accurate segmentation of data from HeLa cell slides across the full FOV and live C166 cells *in vitro*. Given its capability for quantifying cell growth and proliferation based on the multi-cell parameters generated from segmentation results, the proposed approach is expected to provide a strong foundation for subsequent drug development and biological screening.

[DOI: 10.1117/1.OE.63.11.111812]

Keywords: lens-free; high throughout; cell segmentation; Swin Transformer; multi-task learning

Paper 20240254SS received Mar. 21, 2024; revised Aug. 17, 2024; accepted Sep. 5, 2024; published Sep. 30, 2024.

1 Introduction

Quantitative phase imaging (QPI)[1–5] has become an important optical tool in biomedical research by imaging optical thickness changes in live cells and tissues without specific staining. However, most of the QPI methods are based on convolutional microscopes,[6–11] which suffer from the

*Address all correspondence to Chao Zuo, zuochao@njust.edu.cn

†These authors contributed equally to this work.

0091-3286/2024/$28.00 © 2024 SPIE

inherent trade-off between the field-of-view (FOV) and imaging resolution due to the limitation of the spatial bandwidth product.[12] With the technological innovations of photoelectric sensors, the emergence and rapid development of lens-free on-chip digital holographic microscope (LFOCDHM) in the last few years have provided a promising solution to the above-mentioned problems.[13–17] The LFOCDHM can be built directly into the incubator for *in situ* observation of *in vitro* cultured cells due to its compactness.[18]

LFOCDHM is primarily used for pre-experiments and live cell growth experiments, offering unique advantages in drug development and biological screening through its dynamic monitoring capabilities.[1,19,20] It can accurately quantify and monitor the effects of various drug concentrations or classes on live cell cultures in real time, providing dynamic data on changes in cell morphology and activity at different time points. These macroscopic responses to drug effects provide crucial experimental data for further molecular and genetic research. Among them, parameters such as cell number, area, perimeter, concavity, area-perimeter ratio, and aspect ratio are key indicators to describe the cell growth status, which are obtained from cell segmentation results.[21,22] Traditional cell segmentation algorithms include threshold segmentation,[23] Canny edge detection,[24] and watershed algorithms.[25,26] Segmenting morphologically diverse cells in the FOV accurately using traditional methods requires adjustment and optimization based on the actual situation, which can be time-consuming and computationally intensive.

To address the above problem, deep learning-based segmentation methods can automatically learn high-level features in image data through pre-training to obtain an end-to-end cell segmentation model.[27–29] For example, the U-Net-based biomedical image segmentation method, proposed by Ronneberger et al.,[30] treats cell segmentation as a binary classification problem and trains the model to act as an end-to-end classifier to distinguish cells from background pixels. Despite the strong representation learning capability of such U-Net methods, their performance in learning features is limited to their local receptive fields. As a result, this shortcoming in capturing multi-scale information leads to suboptimal segmentation of structures of variable shape and size (e.g., cells of different sizes). Unlike methods such as U-Net that treat cell segmentation as a classification problem, Koyuncu et al.[31] treated cell segmentation as a regression problem and proposed a deep-distance network model based on multi-task learning[32,33] with shared encoder paths. Their work focused on the problem of detecting cells, aiming to identify cell locations in unlabelled images without identifying the exact boundaries of the cells.

To solve the aforementioned problems, we propose Swin Transformer U-Net (STU-Net) with multi-task decoding paths. By symmetrically skip connection to connect coded features at different scales, STU-Net learns the normalized internal distance (NID)[31,34] and pixel-level classification (PLC) to refine cell boundaries using two parallel decoder paths. Based on our proposed network, we successfully realized the precise segmentation of HeLa cells across a large FOV of 19.5 mm^2 (accuracy >0.9743). Finally, we performed consistent segmentation of the dynamic data of C166 cells over an extended period and generated multidimensional cellular parameters based on the segmentation results to quantify cell proliferation and growth. Our method can accurately and stably segment wide-field cell results of LFOCDHM, providing a strong guarantee for drug development and biological screening.

2 Method

2.1 Overview of Method

Our method enables intelligent analysis from a single-shot hologram to wide-field cellular results based on deep learning on the LFOCDHM system. The schematic diagram of LFOCDHM is shown in Fig. 1(a). It does not contain any objective lens and can be placed in an incubator for *in situ* living cell observation. It consists of two fundamental components: a complementary metal oxide semiconductor (CMOS) sensor (5664×4256, pixel size: 0.9 μm, 24,000 pixels, Jiangsu Team one Intelligent Technology Co., Ltd.) and a quasi-monochromatic light-emitting diode (LED) that can emit a wavelength of 623 nm. The LED light wave travels roughly ($Z_1 \sim 90$ mm) to interact with the sample, generating a diffraction pattern. The diffraction pattern is recorded by a CMOS sensor, placed close to the sample ($Z_2 \sim 1000$ μm). The cell slides are directly placed on the sensor plane to achieve diffraction patterns, resulting in twin image interference due to a lack of phase information. The cell slides are directly placed on the sensor plane to achieve diffraction

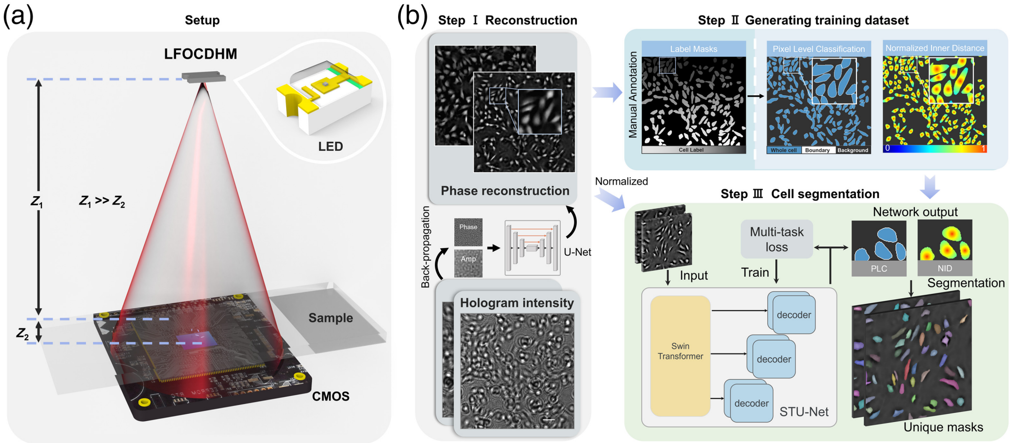


Fig. 1 LFOCDHM setup and overview of cell segmentation process based on lens-free phase images. (a) Schematic diagram of LFOCDHM, where Z_1 is the distance from the LED light source to the sample, and Z_2 is the distance from the sample to the CMOS sensor. (b) Step I: Reconstruction. The amplitude and phase images obtained by free space backpropagation of a single hologram intensity are used as U-Net inputs to implement phase reconstruction. Step II: Generate training datasets. Converting manually annotated cell label masks to PLC and NID. Step III: Cell segmentation. Predict PLC and NID using STU-Net and estimate cell boundary realization based on the prediction using a region-growing algorithm.

patterns, resulting in twin image interference due to a lack of phase information. To solve this problem, we utilized a framework for phase reconstruction based on deep learning,[35] as shown in Fig. 1(b). The initial stage of the framework consists of training U-Net. The training involves learning the statistical transformation between amplitude and phase images obtained from a single hologram intensity through the free-space backpropagation of the object and the same object's image that is reconstructed using a multi-height phase retrieval algorithm (treated as the gold standard for the training phase). Specifically, the multi-height retrieval algorithm first acquires 10 raw images with an axial step size of 3μm and then calculates the reconstruction of phase images through the adaptive pixel-super-resolved lens-free imaging method.[36] The hyperparameters, including a dynamically adjusted learning rate initially set to 0.001, batch size of 16, 100 epochs, and Adam optimizer, were fine-tuned using grid search and manual adjustments based on validation performance. The training process, performed only once, yields a fixed U-Net used for blind reconstruction of phase images of any object using a single hologram intensity.

The pixels of each cell unit were manually annotated within the reconstructed phase results. This manual annotation enabled us to derive NID and PLC based on these label masks. NID is used to locate cells and identify cell boundaries. PLC calibrates the position of each pixel relative to the cell, distinguishing among the whole cell, cell boundary, and background.

To compute NID, first, for each intracellular pixel p, its distance r to the center of mass pixel (x_c, y_c) of the cell unit in which it is located is calculated

$$r = \sqrt{(x_p - x_c)^2 + (y_p - y_c)^2}. \tag{1}$$

After that, the NID of that pixel point is calculated by the distance r

$$\mathrm{NID} = \frac{1}{1 + \beta r/\sqrt{S_c}}, \tag{2}$$

where S_c is the number of pixels in the cell unit where it is located, which makes NID independent of differences in cell size. Based on experimental results and empirical observations, we set β as a hyperparameter to adjust the data distribution during the distance transformation, taking its value as 1.

Finally, we use STU-Net as an end-to-end model, trained with ∼1000 experimentally collected images from living cell experiments, to process the reconstructed phase results and generate two sets of predictions with the same size as the input image: NID and PLC. Then, we apply a Gaussian filter smoothing transformation to NID to remove small abrupt changes. Next, we used a peak-finding algorithm to locate the center of mass of each cell to determine the location of the cell. After that, we combine cell plasmas, PLC, and NID to further optimize the details of cell boundaries by region-growing algorithm.[37,38]

2.2 Swin Transformer U-Net

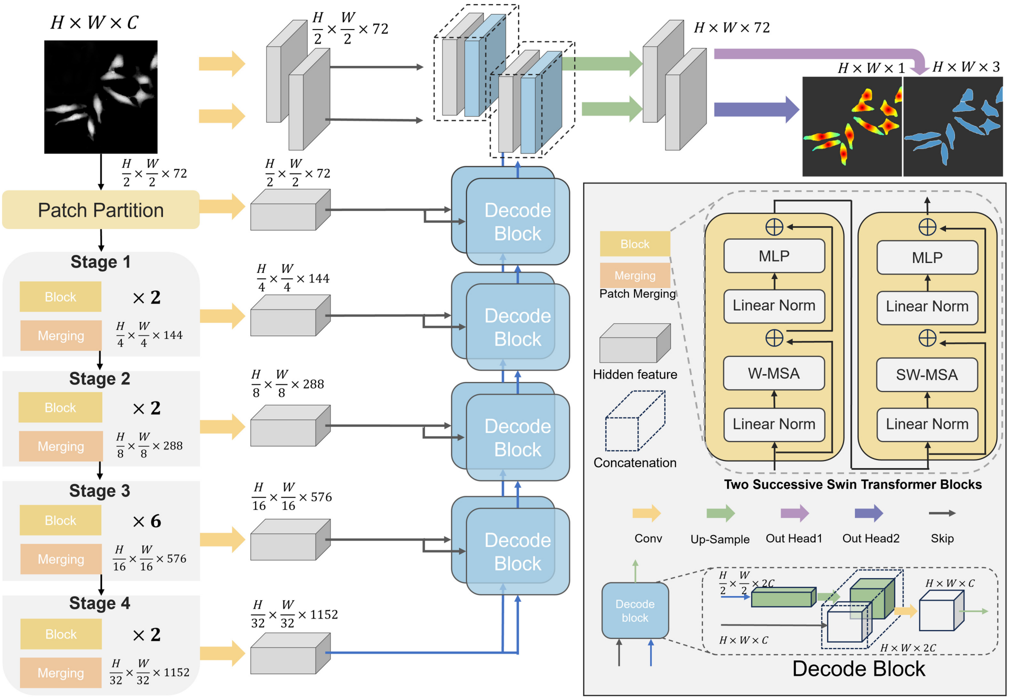
The convolution-based U-Net architecture has an inherent limitation in capturing long-range spatial relations.[39] Similarly, the Swin Transformer architecture has limitations in capturing low-level features.[40] It has been demonstrated that both local and global information are crucial for dense prediction tasks, such as segmentation in challenging contexts.[41,42] Consequently, STU-Net has been proposed as a hybrid model that effectively combines U-Net and Swin Transformer for cell segmentation.

2.2.1 *Encoder*

The self-attention module in Vision Transformer allows for modeling long-range information through the pairwise interaction among token embeddings, leading to more effective local and global contextual representations.[43,44] Swin transformers[40,45] have been proposed as a hierarchical vision transformer that computes self-attention in an efficient shifted window partitioning scheme, which reduces the computational complexity dramatically by decreasing the relationship with the pixel size squared to linear, making it more suitable for processing lens-free wide-field images. The architecture of STU-Net is shown in Fig. 2 using the Swin Transformer as a feature extractor in the encoder. Swin Transformer computes self-attention according to

$$\text{Attention }(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{3}$$

where Q, K, and V denote queries, keys, and values, respectively; d represents the size of the query and key.



$H \times W \times C$
$\frac{H}{2} \times \frac{W}{2} \times 72$
$H \times W \times 72$
$H \times W \times 1$
$H \times W \times 3$
$\frac{H}{2} \times \frac{W}{2} \times 72$
$\frac{H}{2} \times \frac{W}{2} \times 72$
Patch Partition
Stage 1
Block
× 2
Merging
$\frac{H}{4} \times \frac{W}{4} \times 144$
$\frac{H}{4} \times \frac{W}{4} \times 144$
Stage 2
Block
× 2
Merging
$\frac{H}{8} \times \frac{W}{8} \times 288$
$\frac{H}{8} \times \frac{W}{8} \times 288$
Stage 3
Block
× 6
Merging
$\frac{H}{16} \times \frac{W}{16} \times 576$
$\frac{H}{16} \times \frac{W}{16} \times 576$
Stage 4
Block
× 2
Merging
$\frac{H}{32} \times \frac{W}{32} \times 1152$
$\frac{H}{32} \times \frac{W}{32} \times 1152$
Decode Block
Decode Block
Decode Block
Decode Block
Block
Merging
Patch Merging
Hidden feature
Concatenation
MLP
Linear Norm
W-MSA
Linear Norm
MLP
Linear Norm
SW-MSA
Linear Norm
Two Successive Swin Transformer Blocks
Conv
Up-Sample
Out Head1
Out Head2
Skip
Decode block
$\frac{H}{2} \times \frac{W}{2} \times 2C$
$H \times W \times C$
$H \times W \times C$
$H \times W \times 2C$
Decode Block

Fig. 2 Overview of the STU-Net architecture. STU-Net uses Swin Transformer as a feature extractor with U-shaped architecture. The STU-Net creates non-overlapping patches of the input data and uses a patch partition layer to create windows with a desired size for computing the self-attention. The encoded feature representations of different scales in ST are concatenated to multi-task decoding paths through symmetric skip connections. The final output consists of NID and PLC.

The window-based self-attention module lacks interconnections among windows, restricting its modeling capabilities. To incorporate cross-window connections while preserving the efficient computation of non-overlapping windows, Swin Transformer proposes a shifted window partitioning approach, which involves alternating between two partitioning configurations in consecutive Swin Transformer blocks. Two successive Swin Transformer blocks are computed as follows:

$$\begin{aligned} \hat{\mathbf{z}}^l &= \text{W-MSA}(\text{LN}(\mathbf{z}^{l-1})) + \mathbf{z}^{l-1}, \\ \mathbf{z}^l &= \text{MLP}(\text{LN}(\hat{\mathbf{z}}^l)) + \hat{\mathbf{z}}^l, \\ \hat{\mathbf{z}}^{l+1} &= \text{SW-MSA}(\text{LN}(\mathbf{z}^l)) + \mathbf{z}^l, \\ \mathbf{z}^{l+1} &= \text{MLP}(\text{LN}(\hat{\mathbf{z}}^{l+1})) + \hat{\mathbf{z}}^{l+1}, \end{aligned} \tag{4}$$

where W-MSA and SW-MSA are regular and window partitioning multi-head self-attention modules, respectively; $\hat{\mathbf{z}}^l$ and $\mathbf{z}^l$ denote the output features of the (S)W-MSA module and the multi-layer perceptron (MLP) module for block l, respectively; and LN denotes layer normalization.

Swin Transformer constructs a hierarchical representation by starting from small-sized patches and gradually merging neighboring patches in deeper transformer layers. Each stage provides features at different scales. With output scales of $\frac{H}{4} \times \frac{W}{4} \times 144$, $\frac{H}{8} \times \frac{W}{8} \times 288$, $\frac{H}{16} \times \frac{W}{16} \times 576$ and $\frac{H}{32} \times \frac{W}{32} \times 1152$ for stages 1 to 4, respectively, the features can be efficiently combined with the U-Net network, which is also of a multilayered structural type.

2.2.2 *Decoder*

The original U-Net model is designed for single-task learning and consists of a single decoder path. In contrast, the STU-Net model employs two decoder paths for multi-task learning, utilizing the shared features generated by the encoder. These two paths correspond to the tasks of NID and PLC, respectively. Simultaneous skip connections facilitate the recovery of spatial information that is lost during the downsampling process of the encoder. The features extracted by the decoder will undergo distinct processing mechanisms via separate output headers because the outputs of the two tasks vary. The regression map resulting from the NID task will have the same size as the input image, and the output of the PLC task will consist of three category probability maps, each containing three categories.

2.2.3 *Loss function*

The multi-task loss function comprises two primary components: the root mean square error loss function,[46] which is used to calculate NID, and the cross-entropy loss function,[47] which is used for PLC. The two components of the loss function are calculated as follows:

$$\mathcal{L}_{\text{RMSE}} = \sqrt{\frac{1}{N \times M} \sum_{y=1}^{N} \sum_{x=1}^{M} (I(x, y) - \hat{I}(x, y))^2}, \tag{5}$$

where $N \times M$ is the total number of pixels, $I(x, y)$ is the pixel value at the (x, y) position in the truth image, and $\hat{I}(x, y)$ is the pixel value at the (x, y) position in the predicted image.

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N \times M} \sum_{y=1}^{N} \sum_{x=1}^{M} \sum_{i=1}^{K} I(x, y, i) \log \hat{I}(x, y, i), \tag{6}$$

where K is the total number of categories, i.e., it is divided into three categories containing whole cell, cell boundary, and background, and $\hat{I}(x, y, i)$ is the predicted probability of the pixel at the (x, y) location corresponding to category i. The overall loss function is defined as follows:

$$\mathcal{L} = \alpha_1 \mathcal{L}_{\text{RMSE}} + \alpha_2 \mathcal{L}_{\text{CE}}, \tag{7}$$

where $\alpha_{1,2}$ are the weight coefficients. Here, $\alpha_{1,2}$ are set as hyperparameters to 0.8 and 0.2 for balancing the multi-task output loss imbalance problem and predicting NID as the main task, respectively.[33,48] When the discrepancy in loss among different tasks is substantial, the network prioritizes convergence toward the task with a higher loss. Setting $\alpha_{1,2}$ facilitates achieving a balanced weighting for multi-task learning, enabling better management of differences and imbalances among multi-tasks and enhancing the STU-Net performance and prediction capabilities.[48]

3 Results

To evaluate the cell segmentation capability of our method in processing wide-field LFOCDHM images, we performed full FOV segmentation experiments. As shown in Fig. 3(a1), we demonstrated the segmentation results of HeLa cells across the full FOV (19.5 mm^2), with

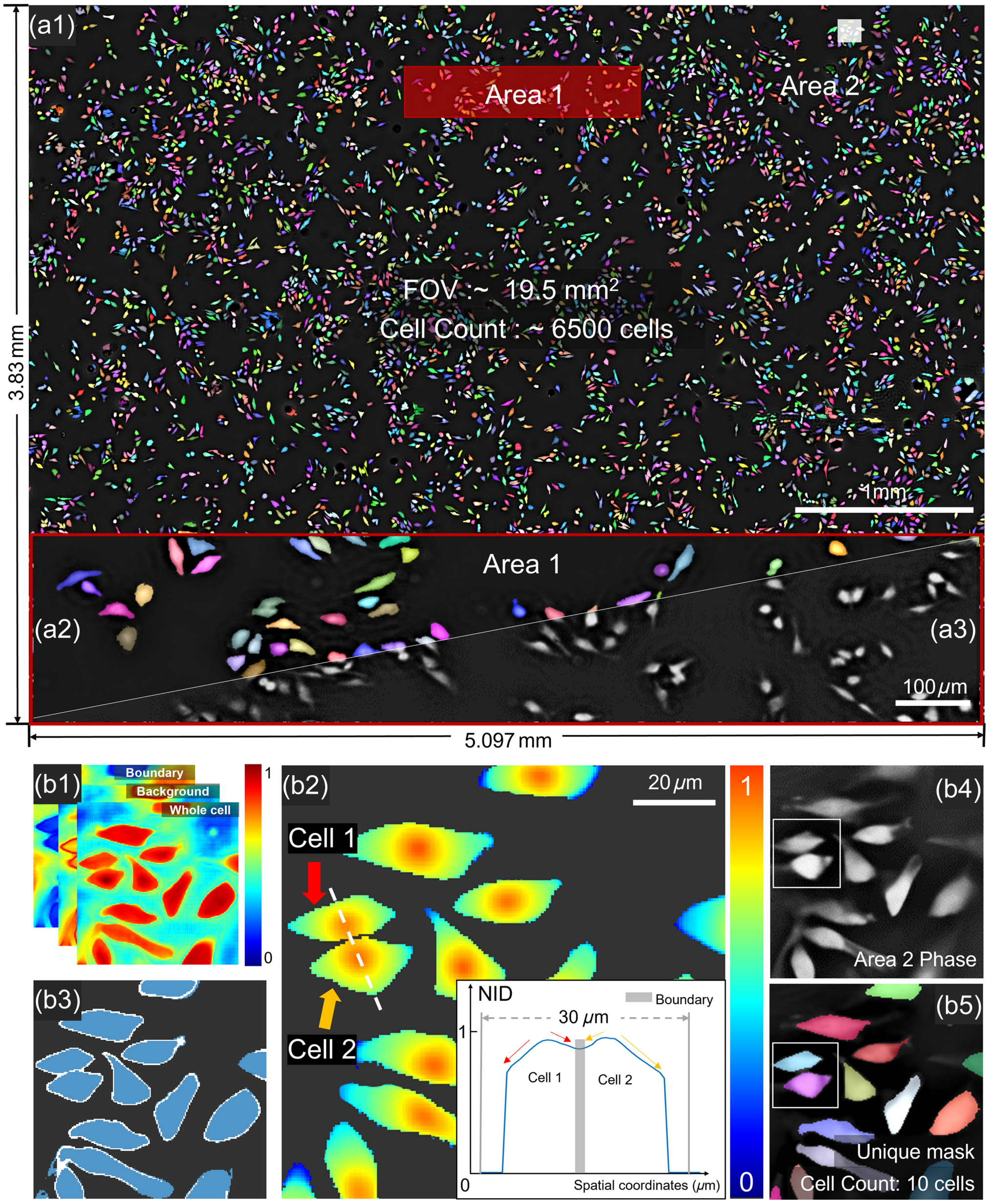


Fig. 3 STU-Net provides accurate segmentation results across the full FOV. (a1) Approximately 6500 cells are identified in the 19.5 mm^2 full FOV. (a2) Each cell can be observed with fine cell edges in area 1. (a3) Reconstructed phase results in area 1 (b1) Corresponding three classification probabilities for area 2. (b2) NID of area 2. (b3) PLC of area 2. (b4) Reconstructed phase results of area 2. (b5) Unique mask for each cell.

~6500 HeLa cells. HeLa cells were cultured in 20 mm glass bottom dishes with 10% fetal bovine serum. The delicate outlines of cells in area 1 were successfully segmented, and the interconnected cells were efficiently separated from each other, as illustrated in Fig. 3(a2). Based on the reconstructed phase results in Fig. 3(b4), its NID and PLC results are calculated using the STU-Net shown in Figs. 3(b1)–3(b3). It is segmented into cell instances by region growth algorithm and a unique mask is assigned to each cell as in Fig. 3(b5).

Figure 4 shows the comparison results of STU-Net and U-Net for live C166 cell segmentation. The cell segmentation results of our proposed cell segmentation method are first compared with the semantic segmentation results using the U-Net output directly. Figures 4(a)–4(c) demonstrate the difference between the segmentation results of the two methods in the four regions with labels. We used true positive, true negative, false positive, and false negative to represent these differences. As indicated by the arrows in the figure, we can observe the presence of the U-Net misidentifying cells as background or misidentifying background as cells. In contrast, our proposed cell segmentation method can perform cell segmentation more accurately. According to Table 1, our method outperforms the traditional U-Net model and Cellpose across[49,50] the board in terms of accuracy, recall, precision, and $F1$ score.[51] The calculation formulas are as follows:

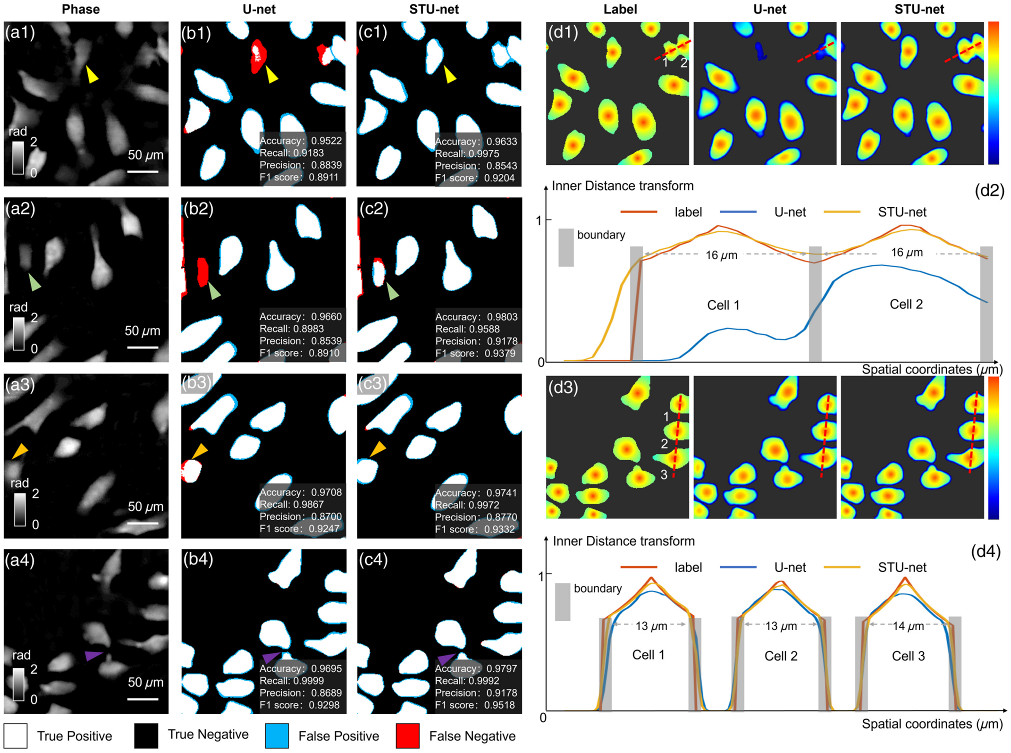


Fig. 4 Comparative experimental results of STU-Net and U-Net. Panels (a)–(c) show the comparison of the results using our cell segmentation method with the traditional U-Net network semantic segmentation within the four regions. Panels (d1) and (d2) are the differences between the inner distance transforms of our STU-Net network and the traditional U-Net outputs within the regions (a1) and (a2). Panels (d3) and (d4) represent the profiles of the two methods with the labels.

Table 1 Comparison of STU-Net, U-Net, and Cellpose segmentation metrics.

	Accuracy	Precision	Recall	$F1$ score
U-Net	0.9646	0.8721	0.9508	0.9096
Cellpose	0.9563	0.8634	0.9213	0.8905
STU-Net	0.9743	0.8895	0.9882	0.9358

$$
\begin{aligned}
\text{accuracy} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}},\\
\text{precision} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\\
\text{recall} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\\
\text{F1 score} &= 2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}},
\end{aligned}
\tag{8}
$$

where true positives (TPs) represent the number of pixels correctly predicted as cells; true negatives (TNs) denote the number of pixels correctly predicted as background; false positives (FPs) refer to the number of pixels incorrectly predicted as cells (which are actually background); and false negatives (FNs) indicate the number of pixels incorrectly predicted as background (which are actually cells).

We then compared the NID output from the STU-Net with those from the U-Net trained on a single task. Figures 4(d1) and 4(d3) show the differences between the NID and the labels for the corresponding regions of Figs. 4(a1) and 4(a4). The results of the NID computed by the U-Net are severely distorted, which results in an inability to further refine the boundaries of the cells.

We performed long-term dynamic live C166 cell segmentation experiments. C166 cells were cultured in 20 mm glass bottom dishes with 10% fetal bovine serum. Our compact system allowed for *in situ* observation by placing it directly in the incubator, shooting one frame every 10 min for a total duration of ~17 h. Figure 5(a) displays the results of cell segmentation for a small area within the full FOV at 00:00:00, comprising ~100 cells.

Figure 5(b) displays the curve of cell count variation in the selected FOV. In addition, we further calculated the multi-dimensional parameters of the cells, such as cell area, perimeter, concavity, perimeter-area ratio, and aspect ratio. Figure 5(c) demonstrates the experimental

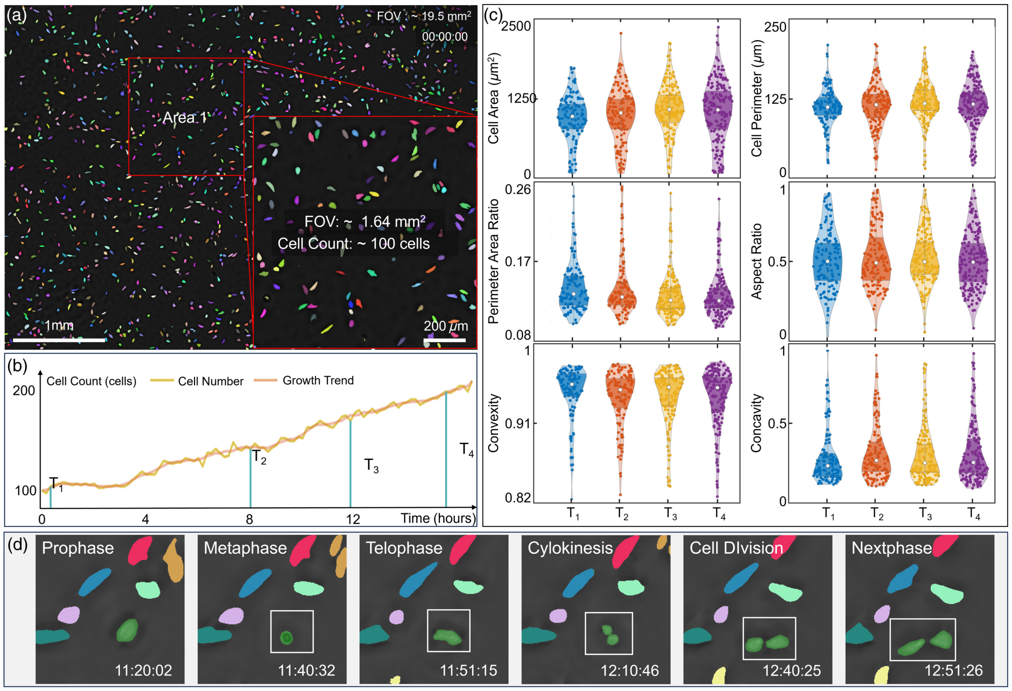


Fig. 5 Long-term dynamic C166 cells segmentation results. A total of 100 frames were taken at 10 min intervals in an area of 1.729 mm^2. The results of the dynamic experiment: Panel (a) shows the segmentation results at the beginning. Panel (b) demonstrates the change in cell number over 16 h. Panel (c) demonstrates that for the cell area, perimeter, concavity, perimeter-area ratio, and aspect ratio of the cells within the area of 1.729 mm^2 at $T_1 = 10$ min, $T_2 = 8$ h, $T_3 = 12$ h, and $T_4 = 16$ h. Panel (d) demonstrates the cell division process within area 1.

results for key frames within the FOV of Fig. 5(a), including $T_1 = 10$ min, $T_2 = 8$ h, $T_3 = 12$ h, and $T_4 = 16$ h. As shown in Fig. 5(d), we can observe the dynamic division process of a cell within Fig. 5(a) area 1. Through the analysis of these indicators, the growth and division of the cells were further investigated. Our method demonstrates the ability to accurately and consistently segment LFOCDHM live cell data with a wide field over an extended period, thereby ensuring reliable support for drug development and biological screening.

4 Conclusions

In this paper, we proposed STU-Net, which achieves multi-tasking through two decoding paths and shares the multi-scale features extracted by Swin Transformer. Based on NID and PLC, the segmentation of cells with complex and variable morphology achieves accurate cell positions and clear cell boundaries (accuracy > 0.9743) across the whole FOV (19.5 mm^2). In addition, multi-dimensional cell parameters can be generated to quantify cell proliferation and growth, thereby improving the accuracy of downstream analysis tasks such as cell tracking and cellular genealogy research. This capability facilitates the analysis of cell morphology, structure, and function, which holds a critical role in the investigation of disease mechanisms, diagnosis, and therapeutic approaches. However, the Swin Transform architecture excels on large datasets due to its ability to capture complex patterns and features as a large model. In future work, we will explore the training of Transformer architectures on larger microscopy datasets with a broader range of cell types.

Code and Data Availability

The data that support the findings of this article are not publicly available due to privacy. They can be requested from the author at liwnenhui@njust.edu.cn.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant No. 2022YFA1205002), the National Natural Science Foundation of China (Grant Nos. 62105151, 62175109, U21B2033, and 62227818), the Leading Technology of Jiangsu Basic Research Plan (Grant No. BK20192003), the Youth Foundation of Jiangsu Province (Grant No. BK20210338), the Biomedical Competition Foundation of Jiangsu Province (Grant No. BE2022847), the Key National Industrial Technology Cooperation Foundation of Jiangsu Province (Grant No. BZ2022039), the Fundamental Research Funds for the Central Universities (Grant Nos. 30920032101 and 30923010206), the Fundamental Scientific Research Business Fee Funds for the Central Universities (Grant No. 2023102001), and the Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense (Grant Nos. JSGP202105 and JSGP202201).

References

1. Y. Park, C. Depeursinge, and G. Popescu, “Quantitative phase imaging in biomedicine,” *Nat. Photonics* **12**(10), 578–589 (2018).
2. G. Popescu, *Quantitative Phase Imaging of Cells and Tissues*, McGraw-Hill Education (2011).
3. Y. Fan et al., “Smart computational light microscopes (SCLMS) of smart computational imaging laboratory (SCILab),” *PhotoniX* **2**, 1–64 (2021).
4. C. Zuo et al., “Transport of intensity equation: a tutorial,” *Opt. Lasers Eng.* **135**, 106187 (2020).
5. L. Lu et al., “Hybrid brightfield and darkfield transport of intensity approach for high-throughput quantitative phase microscopy,” *Adv. Photonics* **4**(5), 056002 (2022).
6. G. Zheng, R. Horstmeyer, and C. Yang, “Wide-field, high-resolution Fourier ptychographic microscopy,” *Nat. Photonics* **7**(9), 739–745 (2013).
7. G. Popescu et al., “Diffraction phase microscopy for quantifying cell structure and dynamics,” *Opt. Lett.* **31**(6), 775–777 (2006).
8. B. Kemper, P. Langehanenberg, and G. Von Bally, “Digital holographic microscopy: a new method for surface analysis and marker-free dynamic life cell imaging,” *Optik Photonik* **2**(2), 41–44 (2007).
9. J. Qian et al., “Structured illumination microscopy based on principal component analysis,” *eLight* **3**(1), 4 (2023).
10. Y. Shu et al., “Adaptive optical quantitative phase imaging based on annular illumination Fourier ptychographic microscopy,” *PhotoniX* **3**(1), 24 (2022).

11. C. Liu et al., “Continuous optical zoom microscope with extended depth of field and 3D reconstruction,” *PhotoniX* **3**(1), 20 (2022).
12. A. W. Lohmann et al., “Space–bandwidth product of optical signals and systems,” *J. Opt. Soc. Am. A* **13**(3), 470–473 (1996).
13. A. Ozcan and U. Demirci, “Ultra wide-field lens-free monitoring of cells on-chip,” *Lab Chip* **8**(1), 98–106 (2008).
14. G. Zheng et al., “The ePetri dish, an on-chip cell imaging platform based on subpixel perspective sweeping microscopy (SPSM),” *Proc. Natl. Acad. Sci.* **108**(41), 16889–16894 (2011).
15. C. Oh et al., “On-chip differential interference contrast microscopy using lensless digital holography,” *Opt. Express* **18**(5), 4717–4726 (2010).
16. J. Zhang et al., “Resolution analysis in a lens-free on-chip digital holographic microscope,” *IEEE Trans. Comput. Imaging* **6**, 697–710 (2020).
17. Y. Chen et al., “Single-shot lensfree on-chip quantitative phase microscopy with partially coherent led illumination,” *Opt. Lett.* **47**(23), 6061–6064 (2022).
18. X. Wu et al., “Wavelength-scanning lensfree on-chip microscopy for wide-field pixel-super-resolved quantitative phase imaging,” *Opt. Lett.* **46**(9), 2023–2026 (2021).
19. X. Lin, X. Li, and X. Lin, “A review on applications of computational methods in drug screening and design,” *Molecules* **25**(6), 1375 (2020).
20. E. R. Polanco et al., “Multiparametric quantitative phase imaging for real-time, single cell, drug screening in breast cancer,” *Commun. Biol.* **5**(1), 794 (2022).
21. K. Yao, N. D. Rochman, and S. X. Sun, “Ctrl—a label-free artificial intelligence method for dynamic measurement of single-cell volume,” *J. Cell Sci.* **133**(7), jcs245050 (2020).
22. D. A. Van Valen et al., “Deep learning automates the quantitative analysis of individual cells in live-cell imaging experiments,” *PLoS Comput. Biol.* **12**(11), e1005177 (2016).
23. M. Sezgin and B. L. Sankur, “Survey over image thresholding techniques and quantitative performance evaluation,” *J. Electron. Imaging* **13**(1), 146–168 (2004).
24. F. Al-Hafiz, S. Al-Megren, and H. Kurdi, “Red blood cell segmentation by thresholding and canny detector,” *Proc. Comput. Sci.* **141**, 327–334 (2018).
25. A. N. Strahler, “Quantitative analysis of watershed geomorphology,” *Eos, Trans. Am. Geophys. Union* **38**(6), 913–920 (1957).
26. J. M. Sharif et al., “Red blood cell segmentation using masking and watershed algorithm: a preliminary study,” in *Int. Conf. Biomed. Eng. (ICoBE)*, IEEE, pp. 258–262 (2012).
27. D. Ciresan et al., “Deep neural networks segment neuronal membranes in electron microscopy images,” in *Adv. in Neural Inf. Process. Syst.*, Vol. 25 (2012).
28. B. Hariharan et al., “Hypercolumns for object segmentation and fine-grained localization,” in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 447–456 (2015).
29. J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 3431–3440 (2015).
30. O. Ronneberger, P. Fischer, and T. Brox, “U-net: convolutional networks for biomedical image segmentation,” *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
31. C. F. Koyuncu et al., “DeepDistance: a multi-task deep regression model for cell detection in inverted microscopy images,” *Med. Image Anal.* **63**, 101720 (2020).
32. R. Caruana, “Multitask learning,” *Mach. Learn.* **28**, 41–75 (1997).
33. Y. Zhang and Q. Yang, “A survey on multi-task learning,” *IEEE Trans. Knowl. Data Eng.* **34**(12), 5586–5609 (2021).
34. Y. Gao et al., “Learning distance transform for boundary detection and deformable segmentation in CT prostate images,” *Lect. Notes Comput. Sci.* **8679**, 93–100 (2014).
35. Y. Rivenson et al., “Phase recovery and holographic image reconstruction using deep learning in neural networks,” *Light: Sci. Appl.* **7**(2), 17141 (2018).
36. J. Zhang et al., “Adaptive pixel-super-resolved lensfree in-line digital holography for wide-field on-chip microscopy,” *Sci. Rep.* **7**(1), 11777 (2017).
37. R. Adams and L. Bischof, “Seeded region growing,” *IEEE Trans. Pattern Anal. Mach. Intell.* **16**(6), 641–647 (1994).
38. T. Pavlidis and Y.-T. Liow, “Integrating region growing and edge detection,” *IEEE Trans. Pattern Anal. Mach. Intell.* **12**(3), 225–233 (1990).
39. K. Alrfou, A. Kordijazi, and T. Zhao, “Computer vision methods for the microstructural analysis of materials: the state-of-the-art and future perspectives,” arXiv:2208.04149 (2022).
40. Z. Liu et al., “Swin transformer: hierarchical vision transformer using shifted windows,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, pp. 10012–10022 (2021).
41. K. Alrfou, T. Zhao, and A. Kordijazi, “Transfer learning for microstructure segmentation with CS-UNet: a hybrid algorithm with transformer and CNN encoders,” arXiv:2308.13917 (2023).

42. M. Raghu et al., “Do vision transformers see like convolutional neural networks?,” in *Adv. in Neural Inf. Process. Syst.*, Vol. 34, pp. 12116–12128 (2021).
43. A. Vaswani et al., “Attention is all you need,” in *Adv. in Neural Inf. Process. Syst.*, Vol. 30 (2017).
44. A. Dosovitskiy et al., “An image is worth 16 × 16 words: transformers for image recognition at scale,” arXiv:2010.11929 (2020).
45. Z. Liu et al., “Video Swin Transformer,” in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 3202–3211 (2022).
46. T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE),” *Geosci. Model Dev. Discuss.* **7**(1), 1525–1534 (2014).
47. P.-T. De Boer et al., “A tutorial on the cross-entropy method,” *Ann. Oper. Res.* **134**, 19–67 (2005).
48. T. Standley et al., “Which tasks should be learned together in multi-task learning?” in *Int. Conf. Mach. Learn.*, PMLR, pp. 9120–9132 (2020).
49. C. Stringer et al., “Cellpose: a generalist algorithm for cellular segmentation,” *Nat. Methods* **18**(1), 100–106 (2021).
50. M. Pachitariu and C. Stringer, “Cellpose 2.0: how to train your own model,” *Nat. Methods* **19**(12), 1634–1641 (2022).
51. M. Sokolova, N. Japkowicz, and S. Szpakowicz, “Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation,” *Lect. Notes Comput. Sci.* **4304**, 1015–1021 (2006).

Wenhui Lin is a master’s student at Nanjing University of Science and Technology and he is pursuing a degree in optoelectronic information engineering under the guidance of Professor Chao Zuo. He received his bachelor of engineering degree in communications engineering from Hohai University in 2022.

Yang Chen is a PhD student at the Nanjing University of Science and Technology and he is pursuing a degree in optical engineering under Dr. Chao Zuo. He received his BS degree in optoelectronic information engineering in electrical engineering from the Nanjing University of Science and Technology in 2020.

Xuejuan Wu is a PhD student at Nanjing University of Science and Technology and she is pursuing a degree in optical engineering under Prof. Chao Zuo. She received her BS degree in physics from Jiangsu Normal University, MS degree in physics from University of Science and Technology of China in 2015.

Yufan Chen is a master’s student in optoelectronic information engineering at Nanjing University of Science and Technology. He received his BE degree in electrical engineering from Wuhan University of Technology in 2021. His main research interest is lens-free quantitative phase microscopy.

Yanyan Gao received her master’s degree in communication engineering from Nanjing University of Science and Technology in 2024 and bachelor’s degree in communication engineering from University of Shanghai for Science and Technology in 2021.

Chao Zuo received his BE and PhD degrees from Nanjing University of Science and Technology (NJUST) in 2009 and 2014, respectively. He was a research assistant at the Centre for Optics and Lasers Engineering, Nanyang Technological University, from 2012 to 2013. He is now a professor at the Department of Electronic and Optical Engineering and principal investigator of the Smart Computational Imaging Laboratory, NJUST. He has broad research interests in computational imaging and high-speed 3D sensing and has authored over 226 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.