Research Article

ADVANCED IMAGING

Snapshot virtual refocusing and three-dimensional reconstruction for microscopy with sequential learning

Xiao Li,[a] Zhenbo Ren,[b,*] Ping Su,[b] Ni Chen,[c] Edmund Y. Lam,[c] Ju Tang,[d,*] Jianglei Di,[b,*] and Jianlin Zhao[a]

[a]Key Laboratory of Light Field Manipulation and Information Acquisition, Ministry of Industry and Information Technology, and Shaanxi Key Laboratory of Optical Information Technology, School of Physical Science and Technology, Northwestern Polytechnical University, Xi'an, China
[b]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[c]Department of Electrical and Electronic Engineering, Faculty of Engineering, The University of Hong Kong, Hong Kong, China
[d]Key Laboratory of Photonic Technology for Integrated Sensing and Communication, Ministry of Education, and Guangdong Provincial Key Laboratory of Information Photonics Technology, Guangdong University of Technology, Guangzhou, China

Abstract. Conventional microscopic systems have limited depth-of-field (DOF), as they can only produce clear images of a three-dimensional (3D) sample located on a specific focal plane. More importantly, once an image is captured, it is difficult to recover images at other distances. In this paper, we propose a sequential learning method for virtual refocusing in microscopy. This approach utilizes a stack of trainable encoder–decoder architectures to purposely learn individual mappings between in-focus and out-of-focus images at different axial positions. By doing so, it enables the prediction of a stack of out-of-focus images in a stepwise manner from a single image, facilitating the implementation of virtual refocusing as well as 3D volume reconstruction. Our approach provides a novel computational way to gradually extend the DOF and to eventually estimate the point spread function of the system and to achieve 3D imaging by yielding high-quality predictions of the image stack. We experimentally validate the performance by training and blindly testing it on various biological samples, proving its superior refocusing capability and generalization. The proposed virtual refocusing approach demonstrates superior performance across various quantitative evaluation metrics and can achieve 3D reconstruction based on predicted out-of-focus images from only one image.

Keywords: computational microscopy; virtual refocusing; snapshot imaging; sequential learning.

Received Oct. 30, 2025; revised manuscript received Jan. 3, 2026; accepted Jan. 20, 2026; published online Mar. 11, 2026.

[DOI: 10.3788/AI.2025.10027]

1. Introduction

In the field of biomedical imaging, low-quality microscope images can significantly impact image analysis and diagnosis[1,2]. To address this issue, confocal microscopy systems utilize point-by-point scanning of samples to achieve high-resolution and tomographic images[3]. Autofocusing, including active and passive techniques, enables rapid observation and image acquisition on a specific focal plane in microscopy systems. However, in reality, biological tissues are rarely uniformly distributed on a single plane. When a portion of the sample lies beyond the focal depth defined by the focusing plane, i.e., depth-of-field (DOF), blurring occurs and prevents the acquisition of high-quality microscopic images[4,5]. These defocused regions complicate automatic biological analysis and observation.

Methods for deblurring microscopic images are primarily categorized into blind deconvolution algorithms [the point spread function (PSF) is unknown] and non-blind image deblurring (the PSF is known). Traditional approaches to image deblurring involve the challenge of estimating the PSF and then deconvolving the blurred image with the estimated PSF to obtain a clear image. Early classical deblurring algorithms include Wiener filtering[6–12], Richardson–Lucy (R–L) deconvolution[13],

*Address all correspondence to Zhenbo Ren, zbren@nwpu.edu.cn; Ju Tang, tangju@gdut.edu.cn; Jianglei Di, jiangleidi@gdut.edu.cn

and total variation methods. With continuous improvement, more convolutional-based image deblurring algorithms have been proposed, such as super-resolution radial fluctuations[14], mean-shift super resolution[15], and deblurring by pixel reassignment[16], all of which have shown promising results.

These methods, however, often fall short of leveraging the prior information within microscopic images, resulting in less precise image restorations mainly suited for non-blind image deblurring scenarios. Given the shortcomings of classical deblurring methods, later traditional approaches have extensively utilized prior information from microscopic images to restore them[17]. This includes the utilization of motion models, multi-image fusion, and other techniques, which have overcome the limitations of the original methods. Advanced methods require additional prior information and complex mathematical models based on conventional approaches[18] to enhance the effectiveness of image deblurring.

In recent years, there has been rapid progress in imaging development based on deep learning. Many emerging deep learning methods have demonstrated remarkable capabilities in various vision tasks, including segmentation, counting, classification, super-resolution, refocusing, and virtual staining[19–30]. Consequently, numerous deep-learning-based approaches have been proposed to address the challenges mentioned. Some notable approaches include the use of generative adversarial networks (GANs) for enhancing surface plasmon resonance microscopy (SPRM). These methods involve training GANs on a diverse set of defocused SPRM images, which enables them to generate focused SPRM images directly from a single defocused image without the need for prior information[31]. One type of deep-learning-based automatic focusing technology for whole slide imaging is the DAFNet, a neural network designed to generate focused images from potentially defocused ones. This network operates with just two images captured at different focal lengths[32]. Additionally, there are other networks, such as the fully connected Fourier convolutional neural network (FCFNN) proposed by Pinkard *et al.*[33], a convolutional neural network (CNN) network introduced by Jiang *et al.*[34], and the GANscan network for continuous scanning microscopy deblurring[35], all of which are capable of producing high-quality images.

However, many of these deep learning methods rely on large manually annotated training data, which can be challenging and laborious to experimentally obtain. The scarcity of labeled data is a significant limitation for deep learning in imaging applications. Researchers have been exploring solutions that combine physical models with data-driven approaches to mitigate this challenge. For example, PhysenNet, as proposed by Situ *et al.*[36], is a method that directly extracts phase information from diffraction images. In subsequent research, they extended this approach to predict the diffraction distance, making it a trainable and predictable parameter[37]. Another notable method is R–L networks (RLNs)[38], which is a fast and lightweight deep learning approach that incorporates the R–L algorithm into the image deconvolution process. It has been applied to three-dimensional (3D) fluorescence microscopy deconvolution and improves network performance by connecting with the image formation process. The application of such methods to specific microscopes includes artificial confocal microscopy, which can achieve confocal levels of depth slicing, sensitivity, and chemical specificity on unlabeled specimens[39]. Furthermore, we previously designed an extended focus imaging algorithm that combines structure tensors and guided filters. This approach utilizes structure tensors for focus measurement and generates initial weight maps, resulting in improved image fusion quality for both incoherent and coherent microscopic systems[40]. However, in more complex scenarios, establishing a physical model for methods like these can be particularly challenging. Moreover, as data-driven approaches continue to improve with advancements in network models, they place higher demands on hardware and computational resources. Therefore, in future research, the integration of mathematical models remains an essential direction.

In this paper, we present an innovative approach utilizing a deep learning technique to achieve virtual refocusing with a single image, thereby synthesizing intentionally induced out-of-focus blur. The proposed method leverages the strong learning capability of neural networks to enhance and expand the microscopic image through the generation of out-of-focus images. In contrast to conventional blur synthesis approaches, our method achieves rapid and precise prediction of out-of-focus images based on experimentally synthesized blur in an end-to-end manner. Besides, by sequentially learning the information among an image stack, the whole 3D volume can be numerically reconstructed without any prior knowledge or regularization constraints.

Within the specific context of this study, sequential learning denotes a stepwise prediction process. The primary objective is to employ the network to generate a series of defocused images with incrementally increasing degrees of blur. This approach enables the model to effectively learn the characteristics of image degradation across the entire defocus range, thereby computationally simulating physical axial displacement. Conceptually, this methodology incorporates the fundamental principles of cascade network architectures.

2. Principle and Method

2.1. Image Formation in Microscopy

As shown in Fig. 1(a), which depicts the optical path diagram of a classic optical microscope and its internal system[41], it can be observed that, when the light source illuminates the sample, the light passes through the microscopic objective lens and the tube lens to form an image on the CCD. When the sample is precisely located at the focal point of the microscopic objective lens, a clear image can be formed. However, when the objective lens is moved, the sample appears blurred in the image plane. If the movement distance exceeds a certain limit, the sample can no longer be clearly rendered. Nonetheless, within a certain limit, the sample can be considered to be in focus, and this limit defines the DOF[42] of the system.

However, when either the object or the detector moves axially by a certain distance, points originating from the object spread out and form a radially symmetric pattern known as the PSF, rather than a single point[43]. The distribution of this pattern results in defocused images on the image plane, causing the loss of object details and lower spatial resolution. Consequently, images captured beyond the focal range become blurred, as illustrated in Fig. 1(b)[44]. This can be explained by the following formula:

$$f = x \otimes h + n, \tag{1}$$

where f represents the observed blurred image, x represents the clear image, $\otimes$ denotes the convolution operator, h represents the PSF, and n is the noise.

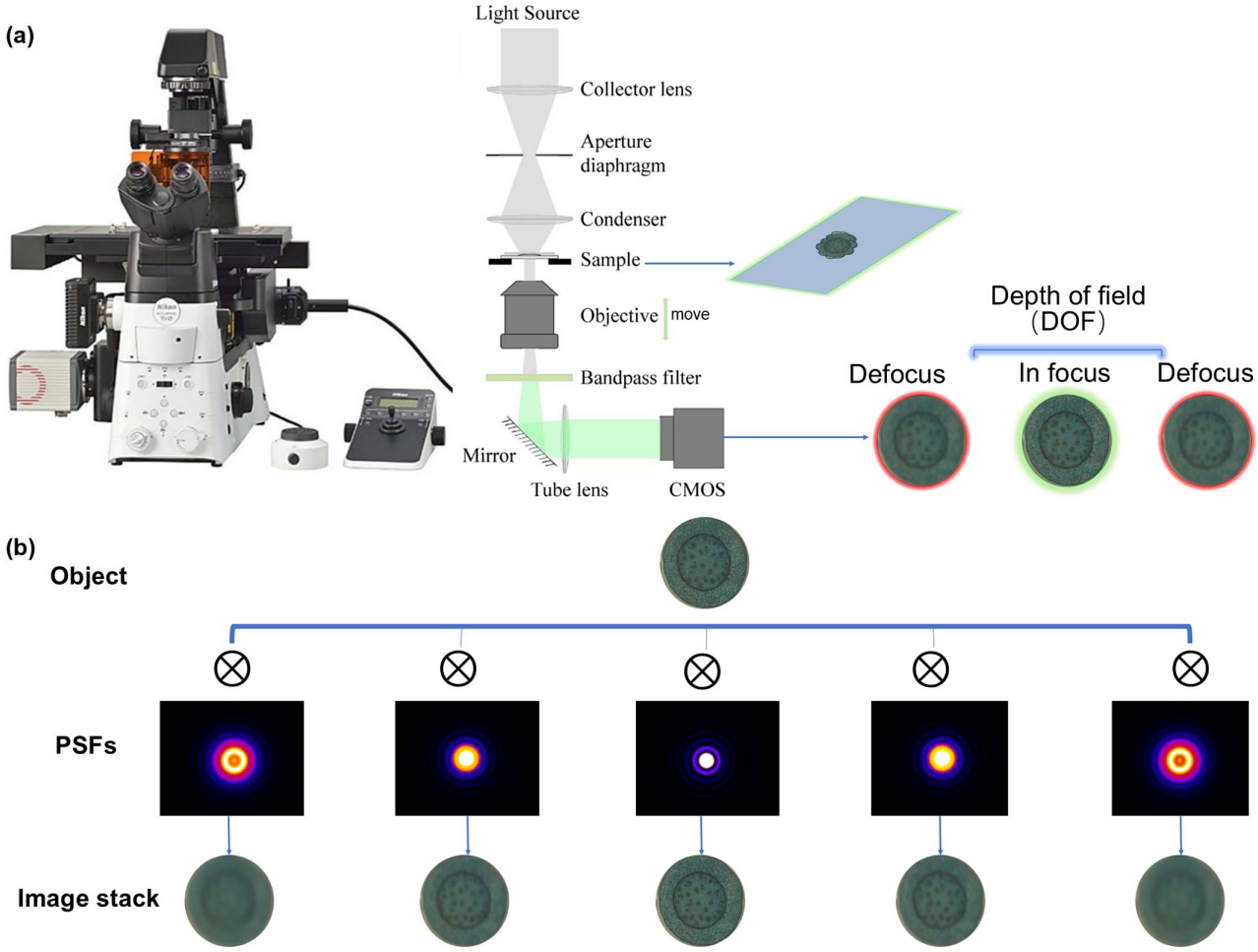

Fig. 1. (a) Image formation of an object with an imaging system. Only objects on the focal plane (or within the DOF) of the lens are in focus, and the image is sharp. (b) Generation of different blurred images through convolution with the object and the blur kernels (i.e., PSFs) at different distances.

According to the principles mentioned above, in microscopy imaging, any sample exceeding the DOF range can result in image blurring. Blurred samples can introduce numerous issues into subsequent 3D imaging, such as a decrease in depth resolution. Defocusing blurs reduce the depth resolution of a microscope, making it challenging to differentiate the positions of objects along the z-axis (the axis perpendicular to the focal plane). This implies that it may not be possible to accurately determine the object's position in 3D space. Furthermore, it leads to the loss of sample details, resulting in insufficient reconstruction accuracy and hindering subsequent analysis. Additionally, it can cause the loss of information from certain depth planes, thereby reducing the ability to present a complete 3D representation of the object. Consequently, whether in 2D or 3D images, defocusing blur represents a critical.

2.2. Proposed Method

The structure of the proposed sequential learning approach[45], as illustrated in Fig. 2(a), can be divided into two main components: encoder and decoder. These two parts together form the U-shaped architecture for which the U-Net is named. Note that the U-Net is not the only structure that can be incorporated in the proposed approach. Any encoder–decoder architecture, such as a variational autoencoder, GAN, or vision Transformer, can also be employed here. The U-Net architecture was selected primarily for its established excellence in biomedical image-to-image translation tasks. Its defining skip-connection mechanism is critical for effectively preserving fine spatial details during the reconstruction process. Furthermore, compared to more complex architectures such as Transformers or GANs, the U-Net possesses a relatively lower parameter count. This computational efficiency makes it particularly well-suited for our proposed sequential prediction task, which requires repeated inference.

Apart from the two components, there is an intermediate feature section called the bottleneck, which includes one or more convolutional layers to further extract advanced features, between the encoder and the decoder. It helps capture more complex information within the image. A key feature of the U-Net is the presence of skip connections, which connect a layer in the encoder to the corresponding layer in the decoder. These connections facilitate the transmission of both low-level and high-level feature information, helping the network capture details and contextual information effectively, thereby enhancing

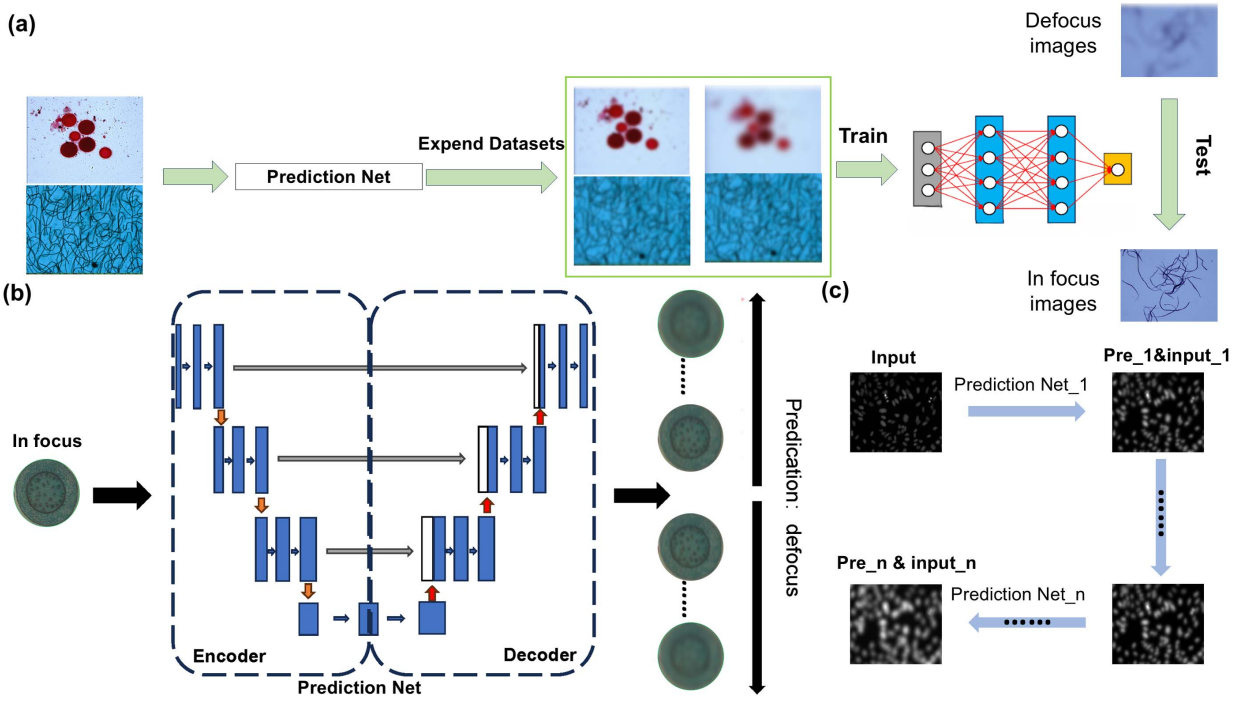

Fig. 2. (a) The acquired in-focus images can be used to generate blurred images through a predictive model for data augmentation. Subsequently, after being trained by the deblurring network, the model can be applied to samples for which it is difficult to obtain large amounts of data. (b) The in-focus images can be utilized to predict blurred images at various defocus distances through a U-shaped network. (c) Iterative prediction and *z*-stack generation.

reconstruction performance. Ultimately, the results are produced as the final output.

The proposed approach verified in this manuscript is based on the well-established U-Net. Through adequate training, it has the capability to learn the mapping relationship between images. A typical deep-learning-based method involves attempting to learn a mapping function R from a large set of labeled data $(\delta_k, I_k), k = 1, \ldots, K$, where δ_k represents the label (in-focus image) and I_k represents the data (out-of-focus image). These labeled data points constitute the training set $S_T = \{(\delta_k, I_k); k = 1, \ldots, K\}$ by solving

$$R_{\theta^*} = \arg\min_{\theta \epsilon \Theta} \|R_\theta(I_k) - \delta_k\|^2 \qquad \forall(\delta_k, I_k) \in S_T, \tag{2}$$

where R_θ represents the mapping function of a neural network defined by a set of weights and biases $\theta \epsilon \Theta$. The training process yields a viable mapping function R_{θ^*} that can map I, which is not in S_T, back to the corresponding $\delta = R_{\theta^*}(I)$.

In typical computational imaging applications, the size k of the training set S_T can range from thousands to even tens of thousands. This allows the network to gradually learn this mapping relationship, enabling end-to-end processing without the need for additional prior knowledge. However, this approach requires a large number of blurred–sharp image pairs for the model to learn from, yet such data are not readily available. If we can leverage the model's learning capability to predict blur using easily obtainable data, these predicted images can be used to augment the dataset for the deblurring network, which in turn can help the model better restore in-focus images. This process is illustrated in Fig. 2(a). Meanwhile, the specific workflow implemented by the blur prediction method is shown in Fig. 2(b): focused images are input into the network, which then progressively generates defocused, blurred images at different positions. We have utilized the fundamental principles and structure of a U-shaped network. The model training workflow is structured as a recursive cascade in Fig. 2(c). Initially, the network accepts an in-focus image (denoted as input) to predict the image at the immediate next defocus position (pre_1). This predicted output then serves as the input (input_1) for the subsequent stage. This process continues iteratively, propagating predictions through successive defocus positions, until the image at the maximum projected defocus distance (pre_n) is reached. In essence, this mechanism generates a comprehensive *z*-axis image stack by repeatedly applying the same network architecture to its own output—a process technically referred to as auto-regressive generation. Consequently, through this chain of inference, a series of sequential prediction models, denoted as pre_net_n, is trained simultaneously.

2.3. Data Preparation and Network Architecture

2.3.1. Data preparation

We utilize the public BBBC006 dataset[46], which is based on human U2OS cells. This dataset comprises images obtained from a 384-well microplate containing U2OS cells stained with Hoechst 33342 (used for labeling cell nuclei). The images were acquired at a 20× magnification and 2× binning, and with exposures of 15 and 1000 ms for Hoechst and Hoechst + Cytoplasm, respectively, at each of the two positions per well.

Laser autofocusing was used to find the optimal focus at the bottom of the well. Subsequently, the automated microscope was programmed to collect z-stacks of 34 image sets, with 2 μm spacing between slices.

Additionally, simulated high content screening (HCS) images are generated using the SIMCEP[46] simulation platform for fluorescence cell population images. Focus blur is simulated by applying a Gaussian filter to the images. Each image has a size of 696 pixel × 520 pixel, in 8-bit TIF format, and the cell nuclei and cell areas match the average cell nucleus and cell area from the BBBC006 (Human U2OS cells) image set. The level of blurriness is indicated by the z-stack levels, ranging from z_1 to z_{16}, representing the degree of blurriness from defocused to focused. The generalization capability of the method is validated by testing these synthetic cell images that are not part of the training set.

The BPAEC dataset[48], an open-source dataset, was acquired using a high-resolution confocal laser scanning fluorescence microscope (Zeiss-LSM880) equipped with a Zyla CMOS camera featuring a pixel size of 6.5 μm. Mitochondria, action, and nuclei in the microscopic sections were stained with MitoTracker® Red CMXRos, Alexa Fluor® 488 phalloidin, and the blue fluorescent DNA stain DAPI, respectively. For each fluorescence excitation channel, image stacks were captured using a 100× magnification objective with a numerical aperture (NA) of 1.4. Each image stack consists of 15 layers, covering a scanning depth of 8.4 μm with an interval of 0.6 μm between layers.

To acquire novel experimental data for further validation, we utilized a Nikon Eclipse Ti2E inverted microscope. Microscopic images of 800 nm diameter microbeads were captured using a 60× objective lens with an NA of 0.7. Given the small scale of the microbeads, a z-axis sampling strategy was adopted: images were acquired at 2 μm intervals, starting from the focal plane and extending to a maximum defocus distance of 20 μm. This process yielded a total of 11 images per FOV. Furthermore, we captured multiple image groups across distinct FOVs. To ensure data integrity, the raw images underwent manual screening and cropping to construct a high-quality dataset, which was subsequently partitioned into training and testing subsets.

Regarding the specific implementation of the dataset, BBBC006, BPAEC, and the microbead dataset were employed for model training. To ensure a rigorous evaluation protocol, each dataset was partitioned according to an 8:1:1 ratio: approximately 80% of the data was allocated for training, 10% for testing, and the remaining 10% for validation. This partitioning strategy is critical for effectively optimizing the model while preventing overfitting.

More specifically, to assess the model's generalization capability, the BBBC005 and *C. elegans* embryo dataset were reserved exclusively for inference and validation. These datasets were evaluated using the model weights pre-trained on the dataset. This implies that the BBBC005 and *C. elegans* embryo data served as completely unseen external test sets and were not involved in the training phase. The experimental 800 nm microbead data underwent an independent training and validation cycle. In contrast, the BPAEC dataset was treated as a distinct group, undergoing a fine-tuning and validation process without cross-validation or data mixing with the other dataset.

2.3.2. Architecture and Training

We conduct data training and testing with the proposed U-Net in Fig. 2. In the U-Net network, during the encoding phase, two convolution operations with 3 × 3 convolution kernels and one 2 × 2 pooling operation are used as a single down-sampling step. This down-sampling is repeated 4 times, compressing the image to a smaller size for feature extraction. Subsequently, during the decoding phase, an up-sampling operation is performed using a single 2 × 2 transposed convolution operation, along with the same convolution operations as in the down-sampling phase. This process is intended to restore the original image size, and a feature fusion operation is conducted during the corresponding up-sampling and down-sampling phases to enhance network reconstruction accuracy. These models are trained for 200 epochs with a batch size of 5. We use the Adam optimizer with a learning rate of 10^{-4} for network optimization. Each image is cropped to a size of 256 × 256, and the dataset is randomly shuffled during preprocessing.

Regarding the optimization strategy, specifically learning rate scheduling and weight decay, we employed a cosine annealing learning rate decay scheme. This choice is motivated by the sequential nature of the task, which involves learning the mapping relationships between blurred images at adjacent defocus positions.

The training process is designed to accommodate the spectral learning characteristics of the network. In the initial phase, the model prioritizes the acquisition of global, low-frequency information (such as general shapes and background intensity). As training progresses, the optimization strategy involves dynamically adjusting the weight distribution within the loss function. This shift compels the model to focus increasingly on reconstructing high-frequency details and fine structures, effectively addressing the "spectral bias" often observed in neural networks.

The mathematical formulations for the optimization schedule and the loss components are defined as follows:

Cosine annealing schedule:

The learning rate η_t at epoch t is updated according to

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{T_{\text{cur}}}{T_{\max}}\pi\right)\right), \tag{3}$$

where $\eta_{\min}$ and $\eta_{\max}$ denote the minimum and maximum learning rates, and T_{cur} and $T_{\max}$ represent the current epoch and the total number of epochs per cycle, respectively.

Loss functions:

To balance pixel-level accuracy, structural fidelity, and spectral consistency, we combined the following loss terms:

MSE loss (mean squared error): used to ensure pixel-wise intensity consistency:

$$L_{\text{mse}} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2, \tag{4}$$

where Y_i is the ground truth and $\widehat{Y}_i$ is the predicted image.

SSIM loss (structural similarity index measure loss): used to maximize structural similarity and perceptual quality:

$$L_{\text{ssim}}(x, y) = 1 - \text{SSIM}(x, y), \tag{5}$$

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{6}$$

Frequency loss: employed to minimize the discrepancy in the frequency domain, ensuring the restoration of high-frequency details:

$$L_{\text{freq}} = \frac{1}{N}\sum_{i=1}^{N} |\text{F}(H(x))_i - \text{F}(y)_i|. \tag{7}$$

The total loss is

$$\text{Loss} = C_1 * L_{\text{mse}} + C_2 * L_{\text{ssim}}(x, y) + C_3 * L_{\text{freq}}, \tag{8}$$

where C_1, C_2, and C_3 are the adjustable weights during the training process.

Compared to vision Transformer (base) architectures (parameters: 86 million, FPS: 50–70), our proposed method is significantly more lightweight, requiring only approximately 3.5 million parameters. The key performance metrics are as follows: total model parameters: ∼34.53; input size 256×256: average FPS: 79.96, inference time per image: 12.51 ms; input size 512×512: average FPS: 28.01, inference time per image: 35.70 ms. Based on the prediction step size defined in this study, these metrics imply that the model is capable of generating the entire z-axis image sequence for a 512×512 input within 2 s. Furthermore, regarding memory efficiency, with a batch size of 8 and spatial dimensions of 512×512, the GPU memory consumption remains under 8 GB.

The specific contributions and experimental findings of this study are summarized as follows: Figs. 1 and 2 primarily illustrate the research background and the proposed methodological framework. The results presented in Figs. 3–7, 10(a), and 11, as well as Table 1, were derived from the model trained on the BBBC006 dataset. Specifically, Figs. 3 and 4 focus on the validation phase, where the efficacy of the proposed method is corroborated through 3D reconstruction of the validation results. Figures 5–7 demonstrate the generalization capability of the model when applied to the BBBC005 dataset. Table 1 provides a quantitative performance comparison of the 3D reconstruction results. Figure 10(a), utilizing the *C. elegans* embryo data, validates the method's performance on biologically complex samples. Figure 11 illustrates the prediction of the system's PSF within the negative defocus range. Additionally, Figs. 8 and 9 present experimental verifications using prepared 800 nm microbead microscopy images, which served as both training and testing data. Finally, Fig. 10(b) displays the results following fine-tuning and validation on the BPAEC dataset, demonstrating the model's robust capacity to learn and predict highly complex structural details.

Table 1. Calculated 3D Volumes Obtained from the z-Stack for Both Hoechst Image and (i) & (ii) Images[a]

Test input image	Data	Volume (pixel)	Percent difference
Hoechst	Predicted image	44,288	0.245%
	G-T image	44,180	
i	Predicted image	62,480	3.711%
	G-T image	60,244	
ii	Predicted image	9670	0.352%
	G-T image	9636	

[a]Volumes are represented in terms of the number of pixels.

3. Results

3.1. Cell Slides with Nuclei Labeled at Different Focal Distances

Figure 3(a) intuitively displays the in-focus image of fluorescently stained Hoechst images selected from the dataset using the proposed method, along with a zoomed-in view of the region of interest (ROI). The intensity values along the dashed line in the ROI image were calculated and plotted as a curve. To compare the differences between the predicted images and the ground-truth (G-T) images, Fig. 3(b) presents portions of the predicted images and the corresponding G-T images at negative defocus, in-focus, and positive defocus positions, showing their respective ROIs. Additionally, the intensity values along the dashed line indicated in Fig. 3(a) were calculated and plotted for comparison, with the red curve representing the intensity values of the G-T image and the green curve representing those of the predicted image.

The line profiles of intensity values [Figs. 3(b) and 3(d)] demonstrate a close match between the predicted and G-T images across all defocus distances, with the curves nearly overlapping. This quantitative alignment, in conjunction with the high SSIM/peak signal-to-noise ratio (PSNR) values reported later, indicates that the predicted images accurately capture both the structural features and intensity distributions of the defocused ground truth. This indicates that the proposed method can effectively establish a mapping relationship from the focused position to defocused positions, enabling the output of images with defocus blur effects without requiring any prior knowledge.

Meanwhile, this dataset includes another more complex phalloidin-stained image type. As shown in Fig. 3(c), this section displays the in-focus fluorescence image of phalloidin staining, where an ROI is similarly selected and enlarged, with intensity values calculated along the dashed line and plotted as a curve. In Fig. 3(d), predicted images and G-T images at different positive and negative defocus distances with the same defocus interval are presented for comparison. The intensity values along the dashed line in the ROI images were calculated and plotted, with blue representing the G-T image intensity curve and pink representing the predicted image intensity curve. The high degree of similarity in the intensity profiles [Fig. 3(d)] is maintained across positive and negative defocus distances. This consistency, supported by subsequent quantitative metrics, further validates the method's effectiveness on this more complex dataset. This further demonstrates the effectiveness of the proposed method on this dataset.

Next, we utilized the "Image Processing and Analysis" plugin in ImageJ to display the 2D slice stacks of blurry images generated by our method. Figures 4(a) and 4(b) present comparisons between the synthetically blurred 3D predicted images and the 3D G-T images along the z-axis stack for Hoechst fluorescence images and phalloidin-stained fluorescence images, respectively. By comparing the features of the two 3D representations, it is evident that they exhibit a high degree of similarity in overall structure and morphology, indicating that the image stacks generated by the network closely resemble real microscopic samples in 3D space. We

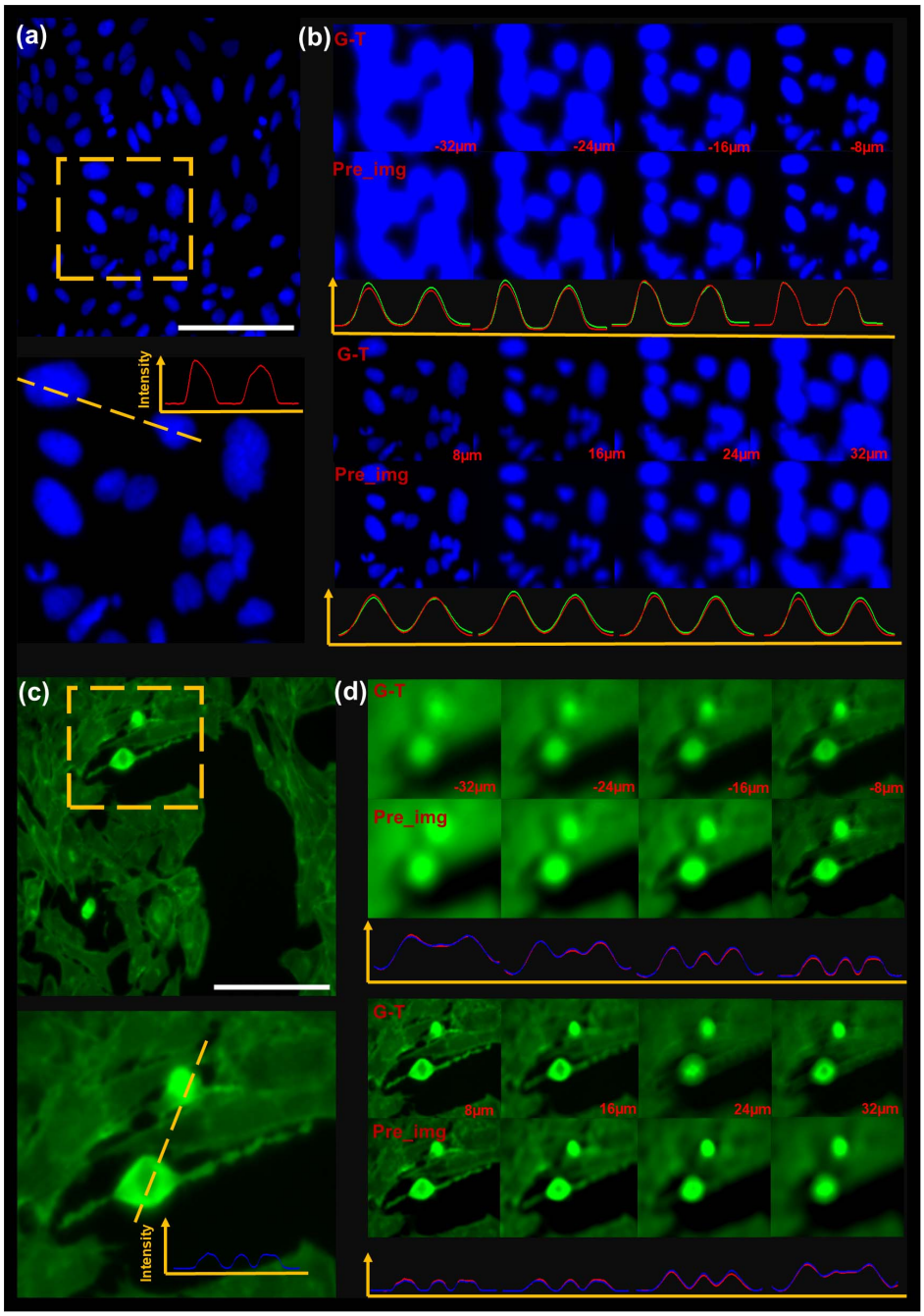

Fig. 3. (a) In-focus fluorescence image of Hoechst staining and the zoomed-in view of the ROI. (b) Comparison of G-T and predicted images in the ROI area under positive and negative defocus, with red curves representing G-T and green curves representing predicted in the intensity profiles. (c) In-focus fluorescence image of phalloidin staining and the zoomed-in view of the ROI. (d) Comparison of G-T and predicted images in the ROI area under positive and negative defocus, with blue curves representing G-T and pink curves representing predicted in the intensity profiles. The scale bars are 300 μm.

specifically extracted and compared the XZ views, calculating intensity values along the yellow dashed line for both G-T and predicted images, and plotted the curves to assess their similarity. The results are shown in Fig. 4(c), where the intensity curves of both fluorescence images are represented by different colors for predicted and G-T. The intensity profiles plotted along the dashed lines in the XZ views [Fig. 4(c)] exhibit nearly identical trends and values for both the predicted and G-T volumes. This close correspondence in the intensity distributions supports the conclusion that the 3D structure generated by our method closely resembles the original.

Furthermore, we performed specific quantitative evaluations by calculating the PSNR and SSIM values for both fluorescence images, as shown in the box plots in Fig. 4(d). The PSNR values

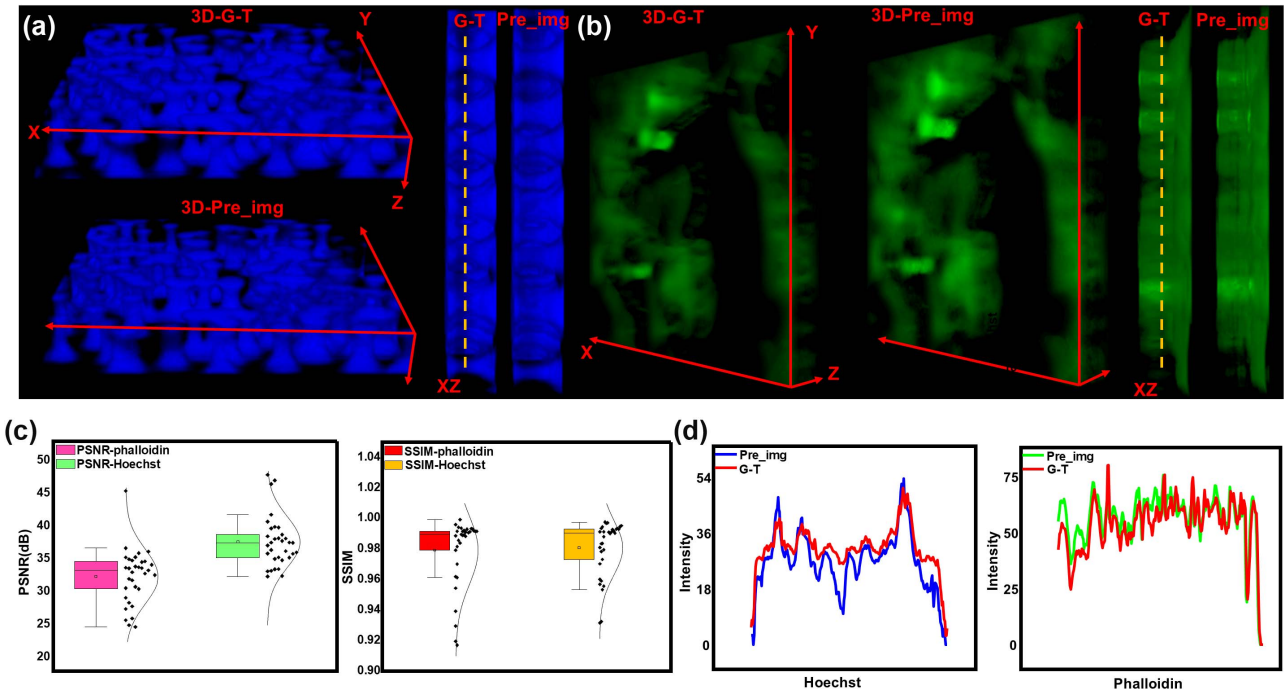

Fig. 4. (a) Synthetically blurred 3D predicted images versus 3D G-T images along the *z*-axis stack for Hoechst fluorescence images. (b) Synthetically blurred 3D predicted images versus 3D G-T images along the *z*-axis stack for phalloidin-stained fluorescence images. (c) Intensity profiles along the dashed lines in the XZ views. (d) Box plots of evaluation metrics for Hoechst and phalloidin-stained fluorescence images.

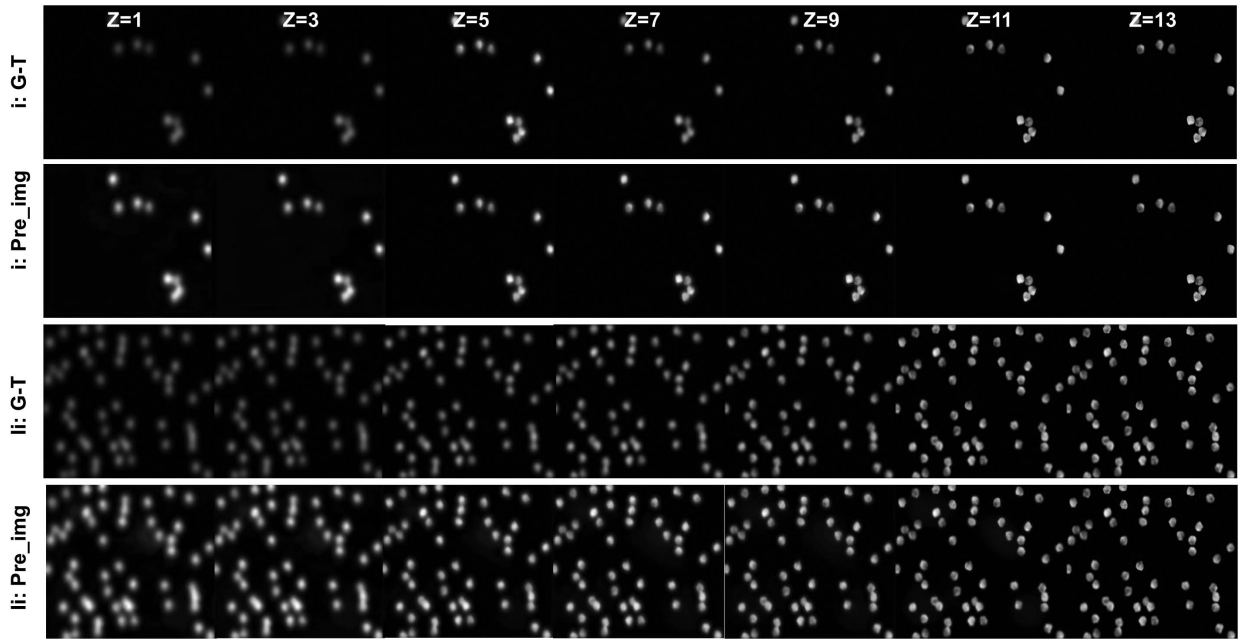

Fig. 5. Generalization test of the proposed method. G-T and predicted images are, respectively, showcased in the first and second rows, displaying images at different defocus positions.

for both image types were all greater than or equal to 25 dB, with Hoechst fluorescence images achieving higher values (above 33 dB), which can be attributed to their simpler features that are easier for the model to learn. For the SSIM metric, both image types achieved excellent values, averaging around 0.98. These numerical results further demonstrate the effectiveness of our method, showing good performance across images with features ranging from simple to complex.

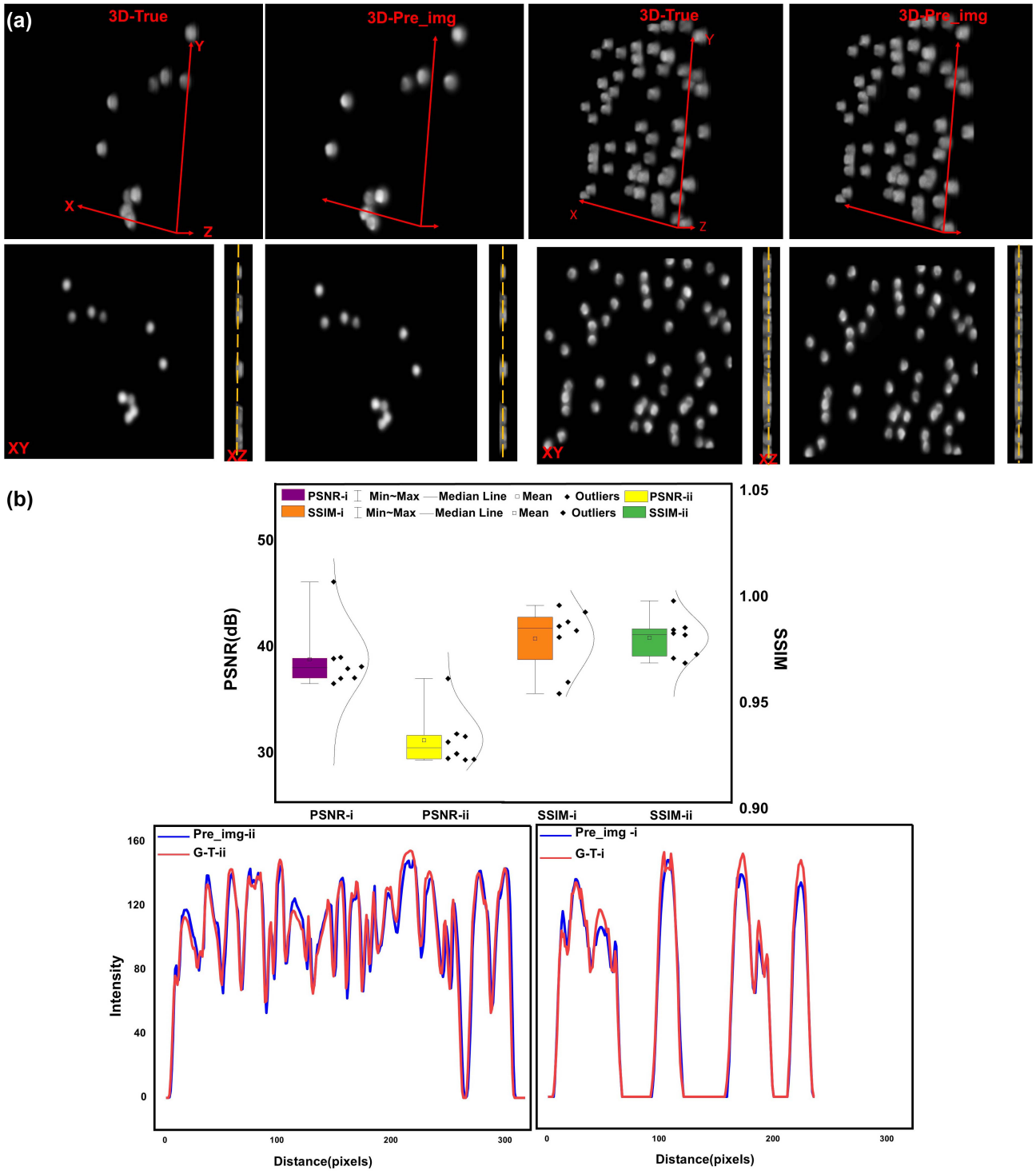

Fig. 6. Volume calculation of image (i) and image (ii). (a) During volume calculation, due to the limited number of features, the entire area is included for calculation. Different colors represent different calculation regions. (b) Selection of ROIs outlined in yellow for calculation.

3.2. Generalization Test and 3D Structure Comparison of Synthetic Data

In the following test, as shown in Fig. 5, we utilize images indicating different levels of blurriness based on the degree of defocus within the z-stack levels ranging from z_1 to z_{13} as the test set. We use the model trained on the dataset, which includes the mapping relationship, to test these synthetic cell images that are not in the training dataset, thereby verifying the generalization capability of our method. Here, we selected two sets of images for testing, designated as (i) and (ii). Figure 5 displays the G-T images at different defocus positions alongside the predicted images generated by our method for both image sets. As shown in

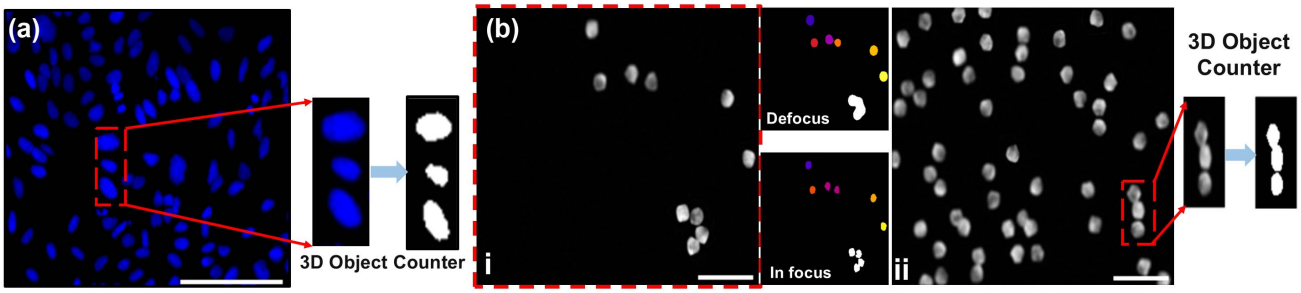

Fig. 7. (a) Voxel count calculation for the 3D Hoechst fluorescence image from the dataset. The red bounding box indicates the ROI used for voxel calculation. (b) Voxel count calculations for samples (i) and (ii) from the BBBC005 dataset. For sample (i), the entire volume was calculated, while for sample (ii), voxel counting was performed within the red-bounded ROI. The scale bars are 300 μm.

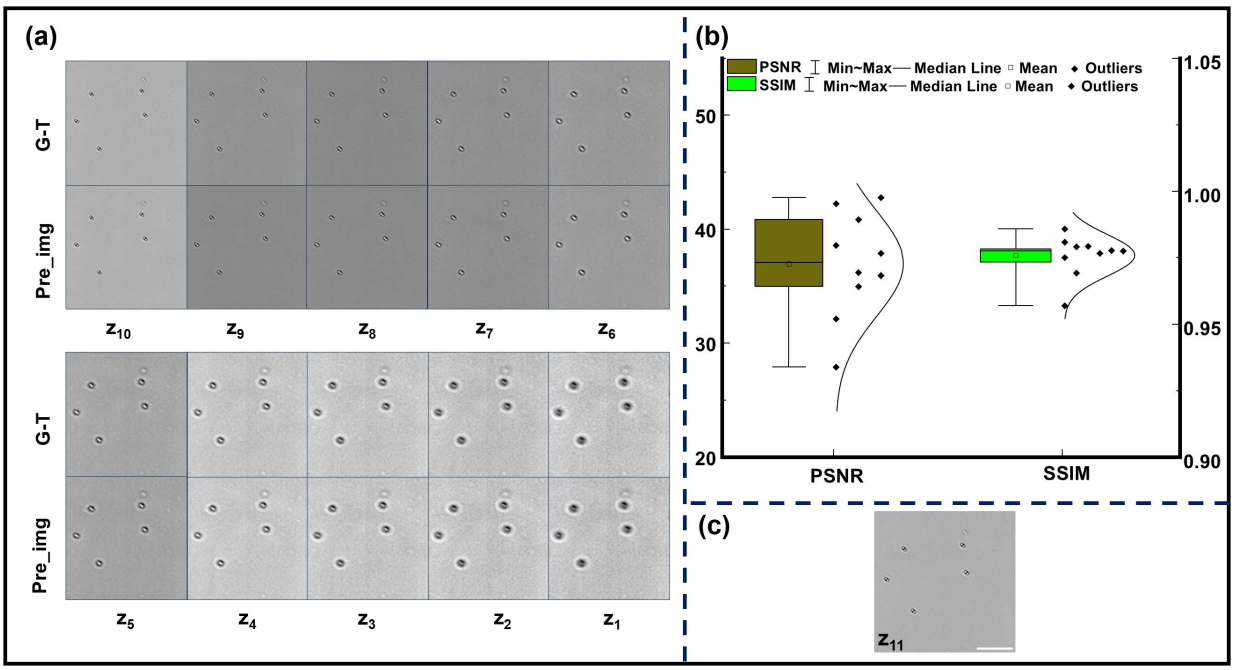

Fig. 8. Validation of the method based on microsphere data. (a) Results of raw images and predicted images from z_{10} to z_1, with the raw images shown above and their corresponding predicted images below. (b) SSIM and PSNR values between the 10 images in (a) are calculated and presented in box plots. (c) In-focus image at position z_{11}. The scale bar is 20 μm.

Fig. 5, both the G-T and predicted image series exhibit a progressive increase in blur extent with increasing defocus distance. This consistent behavior, where the predicted images follow the same defocus progression as the ground truth, validates the model's ability to generalize the blur synthesis process. This observation confirms the validity and demonstrates the generalization capability of our method.

Furthermore, in Fig. 6, we show the 3D representations of the combined z-stack of images for image (i) and image (ii). The figure illustrates the fundamental structures of the 3D images generated from image (i) and Image (ii), with particular emphasis on the XY and XZ cross-sectional views. Comparative analysis between the G-T and predicted images across these views reveals a high degree of similarity. To further evaluate their correspondence, we computed evaluation metrics for the 2D slices of both image (i) and image (ii). Additionally, intensity values along the yellow dashed lines in the XZ cross-sections of the 3D volumetric images were calculated. For clearer visualization, the evaluation metrics are presented as box plots to enable more intuitive comparison of the 2D slice reconstruction quality, while the intensity values are plotted as curves to clearly display the differences between G-T and predicted images. The results are shown in Fig. 6.

Regarding the evaluation metrics, both image (i) and image (ii) achieved PSNR values above 25 dB and SSIM scores exceeding 0.95. These test results align with those obtained from the dataset, where images with simpler features tend to yield higher metrics—for instance, image (i) reached an average

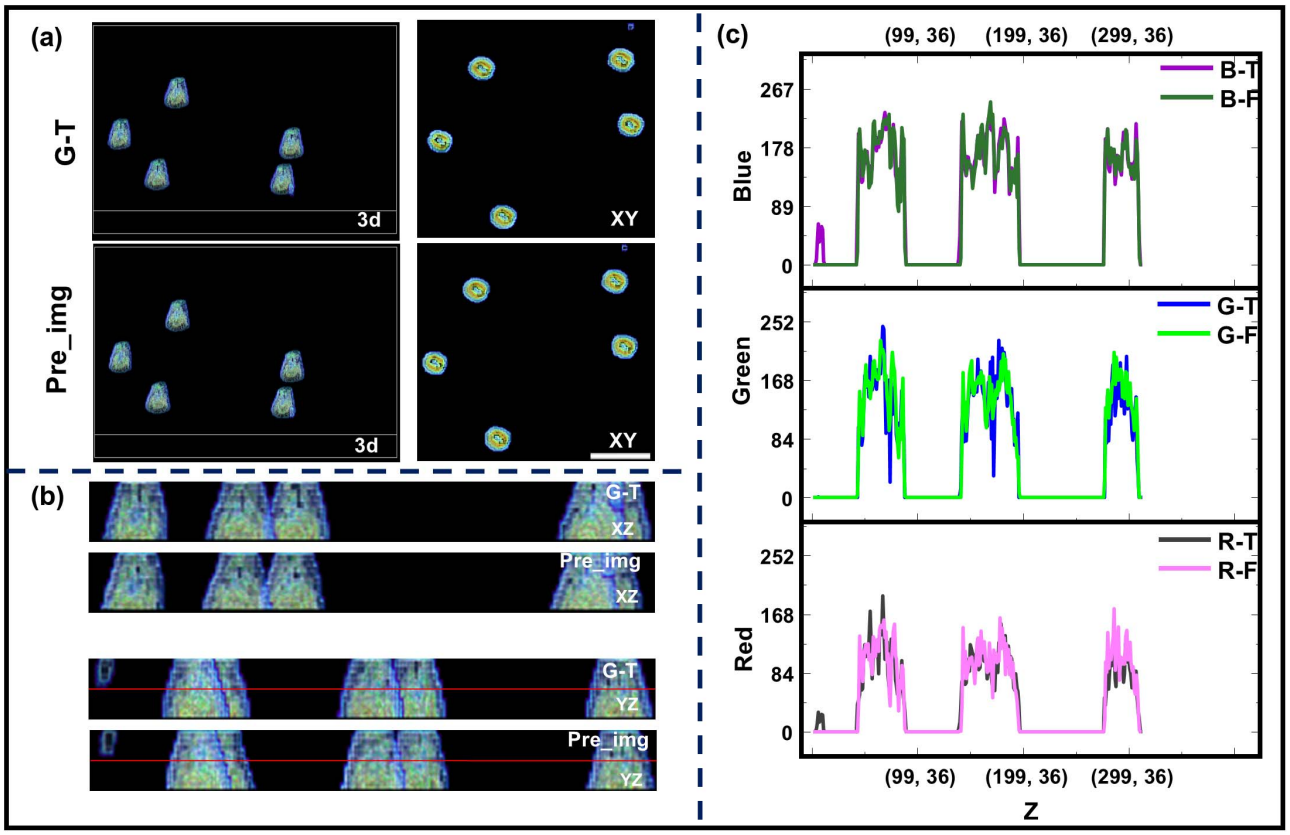

Fig. 9. 3D visualization of the *z*-stack. (a) Comparison between the 3D images of raw and predicted images. Then, the cross-sectional feature images on the XY plane are compared. (b) XZ and YZ cross-sectional images of the 3D image. (c) Intensity profiles along the central axis in the YZ cross-sectional image, delineated across the R–G–B channels. The scale bar is 20 μm.

PSNR of 36 dB in terms of the intensity curves for both image (i) and image (ii), and careful observation and comparison demonstrate only minimal differences between the predicted and G-T images. These visual and quantitative comparisons further confirm the feasibility and robustness of the proposed method.

Furthermore, voxel count calculations were performed on 3D images from the BBBC006 and BBBC005 datasets to further validate the effectiveness of the proposed method. As shown in Fig. 7(a), the ROI of the 3D Hoechst fluorescence image was selected for voxel counting, while Fig. 7(b) presents the voxel count results for the 3D image ROIs of the samples (i) and (ii) from the BBBC005 dataset. Due to its relatively simple features, the entire voxel count of sample (i) was calculated. The final results are summarized in Table 1. Based on the voxel calculation results presented in the table, it can be observed that there is no significant difference between the volumes calculated for the generated images and the G-T images within their respective ROIs, with only a discrepancy of approximately several tens of voxels. This minimal discrepancy relative to the total voxel count demonstrates the method's strong generalization capability. Even when applied to unseen data not included during training, the proposed method still produces satisfactory results.

In summary, the proposed method can predict synthetic blurry images to a certain extent, even on untrained and unseen data, and it can accurately predict moderately blurry images. Due to the architecture of the established network, the approach may face challenges when predicting heavily blurred images with rich feature information.

3.3. Test on Microsphere Data

In this subsection, we capture microscopic images of microspheres with a diameter of 1 μm in the laboratory using a commercial microscope with an objective of 60× for further testing with customized experimental data. Due to the small size of microspheres, images are captured with an interval of 0.5 μm, ranging from the in-focus position to 5 μm away, resulting in a total number of 11 images for testing. Figure 8(a) compares the predicted and G-T images across multiple focal positions. The high fidelity of the predictions is quantified by the SSIM and PSNR values presented in Fig. 8(b), which are above 0.95 and 28 dB, respectively, indicating excellent pixel-wise and structural agreement. We then compute the SSIM and PSNR values between the predicted image and the raw image. As shown by the line in Fig. 8(b), SSIM values are all above 0.95, while the black line represents PSNR values, which are above 28 dB and can reach as high as 42 dB. This indicates a high degree of similarity between the predicted image and the raw image. Figure 8(c) shows the in-focus image at position z_{11} for out-of-focus image prediction. With this in-focus image, the predicted image stack in Fig. 8(a) is numerically generated.

Based on the above results, we perform a 3D visualization of the z-stack. Due to significant background noise in the captured

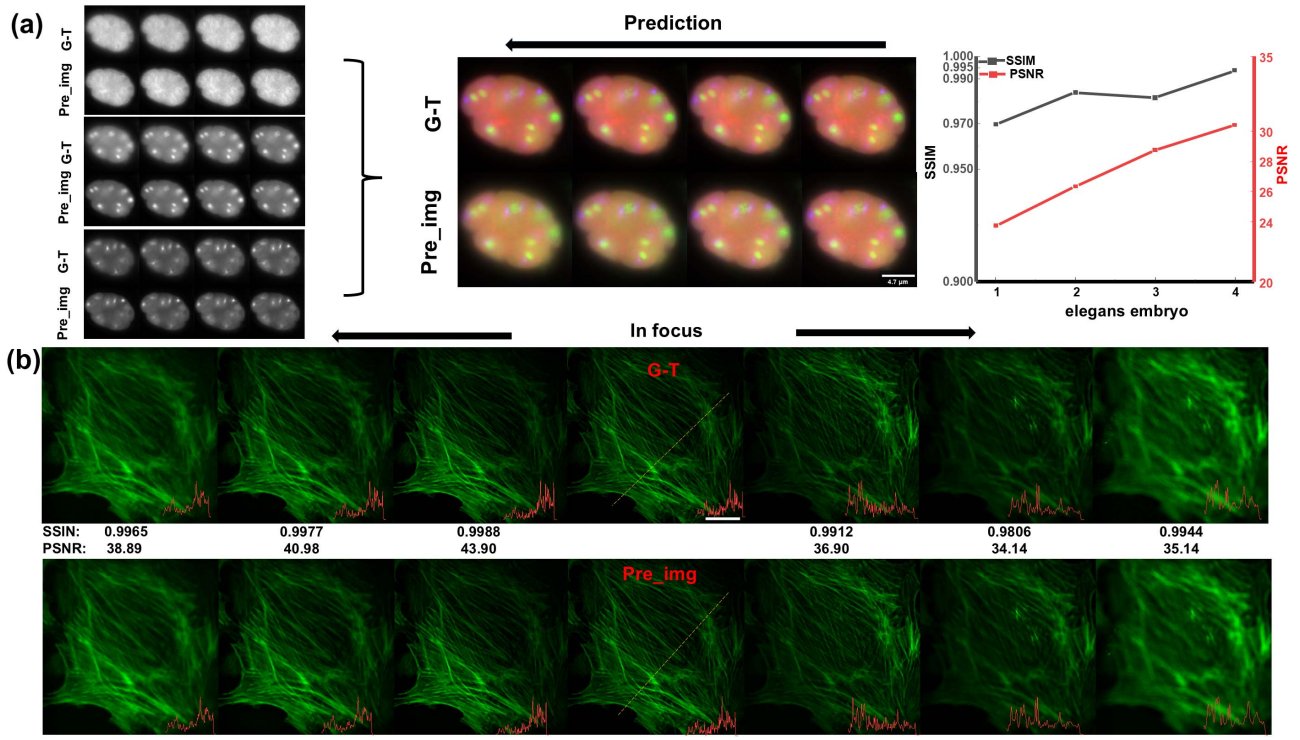

Fig. 10. (a) G-T and predicted images for each of the three channels as well as the synthesized color image, accompanied by the computation of SSIM and PSNR (dB) values for the synthesized color image. (b) Using the focused position images as inputs, the network progressively predicts images at both positive and negative defocus positions, followed by the evaluation of assessment metrics. (a) Scale bar, 4.7 μm; (b) scale bar, 8 μm.

images, it is challenging to achieve a satisfactory 3D display. To address this, we apply a binary process to the z-stack images, separating the background from the features. The 3D viewer results are presented in Fig. 9. First, we showcase the 3D image generated from the z-stack images in Fig. 9(a). Subsequently, we display the distribution of features on the XY plane. The 3D renderings and the XY cross-sectional features of both volumes are presented in Fig. 9(a). A direct comparison shows comparable spatial distributions and shapes of the microsphere clusters in the predicted and G-T volumes. In Fig. 9(b), the XZ and YZ cross-sectional images of the 3D image are displayed. We proceed to compute the intensity distribution curve at a specific location within the YZ cross-sectional image, delineating the calculated range with a red line. As depicted in Fig. 9(c), the intensity distribution curves of the red, green, and blue (R, G, and B) channels are individually presented at this location, and by comparing the intensity curves between the raw and predicted images, it can be observed that they closely match. This demonstrates a high degree of similarity between the two in the 3D display.

3.4. Test on C. Elegans Embryo and BPAEC

To further validate the effectiveness of our method on more complex image structures, we conducted training and testing using the *C. elegans* embryo[47] and BPAEC dataset[48]. For the *C. elegans* embryo, we directly performed testing without including it in the training set. This dataset comprises three channels: the CY3 channel, the DAPI channel, and the FITC channel. Initially, we input individually focused images from each channel and subsequently processed them through the network to obtain predicted defocused images for the three channels. Finally, merging the results from the three channels yielded the combined image as illustrated in Fig. 10(a). The high SSIM and PSNR values plotted adjacent to the images quantitatively demonstrate the high similarity between the G-T and predicted multi-channel composites.

As for the action images, our results, displayed in Fig. 10(b), involve inputting focused images into the network and gradually generating predicted images for both positive and negative defocus. Through SSIM and PSNR metric calculations between the G-T and predicted images, the SSIM hovers around 0.99, while the PSNR consistently remains above 34 dB. In summary, based on the aforementioned data, our method demonstrates good performance even on complex structures.

3.5. PSF Estimation of the Microscopic System

In this section, we utilized an image displaying a 2×2 square-shaped feature at the center of an image with a size of 256×256, as illustrated in Fig. 11(a), as the input to the trained network. Following the network's processing, various degrees of blurriness were produced in the output images, accompanied by intensity curves of PSF images ranging from -4 to -32 μm. Subsequently, a composite 3D image stack was created. Observing the resultant images' post-network processing, the progression from focused to defocused states is evident. The alteration in the blurriness of the square point image is based

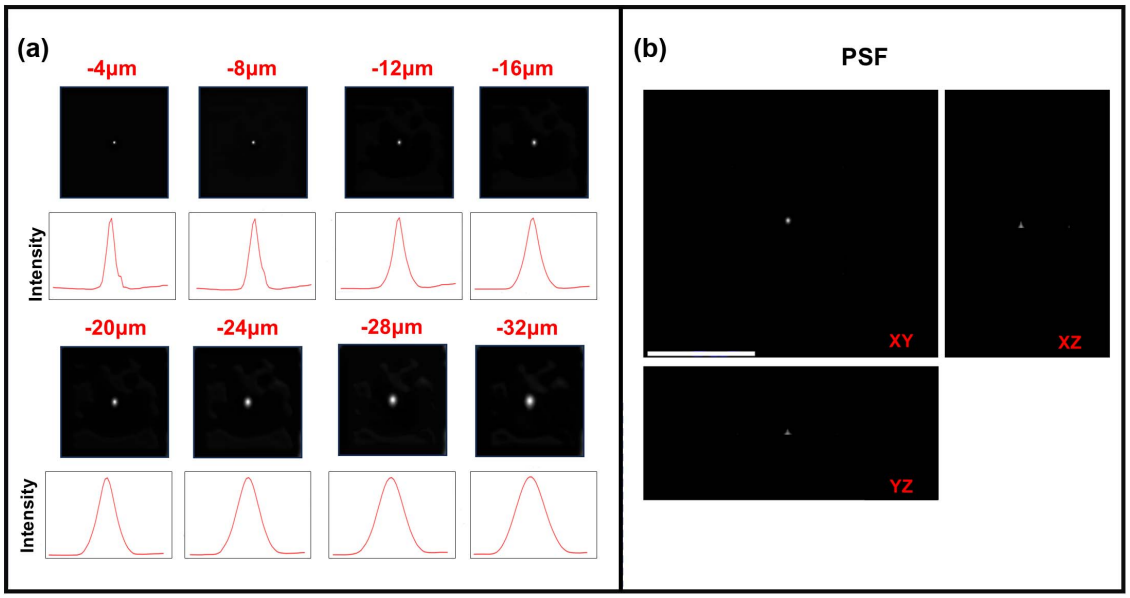

Fig. 11. (a) Focused images alongside the defocused images generated by the network output. (b) 3D model of the synthesized PSF using the image stack and the respective slices in the XY, XZ, and YZ directions. The scale bar is 64 μm.

on the estimated PSF using this methodology. The 3D composite image stack is employed to demonstrate XY, XZ, and YZ cross-sections, as depicted in Fig. 11(b).

4. Conclusion

In this paper, we propose a sequential learning method that utilizes deep learning to synthesize out-of-focus images at different axial positions for virtual refocusing. The classical encoder–decoder network is cascaded to gradually predict out-of-focus images in a stepwise way. Based on experimental results, we can conclude that this method is feasible. The approach harnesses the powerful learning capability of the neural network to learn the mapping knowledge between the individual in-focus and out-of-focus positions. When a focused image is fed into the network, it generates an adjacent out-of-focus image away from the focal plane. Additionally, the proposed virtual refocusing method requires only a single input image, yet it can reconstruct 3D volume by leveraging the predicted out-of-focus image stack. Tremendous results show that the generated predicted image exhibits high SSIM and PSNR compared to the G-T image. We also conduct a generalization test on an untrained synthetic cell dataset. To some extent, we achieve satisfactory results similar to the original training dataset, indicating the method's stability and generalization capability.

In summary, our method holds significant promise for virtual refocusing in microscopy, potentially mitigating challenges associated with acquiring tremendous data in some complex scenarios. With only one image, supervised training data augmentation and 3D volume can be reconstructed. Future research may explore the extension of this method by combining physics- and data-driven models, enabling more precise refocusing and further enhancement of deblurring capability in real time. Besides, the current model assumes the use of thin samples. In the context of thicker specimens, axial crosstalk may occur. To address such scenarios, future iterations of the model could integrate three-dimensional PSF modeling.

Author Contributions

Xiao Li: Conceptualization, Methodology, Software, Data curation, Visualization, Writing–original draft. Zhenbo Ren: Conceptualization, Methodology, Writing–review & editing, Funding acquisition. Ping Su: Writing–review & editing. Edmund Lam: Writing–review & editing. Ju Tang: Methodology, Visualization, Writing–review & editing. Jianglei Di: Supervision, Writing–review & editing, Funding acquisition. Jianlin Zhao: Supervision, Writing–review & editing, Funding acquisition.

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.

Acknowledgments

This work was supported by the National Key R&D Program of China (No. 2021YFB2900900), the National Natural Science Foundation of China (No. 62275218), the China Postdoctoral Science Foundation (No. 2022M712586), and the Guangdong Introducing Innovative and Entrepreneurial Teams of "The Pearl River Talent Recruitment Program" (No. 2021ZT09X04).

References

1. S. J. Yang *et al.*, "Assessing microscope image focus quality with deep learning," *BMC Bioinform.* **19**, 77 (2018).
2. Y. Zhang *et al.*, "Large depth-of-field ultra-compact microscope by progressive optimization and deep learning," *Nat. Commun.* **14**, 4118 (2023).

3. N. Kumar, R. Gupta, and S. Gupta, "Whole slide imaging (WSI) in pathology: current perspectives and future directions," *J. Digit. Imaging.* **33**, 1034 (2020).
4. J. Tang *et al.*, "Single-shot image restoration via a model-enhanced network with unpaired supervision in an optical sparse aperture system," *Opt. Lett.* **48**, 4849 (2023).
5. Y. Wu *et al.*, "Three-dimensional virtual refocusing of fluorescence microscopy images using deep learning," *Nat. Methods* **16**, 1323 (2019).
6. S. Liu and H. Hu, "Extended depth-of-field microscopic imaging with a variable focus microscope objective," *Opt. Express* **19**, 353 (2010).
7. W. Lin *et al.*, "Multi-focus microscope with HiLo algorithm for fast 3-D fluorescent imaging," *PLoS One* **14**, e0222729 (2019).
8. J. R. Alonso *et al.*, "Computational multifocus fluorescence microscopy for three-dimensional visualization of multicellular tumor spheroids," *J. Biomed. Opt.* **27**, 066501 (2022).
9. Y. Rivenson *et al.*, "Deep learning microscopy," *Optica* **4**, 1437 (2017).
10. B. Seong *et al.*, "E2E-BPF microscope: extended depth-of-field microscopy using learning-based implementation of binary phase filter and image deconvolution.," *Light Sci. Appl.* **12**, 269 (2023).
11. L. Jin *et al.*, "Deep learning extended depth-of-field microscope for fast and slide-free histology," *Proc. Natl. Acad. Sci. U.S.A.* **117**, 33051 (2020).
12. V. Pronina, *et al.*, "Microscopy image restoration with deep Wiener-Kolmogorov filters," in *European Conference on Computer Vision (ECCV)* (2020), p. 185.
13. H. Zhao *et al.*, "A new deep learning method for image deblurring in optical microscopic systems," *J. Biomed. Opt.* **13**, e201960147 (2020).
14. N. Gustafsson *et al.*, "Fast live-cell conventional fluorophore nanoscopy with ImageJ through super-resolution radial fluctuations," *Nat. Commun.* **7**, 12471 (2016).
15. E. Torres-García *et al.*, "Extending resolution within a single imaging frame," *Nat. Commun.* **13**, 7452 (2022).
16. B. Zhao and J. Mertz, "Resolution enhancement with deblurring by pixel reassignment," *Adv. Photonics* **5**, 066004 (2023).
17. Z. Wang *et al.*, "Real-time volumetric reconstruction of biological dynamics with light-field microscopy and deep learning," *Nat. Methods* **18**, 551 (2021).
18. S. Koho *et al.*, "Fourier ring correlation simplifies image restoration in fluorescence microscopy," *Nat. Commun.* **10**, 3103 (2019).
19. E. Park *et al.*, "Unsupervised inter-domain transformation for virtually stained high-resolution mid-infrared photoacoustic microscopy using explainable deep learning," *Nat. Commun.* **15**, 10892 (2024).
20. C. Ma *et al.*, "Pretraining a foundation model for generalizable fluorescence microscopy-based image restoration," *Nat. Methods* **21**, 1558 (2024).
21. M. Guo *et al.*, "Deep learning-based aberration compensation improves contrast and resolution in fluorescence microscopy," *Nat. Commun.* **16**, 313 (2025).
22. L. Chen *et al.*, "A physics-informed blur learning framework for imaging systems," in *Proceedings of the Computer Vision and Pattern Recognition Conference* (2025), p. 10913.
23. R. Cao *et al.*, "Dark-based optical sectioning assists background removal in fluorescence microscopy," *Nat. Methods* **22**, 1299 (2025).
24. Z. Lu *et al.*, "Physics-driven self-supervised learning for fast high-resolution robust 3D reconstruction of light-field microscopy," *Nat. Methods* **22**, 1545 (2025).
25. J. Han *et al.*, "System- and sample-agnostic isotropic three-dimensional microscopy by weakly physics-informed, domain-shift-resistant axial deblurring," *Nat. Commun.* **16**, 745 (2025).
26. X. Yang *et al.*, "Deep-learning-based virtual refocusing of images using an engineered point-spread function," *ACS Photonics* **8**, 2174 (2021).
27. B. Bai *et al.*, "Deep learning-enabled virtual histological staining of biological samples," *Light Sci. Appl.* **12**, 57 (2023).
28. J. Winnik *et al.*, "Versatile optimization-based speed-up method for autofocusing in digital holographic microscopy," *Opt. Express* **29**, 33297 (2021).
29. Z. Yuan *et al.*, "Digital refocusing based on deep learning in optical coherence tomography," *Biomed. Opt. Express* **13**, 3005 (2022).
30. J. Tang *et al.*, "Phase aberration compensation via a self-supervised sparse constraint network in digital holographic microscopy," *Opt. Lasers Eng.* **168**, 107671 (2023).
31. Y. Xu *et al.*, "A single-shot autofocus approach for surface plasmon resonance microscopy," *ACS Photonics* **93**, 2433 (2021).
32. Q. Li *et al.*, "Rapid whole slide imaging via dual-shot deep autofocusing," *IEEE Trans. Comput. Imagine* **7**, 124 (2021).
33. H. Pinkard *et al.*, "Deep learning for single-shot autofocus microscopy," *Optica* **6**, 794 (2019).
34. S. Jiang *et al.*, "Transform- and multi-domain deep learning for single-frame rapid autofocusing in whole slide imaging," *Biomed. Opt. Express* **9**, 1601 (2018).
35. M. J. Fanous and G. Popescu, "GANscan: continuous scanning microscopy using deep learning deblurring," *Light Sci. Appl.* **11**, 265 (2022).
36. F. Wang *et al.*, "Phase imaging with an untrained neural network," *Light Sci. Appl.* **9**, 77 (2020).
37. X. Zhang, F. Wang, and G. Situ, "BlindNet: an untrained learning approach toward computational imaging with model uncertainty," *J. Phys. D: Appl. Phys.* **55**, 034001 (2021).
38. Y. Li *et al.*, "Incorporating the image formation process into deep learning improves network performance," *Nat. Methods* **19**, 1427 (2022).
39. X. Chen *et al.*, "Artificial confocal microscopy for deep label-free imaging," *Nat. Photonics* **17**, 250 (2023).
40. Z. Ren *et al.*, "Extended focused imaging in microscopy using structure tensor and guided filtering," *Opt. Lasers Eng.* **140**, 106549 (2021).
41. Z. Ren, E. Y. Lam, and J. Zhao, "Acceleration of autofocusing with improved edge extraction using structure tensor and Schatten norm," *Opt. Express* **28**, 14712 (2020).
42. J. Liao *et al.*, "Rapid focus map surveying for whole slide imaging with continuous sample motion," *Opt. Lett.* **42**, 3379 (2017).
43. L. Silvestri *et al.*, "Universal autofocus for quantitative volumetric microscopy of whole mouse brains," *Nat. Methods* **18**, 953 (2021).
44. Z. Zhang, R. K. Y. Chan, and K. K. Y. Wong, "Spiral phase modulation based deep learning for autofocusing," in *Conference on Lasers and Electro-Optics, Technical Digest Series* (Optica Publishing Group, 2022), p. JW3B.7.
45. 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 (MICCAI)* (2015), p. 234.
46. V. Ljosa, K. L. Sokolnicki, and A. E. Carpenter, "Annotated high-throughput microscopy image sets for validation," *Nat. Methods* **9**, 637 (2012).
47. A. Griffa, N. Garin, and D. Sage, "Comparison of deconvolution software in 3D microscopy: a user point of view—part 1," *G.I.T. Imaging Microsc.* **12**, 43 (2010).
48. C. Zhang *et al.*, "Correction of out-of-focus microscopic images by deep learning," *Comput. Struct. Biotechnol. J.* **20**, 1957 (2022).