PHOTONICS Research

Learning the imaging mechanism directly from optical microscopy observations

Ze-Hao Wang,[1,2,†] Long-Kun Shan,[1,2,†] Tong-Tian Weng,[1,2] Tian-Long Chen,[3] Xiang-Dong Chen,[1,2,4] Zhang-Yang Wang,[3] Guang-Can Guo,[1,2,4] and Fang-Wen Sun[1,2,4,*]

[1]CAS Key Laboratory of Quantum Information, University of Science and Technology of China, Hefei 230026, China
[2]CAS Center for Excellence in Quantum Information and Quantum Physics, University of Science and Technology of China, Hefei 230026, China
[3]University of Texas at Austin, Austin, Texas 78705, USA
[4]Hefei National Laboratory, University of Science and Technology of China, Hefei 230088, China
[†]These authors contributed equally to this work.
*Corresponding author: fwsun@ustc.edu.cn

Received 22 February 2023; revised 7 September 2023; accepted 9 October 2023; posted 16 October 2023 (Doc. ID 488310); published 8 December 2023

The optical microscopy image plays an important role in scientific research through the direct visualization of the nanoworld, where the imaging mechanism is described as the convolution of the point spread function (PSF) and emitters. Based on *a priori* knowledge of the PSF or equivalent PSF, it is possible to achieve more precise exploration of the nanoworld. However, it is an outstanding challenge to directly extract the PSF from microscopy images. Here, with the help of self-supervised learning, we propose a physics-informed masked autoencoder (PiMAE) that enables a learnable estimation of the PSF and emitters directly from the raw microscopy images. We demonstrate our method in synthetic data and real-world experiments with significant accuracy and noise robustness. PiMAE outperforms DeepSTORM and the Richardson–Lucy algorithm in synthetic data tasks with an average improvement of 19.6% and 50.7% (35 tasks), respectively, as measured by the normalized root mean square error (NRMSE) metric. This is achieved without prior knowledge of the PSF, in contrast to the supervised approach used by DeepSTORM and the known PSF assumption in the Richardson–Lucy algorithm. Our method, PiMAE, provides a feasible scheme for achieving the hidden imaging mechanism in optical microscopy and has the potential to learn hidden mechanisms in many more systems.

https://doi.org/10.1364/PRJ.488310

1. INTRODUCTION

Optical microscopy is of great importance in scientific research to observe the nanoworld. The common view is that the Abbe diffraction limit describes the lower bound of the spot size and thus limits the microscopic resolution. However, recent studies have demonstrated that by designing and measuring the point spread function (PSF) or equivalent PSF of microscopy, it is possible to achieve subdiffraction limit localization of emitters. Techniques such as photoactivated localization microscopy [1] and stochastic optical reconstruction microscopy [2] attain superresolution molecular localization through selective excitation and reconstruction algorithms that are based on the microscopy PSF. The spatial mode sorting-based microscopic imaging method (SPADE) [3] can be treated as a deconvolution problem using higher-order modes as the equivalent PSF. Stimulated-emission depletion microscopy achieves superresolution imaging by introducing illumination with donut-shaped PSFs to selectively deactivate fluorophores [4,5]. Additionally, deep-learning-based methods, such as DeepSTORM [6] and DECODE [7], use deep neural networks (DNNs) to predict emitters in raw images by synthesizing training sets with the same PSFs as those used in actual experiments. In all of these microscopic imaging techniques, prior knowledge of the PSF is crucial, making it of great interest to develop a method for directly estimating the PSF from raw images.

Currently, some traditional algorithms such as Deconvblind [8] use maximum likelihood estimation to infer the PSF and emitters from raw images [9–18]. However, these algorithms face two challenges. First, they struggle to estimate PSFs with complex shapes. Second, they can lead to trivial solutions where the PSF is a δ function and the image of the emitters is equal to the raw image. To tackle these issues, researchers have turned to using DNNs [19]. However, this requires a library of PSFs and a large number of sharp microscope images to generate the training data set, which limits the application of these algorithms.

We use self-supervised learning to overcome the above challenges. Here, we treat the PSF as the pattern hidden in the raw images and the emitters as the sparse representation of the raw image. As a result, we propose a physics-informed masked

autoencoder (PiMAE, Fig. 1) that estimates the PSF and emitters directly from the microscopy raw images. Using raw data synthesized by various simulated PSFs, we compare the results of PiMAE and Deconvblind [8] for estimating PSF, as well as PiAME, the Richardson–Lucy algorithm [20], and DeepSTORM [6] for localizing emitters. Our proposed self-supervised learning approach, PiMAE, outperforms existing algorithms without the need for data annotation or PSF measurement. PiMAE demonstrates a significant performance improvement, as measured by the normalized root mean square error (NRMSE) metric, and is highly resistant to noise. In tests with real-world experiments, PiMAE resolves wide-field microscopy images of standard PSF, out-of-focus PSF, and aberrated PSF with high quality, and the results achieve a resolution comparable to structured illumination microscopy (SIM) results. Also, we demonstrate that five raw images can satisfy the requirements of self-supervised training. This approach, PiMAE, shows wide applicability in synthetic data testing and real-world experiments. We expect its usage for the estimation of hidden mechanisms in various physical systems.

2. METHOD

Self-supervised learning leverages the inherent structure or patterns in data to learn meaningful representations. There are two main categories: contrastive learning [21–24] and pretext task learning [25–29]. Mask image modeling (MIM) [25,30–33] is a pretext task-learning technique that randomly masks portions of an input image. Recently, MIM has been shown to learn transferable, robust, and generalized representations from visual images, improving performance in downstream computer vision tasks [34]. PiMAE is an MIM-based method that reconstructs raw images according to the imaging principle of optical microscopy, which is formulated by the convolution of the PSF and the emitters.

A. PiMAE Model

The PiMAE model (Fig. 1) consists of three key components: (1) a vision transformer-based (ViT) [35] encoder–decoder architecture with a mask layer to prevent trivial solutions while estimating emitters, (2) a convolutional neural network as a prior for PSF estimation [36], and (3) a microscopic imaging simulator that implements the imaging principle formulated by PSF and emitter convolution. Appendix A provides detailed information on the network architecture and the embedding of physical principles. PiMAE requires only a few raw images for training, which is attributed to the carefully designed loss function. The loss function consists of two parts: one measures the difference between the raw and the reconstruction images, including the mean of the absolute difference and the multiscale structure similarity; the other part is a constraint on the PSF, including the total variation loss measuring the PSF continuity and the offset distance of the PSF's center of mass. Appendix B contains the expressions for the loss functions.

B. Training

The ViT-based encoder in PiMAE is pretrained on the COCO data set [37] to improve performance. The pretraining is based

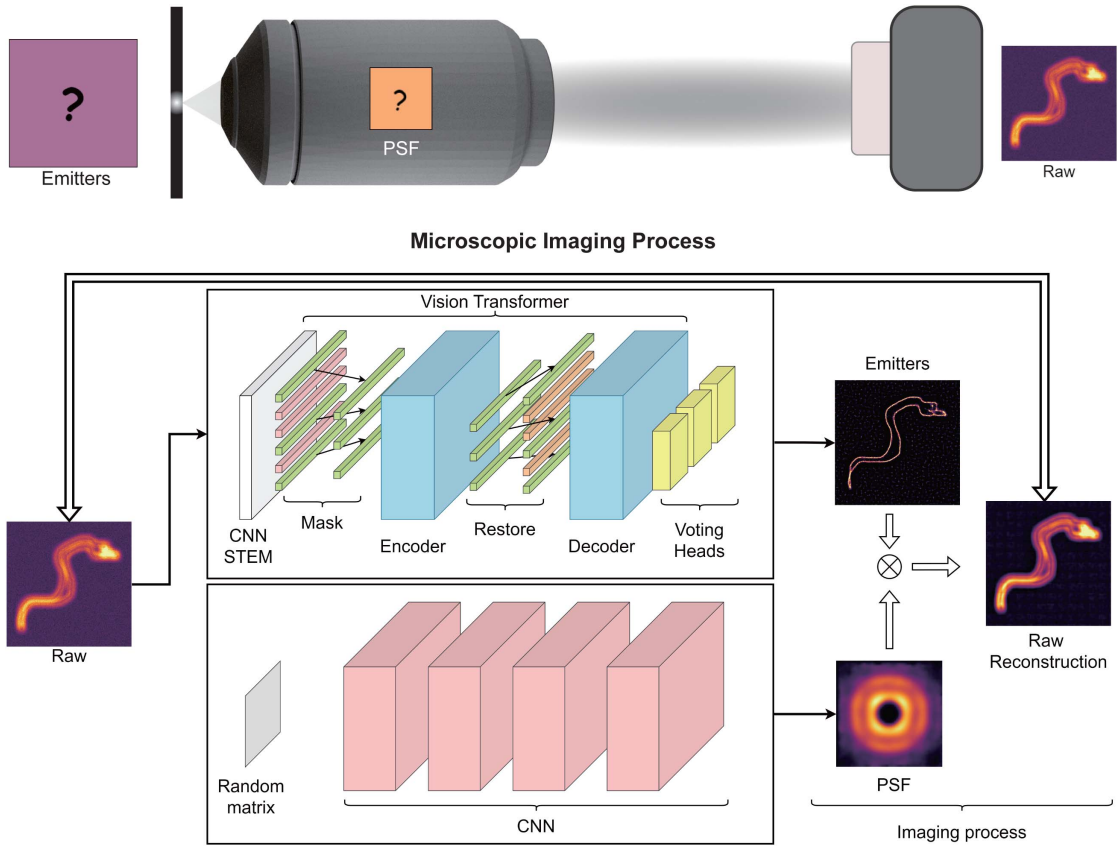

Fig. 1. PiMAE overview. PiMAE, a physics-informed masked autoencoder, is proposed to learn the imaging mechanism of an optical microscope.

on self-supervised learning on a masked autoencoder that does not include a physical simulator module (see Ref. [8] for details). After pretraining, PiMAE loads the trained encoder parameters and undergoes self-supervised training using raw microscopic images. The input image size is 144 × 144 pixels, and we use the RAdam optimizer [38] for training with a learning rate of 10^{-4} and a batch size of 18. The training runs for 5×10^4 steps.

Within PiMAE, the convolutional neural network shown in Fig. 1 is randomly initialized, takes a fixed random vector as input, and outputs the predicted PSF. Relevant details can be found in Appendix A. As PiMAE undergoes self-supervised training, the predicted PSF of the convolutional neural network (CNN) becomes more accurate and closer to the true PSF, as shown in Fig. 2. The experimental setup is shown in Fig. 3.

C. Synthetic Data Design and Evaluation

To evaluate PiMAE's performance, synthetic data sets were designed considering the following factors: (1) PiMAE's requirement for sparse emitter data, (2) the need for the emitter data without discrete points for more challenging PSF estimation tasks, (3) evaluation on standard Gaussian PSF and other challenging PSFs, (4) evaluation at various noise levels, and (5) evaluation at various emitter sparsity levels. Therefore, the Sketches data set [39] was chosen as the emitter, as described in Appendix D.1.A, and various commonly used PSFs were designed in Appendix D.2. The noise robustness is evaluated by adding noise to the raw images at different levels. Moreover, images with sparse lines of varying densities were generated as emitters to assess the impact of sparsity on PiMAE, as described in Appendix D.1.B.

For each scenario, we sample 1000 images as the training set and 100 images as the test set. For PSF estimation, we use Deconvblind [8] as a benchmark. For emitter localization, we use the Richardson–Lucy algorithm [20] and DeepSTORM [6] as reference methods. The results are measured by NRMSE [see Appendix F for definition and Appendix J for multiscale structural similarity (MS-SSIM) results]. Note that for the Richardson–Lucy and DeepSTORM tests, the PSF is assumed *a priori*, while for PiMAE, the PSF is treated as unknown.

D. Real-World Experiments

We evaluate PiMAE's performance in handling both standard and nonstandard PSF microscopy images in real-world experiments. Since the true emitter positions cannot be obtained, we use the BioSR [40] data set to evaluate PiMAE's handling of standard PSF microscopy images and compare it with SIM. Then, we use our custom-made wide-field microscope to produce out-of-focus and distorted PSF microscopy images to analyze PiMAE's performance in handling nonstandard PSF microscopy images.

In the experiment of wide-field microscopic imaging of nitrogen vacancy (NV) color centers, a 532 nm (Coherent Vendi 10 single longitudinal mode laser) laser passes through a customized precision electronic timing shutter, which controls the duration of the laser beams flexibly. The laser is then

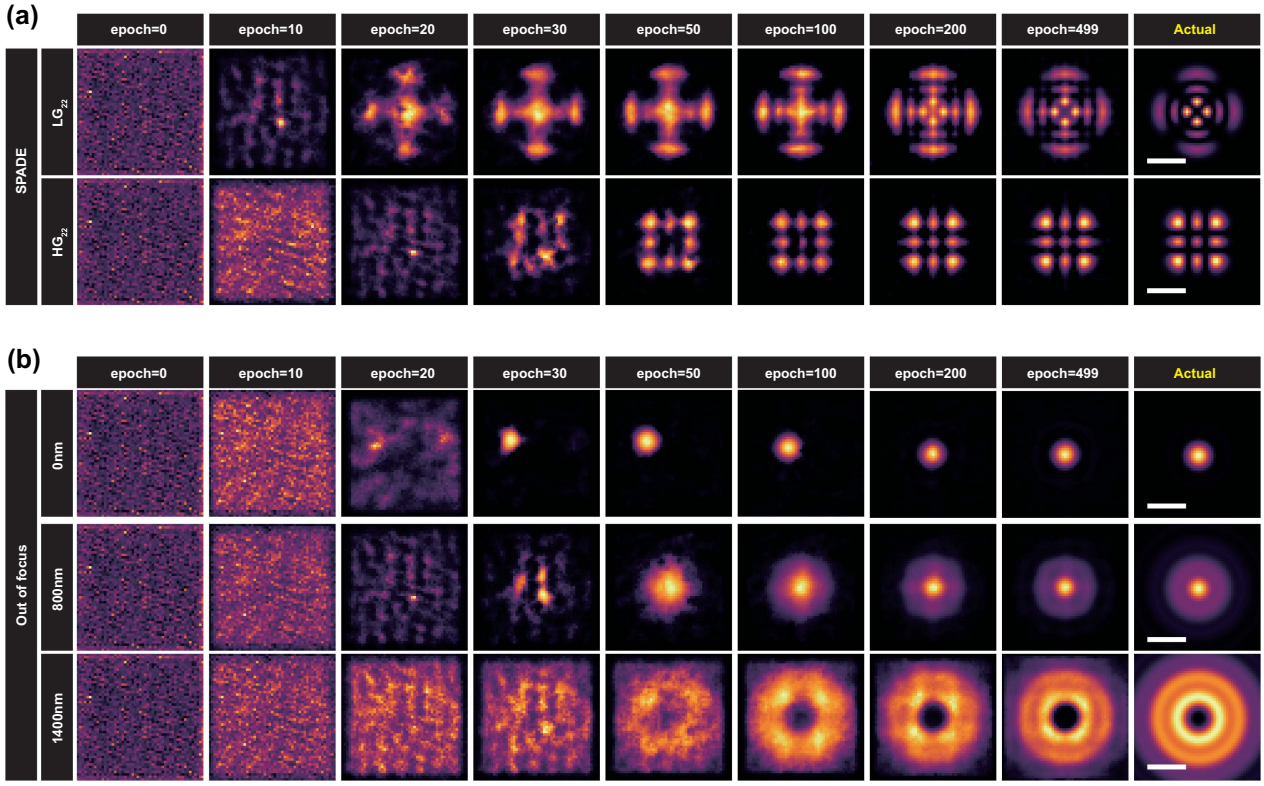

Fig. 2. PSF learning. The results demonstrate that PiMAE can successfully learn the PSF from raw images through the training process. (a) The figure displays the PSF of SPADE, including LG mode LG_{22} and HG mode HG_{22}. The scale bar is 0.5 μm. (b) Out-of-focus (800 and 1400 nm) images under a wide-field microscope imaging setup, along with the in-focus (0 nm) image. The scale bar is 0.5 μm.

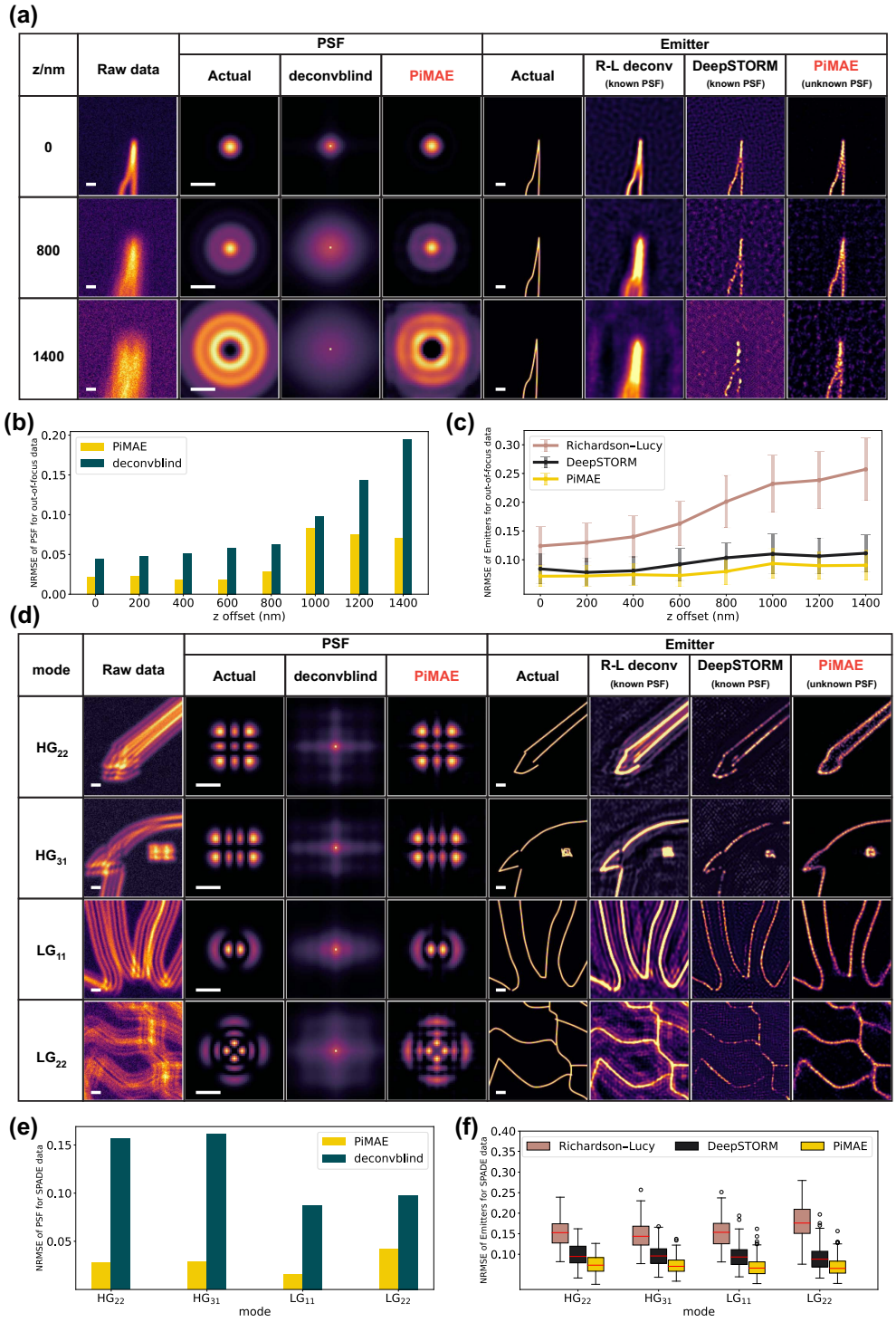

Fig. 3. Evaluation in synthetic data sets. (a) Results of estimated PSF and emitters from out-of-focus synthetic data. The scale bar is 0.5 μm. (b) NRMSE of the results of estimated PSF from out-of-focus synthetic data; (c) NRMSE of the results of estimated emitters from out-of-focus synthetic data. (d) Results of estimated PSF and emitters from synthetic data with HG mode and LG mode (HG/LG) as PSF. The scale bar is 0.5 μm. (e) NRMSE of the results of estimated PSF from HG/LG synthetic data; (f) NRMSE of the results of estimated emitters from HG/LG synthetic data. The noise scale in the above evaluations is $\text{noise}_{\text{std}}/\text{raw}_{\text{mean}} = 0.5$.

expanded and sent to a polarization mode controller that consists of a polarizing film (LPVISE100-A) and a half-wave plate (Thorlabs WPH10ME-532). The extended laser is focused on the focal plane behind the objective lens (Olympus, UPLFLN100XO2PH) by a fused quartz lens with a focal length of 150 mm. The fluorescence signals are collected by a scientific complementary metal oxide semiconductor (sCMOS) camera (Hamamatsu, Orca Flash 4.0 v.3). We use

a manual zoom lens (Nikon AF 70-300 mm, f/4-5.6G, focal length between 70 and 300 mm, and the field of view of 6.3) as a tube lens to continuously change the magnification of the microscopic system.

3. RESULT

A. PiMAE Achieves High Accuracy on Synthetic Data Sets

Being out-of-focus is one of the most common factors that can affect the quality of microscope imaging. PiMAE is capable of addressing this issue, and we demonstrate this by simulating a range of wide-field microscopy PSFs with out-of-focus distances that vary from 0 to 1400 nm. We also add Gaussian noise with a scale of $\text{noise}_{\text{std}}/\text{raw}_{\text{mean}} = 0.5$ to raw images, where $\text{noise}_{\text{std}}$ is the standard deviation of Gaussian noise [41] and raw_{mean} is the mean value of the raw image. First, we evaluate the performance of estimated PSFs. Figure 3(a) shows the actual PSFs and those estimated by Deconvblind and PiMAE. The PiMAE estimated PSF is similar to the actual PSF for all out-of-focus distances, while most of Deconvblind's estimated PSFs are far from the truth, indicating that Deconvblind cannot resolve raw images with complex PSFs. Furthermore, the estimated PSF by Deconvblind converges to the δ function after several iterations (see Appendix G). The NRMSE of the estimated PSFs at different out-of-focus distances is quantified in Fig. 3(b), with PiMAE achieving much better results than Deconvblind. Second, we evaluate the performance of estimated emitters. Figure 3(a) also shows the actual emitters and those estimated by the Richardson–Lucy algorithm, DeepSTORM (see Appendix H for implementation details), and PiMAE. When the out-of-focus distance is large, PiMAE and DeepSTORM significantly outperform the Richardson–Lucy algorithm. The NRMSE at different blur distances is shown in Fig. 3(c), where PiMAE achieves the best performance despite not knowing the actual PSF.

Recently, researchers have found that imaging resolution can be improved using a spatial pattern sorter [3,19,42], a method called SPADE. Using SPADE for confocal microscopy is equivalent to using PSFs corresponding to spatial modes [3], such as Zernike modes, Hermite–Gaussian (HG) modes, and Laguerre–Gaussian (LG) modes. However, SPADE faces several challenges, including the need for an accurate determination of the spatial mode (i.e., the equivalent PSF), high sensitivity to noise, and a lack of reconstruction algorithms for complex spatial modes. PiMAE can solve these problems. Figures 3(d)–3(f) show the SPADE imaging results with noise scale $\text{noise}_{\text{std}}/\text{raw}_{\text{mean}} = 0.5$. PiMAE can accurately estimate the equivalent PSF and emitters, and the performance is much better than that of the Deconvblind, Richardson–Lucy algorithm, and DeepSTORM. Therefore, PiMAE can significantly improve the performance of SPADE. These experiments demonstrate that PiMAE is effective for scenarios with unknown and complex imaging PSFs.

B. Noise Robustness

Noise robustness is a crucial metric for evaluating reconstruction algorithms. We evaluate noise robustness in three scenarios: (1) in-focus wide-field microscopy; (2) wide-field microscopy at 600 nm out-of-focus distance; and (3) Laguerre–Gaussian mode LG_{22} SPADE imaging. The raw image of each scenario contains Gaussian noise (the speckle noise results are shown in Appendix I) at scales ($\text{noise}_{\text{std}}/\text{raw}_{\text{mean}}$) of 0.01, 0.1, 0.5, 1, and 2, as shown in Fig. 4 (see Appendix J for MS-SSIM results). We first compare the results of Deconvblind and PiMAE for estimating PSF. We find that PiMAE shows excellent noise immunity, substantially outperforming Deconvblind in all tests. We then compare the results of the Richardson–Lucy algorithm, DeepSTORM, and PiMAE for estimating the emitters. Overall, PiMAE performs the best, only slightly behind DeepSTORM in the standard PSF scenario at low noise. The Richardson–Lucy algorithm performs similarly to DeepSTORM and PiMAE when the noise scale is very small. However, when the noise scale slightly increases, its performance significantly decreases. This shows the advantage of deep-learning-based methods over traditional algorithms in terms of noise robustness. Moreover, the advantage of PiMAE over the other two algorithms increases as the scale of the noise becomes larger and the shape of the PSF becomes more complex.

C. PiMAE Enables Superresolution Imaging for Wide-Field Microscopy Comparable to SIM

The endoplasmic reticulum (ER) is a system of tunnels surrounded by membranes in eukaryotic cells. In the data set BioSR [40], the researchers imaged the ER in the same field of view using wide-field microscopy and SIM, respectively. Figure 5(a) shows the results of PiMAE-resolved wide-field microscopy raw images (more images of the results are in Appendix K). We find that the resolution of the PiMAE-estimated emitter is comparable to that of SIM, which has a resolution twice that of the diffraction limit. Figure 5(b) shows the cross-sectional results, where the peak positions of the PiMAE-estimated emitter match the peak positions of the SIM results, corresponding to indistinguishable wide-field imaging results. This indicates that the resolvability of the results of wide-field microscopy with PiMAE-estimated emitters is improved to a level similar to that of SIM. Figure 5(c) shows the results of the PiMAE-estimated PSF with FWHM of 230 nm. The fluorescence wavelength of the raw image is 488 nm, the numerical aperture (NA) is 1.3, and its diffraction limit is $0.61 \times \frac{\lambda}{\text{NA}} = 0.61 \times \frac{488\ \text{nm}}{1.3} \approx 229$ nm, which is very close to the FWHM of the PiMAE-estimated PSF. This experiment shows that PiMAE can be applied to real-world experiments to estimate PSF from raw microscopy data and further improve resolution.

D. PiMAE Enables Imaging for Nonstandard Wide-Field Microscopy

The NV color center is a point defect in diamond that is widely used in superresolution microscopy [5,43] and quantum sensing [44,45]. We make a home-built wide field microscope to image the NV center in fluorescent nanodiamonds (FNDs) at out-of-focus distances of 0, 400, and 800 nm. We take 10 raw images with a size of 2048 pixels and a field-of-view size of 81.92 μm at each out-of-focus distance. Figure 6(a) shows that we image NV color centers in the same field of view at different out-of-focus distances, and Fig. 6(b) shows the corresponding PiMAE-estimated emitters. This is a side-by-side demonstration of the accuracy of the PiMAE-estimated emitters. The

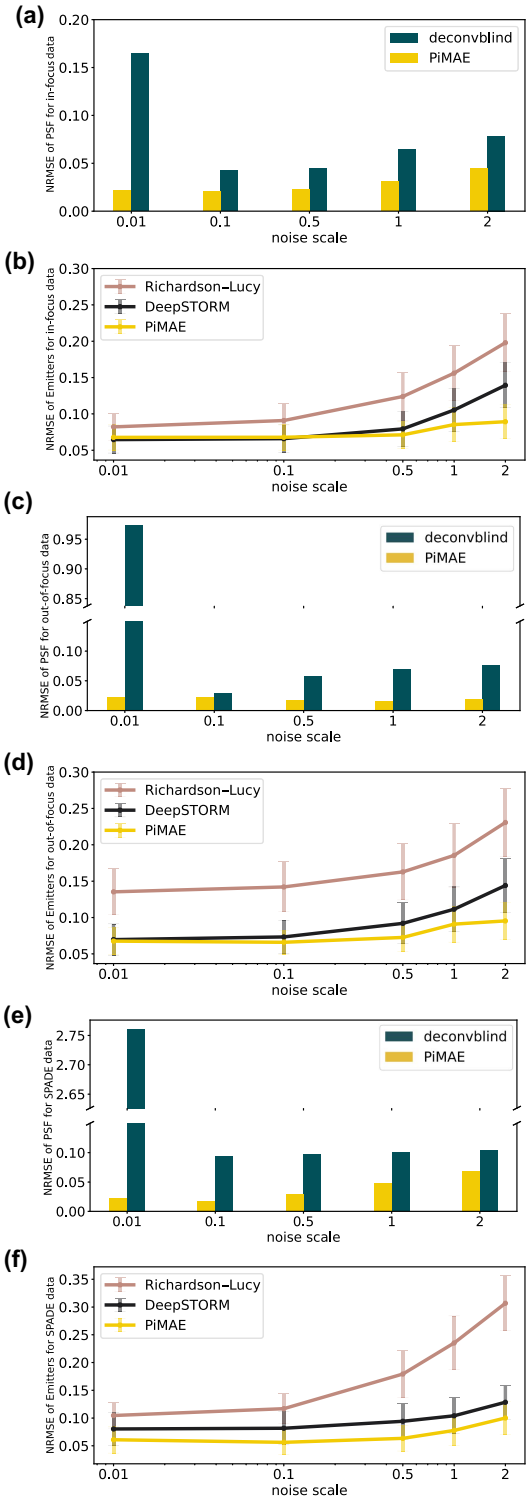

Fig. 4. Evaluation of noise robustness. (a) NRMSE of the results of estimated PSF from in-focus synthetic data; (b) NRMSE of the results of estimated emitters from in-focus synthetic data; (c) NRMSE of the results of estimated PSF from 600 nm out-of-focus synthetic data. (d) NRMSE of the results of estimated emitters from 600 nm out-of-focus synthetic data; (e) NRMSE of the results of estimated PSF from LG_{22} synthetic data; (f) NRMSE of the results of estimated emitters from LG_{22} synthetic data; the noise scale is $\text{noise}_{\text{std}}/\text{raw}_{\text{mean}}$.

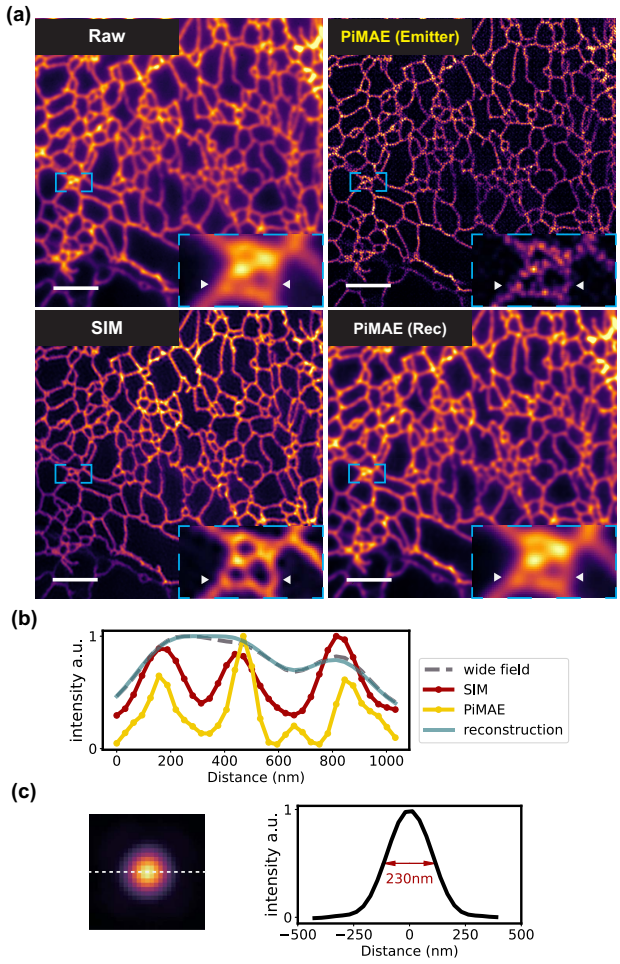

Fig. 5. Superresolution imaging of ER. (a) The figures are the raw image of wide-field microscopic imaging of ER, the result of estimating the emitter from wide-field microscopic imaging using PiMAE, the result of SIM of the same field of view, and the result of wide-field microscopic imaging reconstructed by PiMAE. Data from BioSR data set [40]. The scale bar is 2.50 μm. (b) Comparison of the cross section of the PiMAE estimated emitters and SIM results; it shows that the resolution of the results obtained by PiMAE is comparable to that of SIM. (c) PiMAE estimated wide-field microscope PSF with an FWHM of 230 nm, where the diffraction limit is 229 nm.

out-of-focus distance changes during the experiment, but the field of view is invariant. Therefore, the PiMAE-estimated emitter position should be constant at each out-of-focus distance, as we observe in Figs. 6(b) and 6(c). Figure 6(d) shows the variation of the PSF. The asymmetry of the PSF comes from the slight tilt of the carrier stage. Also, we show the PSF cross section for each scene. The FWHM of the estimated PSF at focus is 382 nm, which corresponds to a diffraction limit of 384 nm. This suggests that PiMAE can be applied in real-world experiments to improve the imaging capabilities of microscopes suffering from out-of-focus.

Moreover, we construct a nonstandard PSF for wide-field microscopic imaging of NV color centers by making the objective mismatch with the coverslip (see Appendix K.2); the results are shown in Figs. 6(e)–6(g). Figure 6(e) shows the imaging

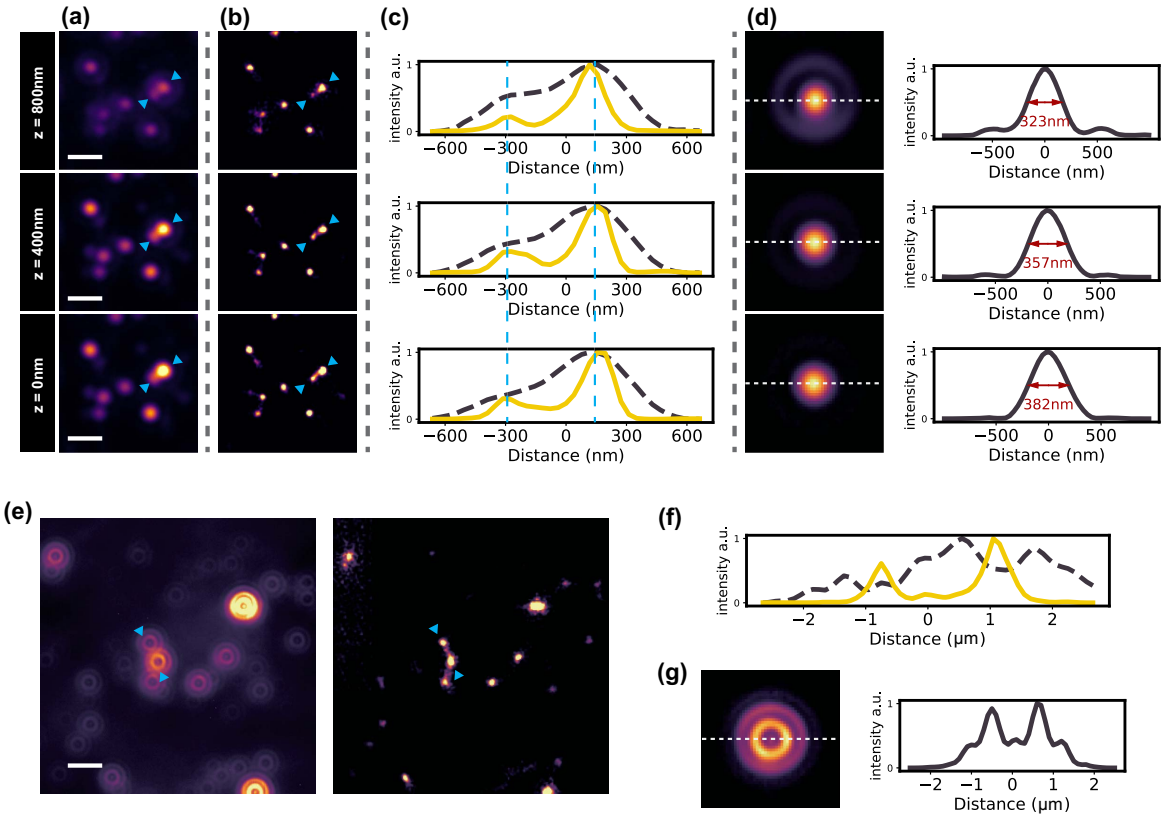

Fig. 6. Wide-field microscopy imaging of NV color centers. (a)–(d) Results of wide-field microscopy imaging of NV color centers at different out-of-focus distances; (a) raw images; the scale bar is 1.25 μm. (b) PiMAE estimated emitters; (c) comparison of the cross section of the raw images and the PiMAE-estimated emitters, where the black dashed line represents the raw images and the yellow solid line represents the PiMAE-estimated emitters; the peak positions of the PiMAE-estimated emitter results are constant for different out-of-focus distances, as seen from the blue dashed line. (d) PiMAE estimated PSF; FWHM of in-focus PSF is 382 nm, where the diffraction limit is 384 nm; the larger the out-of-focus distance, the larger the paraflap of the PSF, despite the decrease of the FWHM in the center region. (e) Comparison of nonstandard microscopic imaging and PiMAE estimated emitters. The scale bar is 3.2 μm. (f) Cross section of the nonstandard microscopic imaging and PiMAE estimated emitters; (g) PiMAE-estimated nonstandard microscopy PSF.

results and PiMAE-estimated emitters. Figure 6(f) shows the results of the cross-sectional comparison. Figure 6(g) shows the PiMAE-estimated PSF. This experiment demonstrates that PiMAE enables researchers to use microscopy with nonstandard PSFs for imaging. And PiMAE's ability to resolve nonstandard PSFs expands the application scenarios of NV color centers in fields such as quantum sensing and bioimaging.

E. PiMAE Enables Microscopy Imaging with Widely Spread Out PSFs

Further testing the capabilities of PiMAE, we evaluate the performance of PiMAE on complex widely spread out PSFs, represented by the character "USTC." We use 1000 images as the training set and 100 images as the test set. The noise level is set at $\text{noise}_{\text{std}}/\text{raw}_{\text{mean}} = 0.01$. The results of the raw images, the PiMAE processed images, and the evaluation of the NRMSE metric are depicted in Fig. 7. PiMAE performed exceptionally well, demonstrating its effectiveness in difficult scenarios.

F. Evaluation of the Influence of Emitter Sparsity

Dense samples can pose challenges for estimating both the PSF and the emitters. We designed emitters with varying densities, as outlined in Appendix D.1.B, and employed LG_{22} as the PSF.

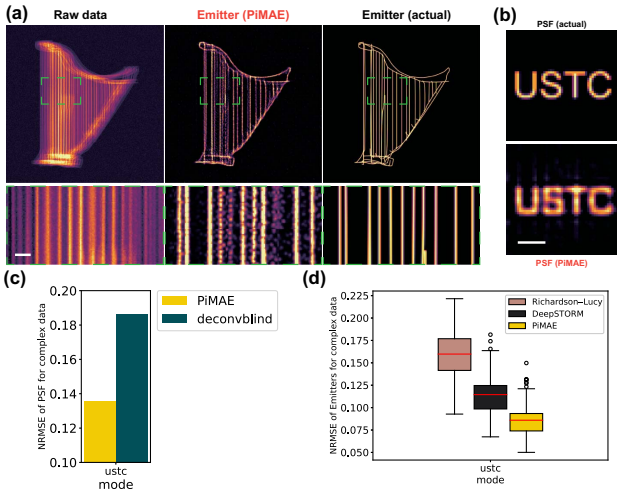

Fig. 7. Evaluation using synthetic data based on PSF of the shape "USTC." (a) Comparison of the raw image, the PiMAE estimated emitters, and the actual emitters; the scale bar is 0.5 μm. (b) Comparison of the actual PSF and the PiMAE-estimated PSF; the scale bar is 0.5 μm. (c) NRMSE of the estimated PSF; (d) NRMSE of the estimated emitters.

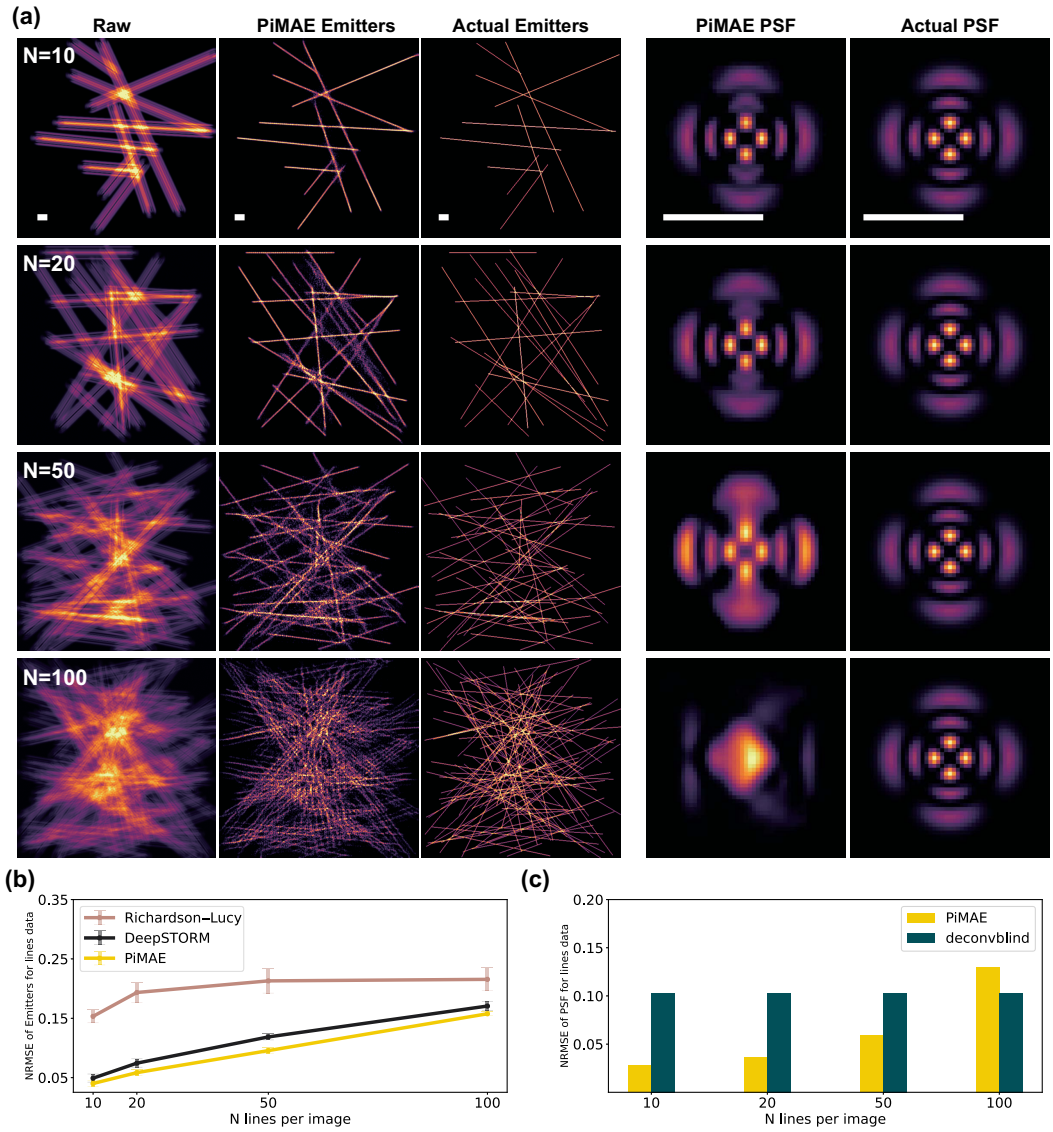

Fig. 8. Influence of emitter sparsity. (a) Comparison of the raw image, the PiMAE estimated emitters, and the actual emitters, and comparison of the actual PSF and the PiMAE-estimated PSF; N refers to the number of sparse lines. The scale bar is 1.0 μm. (b) NRMSE of the estimated emitters; (c) NRMSE of the estimated PSF.

As shown in Fig. 8, we observe that as the number of lines in each image (512 × 512) increases, PiMAE's performance in estimating both the PSF and emitters deteriorates. Intuitively, when the number of lines in each image is less than or equal to 50, PiMAE performs well, while performance is poor when the number of lines is greater than 50. This process allows us to evaluate the influence of emitter sparsity on PiMAE.

G. Computational Resource and Speed

In this work, the code is based on the Python library PyTorch, as we show in Code File 1 [46]. PyTorch is a prominent open-source deep-learning framework that offers an efficient and user-friendly platform for building and deploying deep-learning models. In terms of model training, we utilize three Nvidia Tesla A100 40 GB graphics cards in parallel, which is necessary due to ViT's substantial computational and memory requirements. The training time for each task is 11 h, and the inference time for a single 512 × 512 image is approximately 4 s with the trained model. Compared to supervised models such as DeepSTORM, which takes about 1 h for training and 0.1 s for inference, PiMAE is slower but more powerful. As for the data set size requirement, we show in Appendix E that PiMAE achieves good training results, even with a minimum of five images in the training set.

4. DISCUSSION

In this study, we introduce PiMAE, a novel approach for estimating PSF and emitters directly from raw microscopy images. PiMAE addresses several challenges: it allows for direct identification of the PSF from raw data, enabling deep-learning model training without the need for real-world or synthetic annotation; it has excellent noise resistance; and it is convenient

and widely applicable, requiring only about five raw images to resolve the PSF and emitters.

Our method, PiMAE, extracts hidden variables from raw data using physical knowledge. By recognizing PSF as a hidden variable in a linear optical system, the underlying physical principle involves the decomposition of raw data through the convolution of the emitters with the PSF. Hidden variables are ubiquitous in real-world experiments, by integrating masked autoencoder and physical knowledge, PiMAE provides a framework to solve hidden variables in physical systems through self-supervised learning.

However, it should be noted that PiMAE is an emitter localization algorithm, which means that it requires a sufficient degree of sample sparsity to perform effectively. We conducted an evaluation using synthetic data experiments, and while PiMAE performed reasonably well, there is still room for improvement. There is ambiguity in extracting the PSF and emitters directly from the raw images, so PiMAE opts for a simpler emitter distribution to learn the real PSF, which might result in artifacts. As PiMAE supplies the PSF needed for RL-deconv and DeepSTORM, potential solutions may be to integrate PiMAE with the aforementioned methods or to perform unmasked self-supervised training after masked self-supervised training within PiMAE. Therefore, future work could focus on further enhancing the robustness of PiMAE for use in dense scenarios.

5. CONCLUSION

In conclusion, we have presented PiMAE, a novel solution for directly extracting the PSF and emitters from raw optical microscopy images. By combining the principles of optical microscopy with self-supervised learning, PiMAE demonstrates impressive accuracy and noise robustness in synthetic data experiments, outperforming existing methods such as DeepSTORM and the Richardson–Lucy algorithm. Appendix L shows the full range of synthetic data evaluation metrics. Moreover, our method has been successfully applied to real-world microscopy experiments, resolving wide-field microscopy images with various PSFs. With its ability to learn the hidden mechanisms from raw data, PiMAE has a wide range of potential applications in optical microscopy and scientific studies.

APPENDIX A: NETWORK ARCHITECTURE

The principle of microscopic imaging is

$$\text{raw image} = \text{noise}(\text{emitters} \otimes \text{PSF}) + \text{background}, \quad \textbf{(A1)}$$

where the raw image is the result of convolving the emitters and the PSF with the presence of noise and background. To put this principle into practice, we have developed the PiMAE method, which consists of three modules: emitter inference from raw images, PSF generation, and background separation.

1. Emitter Inference

We have improved the original masked autoencoder for use in microscopic imaging by integrating a voting head into its transformer-based decoder. The head predicts the position and intensity of emitters, respectively. Specifically, the decoder produces 9×9 feature patches, which serve as the input for the voting head. For the emitter position, the voting head employs a two-step process: (1) a multilayer perceptron (MLP) predicts 64 density maps from each feature patch, and (2) the emitter positions are obtained by computing the center of mass of each density map. For emitter intensity, an MLP predicts 64 intensities. The predicted emitter image is generated by placing a Gaussian-type point tensor with $\sigma = 1$ scaled by its corresponding intensity at the predicted position, similar to the design in crowd-counting methods [47]. The mask layer is an essential element in the design of a masked autoencoder. Its main function is to prevent the model from learning trivial solutions and instead encourage it to focus on the relevant features of the input data. This is achieved by randomly blocking out specific parts of the input tensor. To improve the training efficiency, we introduced a CNN stem consisting of four convolutional layers placed before the mask layer [48]. The input image size of 144×144 is reduced to 9×9 after the CNN stem, with each pixel encoding a 384-dimensional vector. We refer to this model as the point predictor, as shown in Fig. 9.

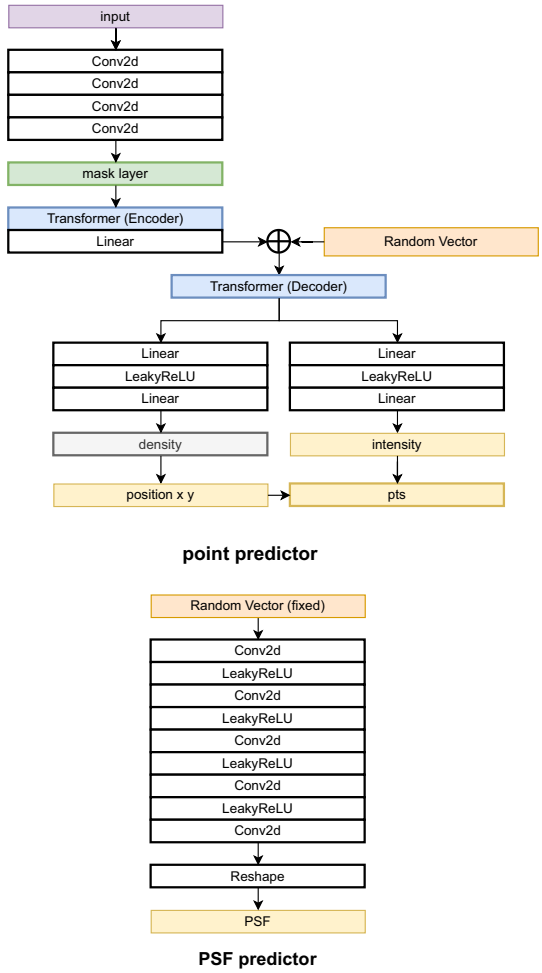

Fig. 9. Network architecture. PiMAE consists of two predictors, namely, a PSF predictor and a point predictor. The point predictor outputs the location and intensity of the points.

2. PSF Generation

Motivated by the observation that a CNN can function as a well-designed prior and deliver outstanding results in typical inverse problems, as evidenced by Deep Image Prior [36], we constructed the PSF generator, as illustrated in Fig. 9. The neural network's parameters are adjusted through self-supervised learning to produce the PSF, with a random matrix as the input, which remains constant throughout the learning process.

3. Background Separation

To isolate the background component from the raw image, we employ a new point predictor (Fig. 9). We assume that the background has a low spatial variability and approximate it by drawing the output points from the point predictor following a Gaussian distribution with $\sigma = 16$.

APPENDIX B: DESIGN OF LOSS FUNCTION

The loss function in our approach is composed of four components, divided into two categories.

The first category measures the similarity between the reconstructed image and the raw image. It consists of the mean absolute difference (L1) and the MS-SSIM, as expressed in Eq. (F2). The combination of these two functions has been demonstrated to perform better than individual functions such as L1 and mean squared error (MSE) in image restoration tasks [49].

The second category concerns the constraint on the generated PSF. To ensure that the center of mass of the generated PSF is at the center of the PSF tensor, we calculate the center distance loss as follows:

$$\text{Center distance loss} = \left| \frac{\sum_{i,j} \text{Intensity}_{ij} \cdot \overrightarrow{\text{Coordinate}_{ij}}}{\sum_{i,j} \text{Intensity}_{ij}} - \overrightarrow{\text{Center position}} \right|. \tag{B1}$$

Additionally, to ensure that the generated PSF is spatially continuous, we use the total variation (TV) loss to quantify the smoothness of the image,

$$\text{TV loss} = \sum_{i,j} (\text{Intensity}_{i,j-1} - \text{Intensity}_{i,j})^2 + (\text{Intensity}_{i+1,j} - \text{Intensity}_{i,j})^2. \tag{B2}$$

Finally, the loss function is defined as

$$\text{Loss function} = \alpha_1 \cdot \text{L1} + \alpha_2 \cdot \text{MS-SSIM} + \alpha_3 \cdot \text{Center distance} + \alpha_4 \cdot \text{TV}, \tag{B3}$$

where $\alpha_1 = 0.95$, $\alpha_2 = 0.05$, $\alpha_3 = 0.001$, and $\alpha_4 = 0.001$.

APPENDIX C: PRETRAINING WITH COCO DATA SET

Recent research has shown that self-supervised pretraining is effective in improving accuracy and robustness in computer vision tasks. In this study, we employed a masked autoencoder [shown in Fig. 10(b)] to pretrain the encoder of PiMAE on the COCO data set [37] (unlabeled), a large-scale data set containing 330,000 RGB images of varying sizes for object detection, segmentation, and captioning tasks.

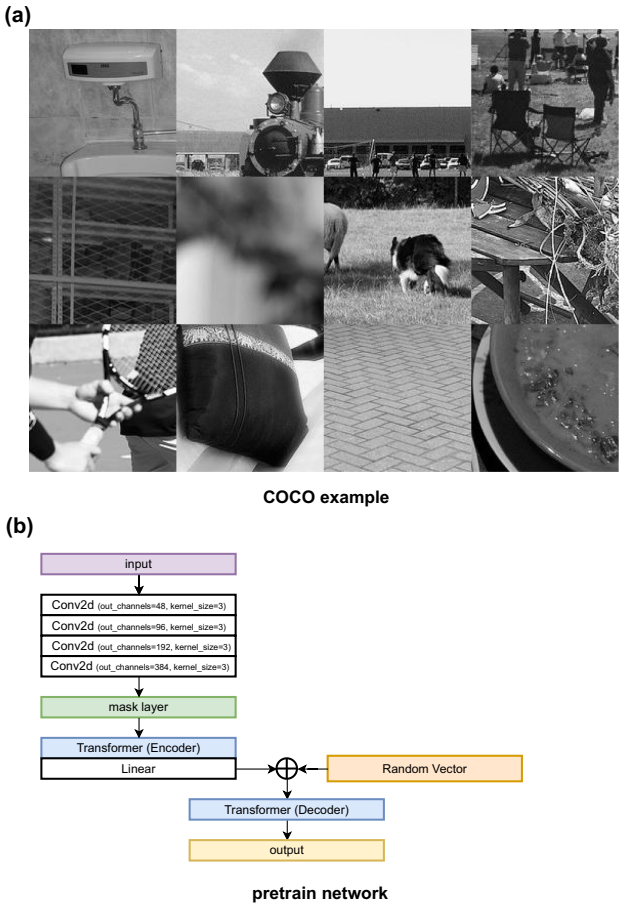

Fig. 10. Pretraining with COCO data set.

For pretraining, we randomly cropped 144×144 portions from the images and transformed them into gray-scale images to form the training set. Examples of the cropped images are shown in Fig. 10(a).

We use the MSE as the loss function during the training process, with a learning rate of 10^{-4} and 500 training epochs. A masking rate of 75% is implemented, and the RAdam optimizer is used. The results of the MAE reconstruction after pretraining can be seen in Fig. 11. Figure 12 demonstrates that the pretraining process has significantly contributed to the enhancement of the localization of emitters' performance.

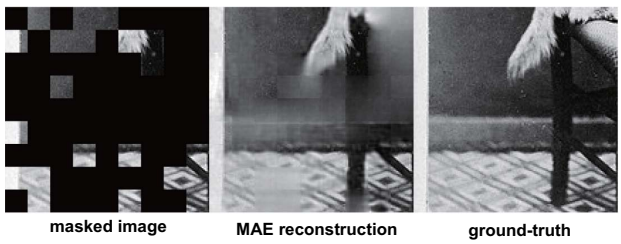

Fig. 11. Example results on COCO. We show the masked image, MAE reconstruction, and the ground truth. The masking ratio here is 0.75.

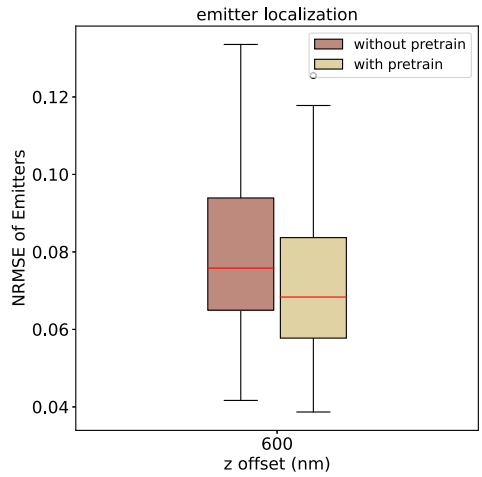

Fig. 12. Pretraining enhancements. Comparison of NRMSE metrics for emitter localization of pretrained and non-pretrained models. Using 600 nm out-of-focus data as an example, after 500 rounds of training, the learning rate is 3×10^{-4}.

When the MAE pretraining is finished, the parameters of the encoder and decoder are stored for the subsequent training of PiMAE.

APPENDIX D: SYNTHETIC DATA GENERATION

In this section, we present the construction method of the synthetic data used to evaluate PiMAE, including emitters and PSFs.

1. Emitters

A. Sketches

Sketches data set [39] is a large-scale exploration of human sketches containing a wide variety of morphologies. To evaluate the performance of the method, emitters of synthetic data are sampled from the Sketches data set. Figure 13 illustrates examples from the Sketches data set.

B. Random Lines

To evaluate the performance of the model under various levels of sparsity, we implement an algorithm to generate images containing N randomly generated lines.

1. A black image of size 512×512 is created.
2. A loop is executed N times to randomly draw lines on the image. In each iteration:

   a. The starting and ending points of a line are randomly generated.
   b. The intensity of the line is randomly generated.
   c. The line is drawn on the image.

3. The image is smoothened to remove jaggedness.

The resulting emitters are shown in Fig. 14.

2. PSFs

A. Out-of-Focus

We simulate the imaging results of a wide-field microscope when the sample is out of focus. The near-focus amplitude can be described using the scalar Debye integral [50],

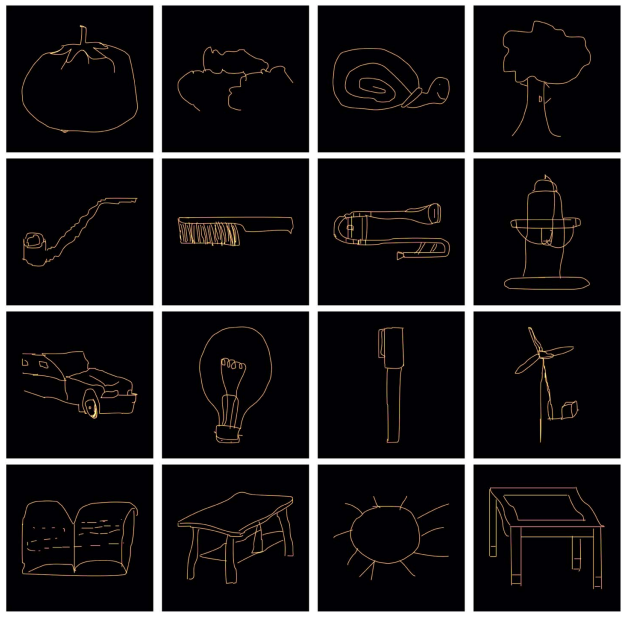

Fig. 13. Sketches data set examples.

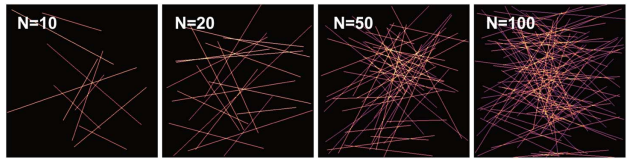

Fig. 14. Randomly generated lines.

$$h(x,y,z;\lambda) = C_0 \int_0^{\alpha} \sqrt{\cos\theta}\, J_0(k\rho \sin\theta) e^{-ikz\cos\theta} \sin\theta \mathrm{d}\theta, \tag{D1}$$

where C_0 is a complex constant, J_0 is the zeroth-order Bessel function of the first kind, $\rho = \sqrt{x^2 + y^2}$, the refractive index is n, the numerical aperture $\mathrm{NA} = n \sin\alpha$, and the wavenumber $k = n(2\pi/\lambda)$. The PSF of the wide-field microscopy is

$$\mathrm{PSF}(x,y,z) = |h(x,y,z;\lambda_{\mathrm{em}})|^2. \tag{D2}$$

The values of the parameters in this experiment are $C_0 = 1$, $n = 1$, $\lambda_{\mathrm{em}} = 400$ nm, $\mathrm{NA} = 0.7$, and each pixel has a size of 39 nm. λ_{em} represents the fluorescence emission wavelength.

B. SPADE

We simulated four scenarios in the SPADE, corresponding to PSFs as Hermite–Gaussian modes HG_{22}, HG_{31}, and Laguerre–Gaussian modes LG_{11}, LG_{22}, respectively. Here we set the wavelength to 500 nm, the PSF size to 51×51 pixels and 15 mm $\times$ 15 mm range, and rescaled to a 39 nm pixel size. The definitions for the amplitude of the Hermite–Gaussian modes and Laguerre–Gaussian modes are [51]

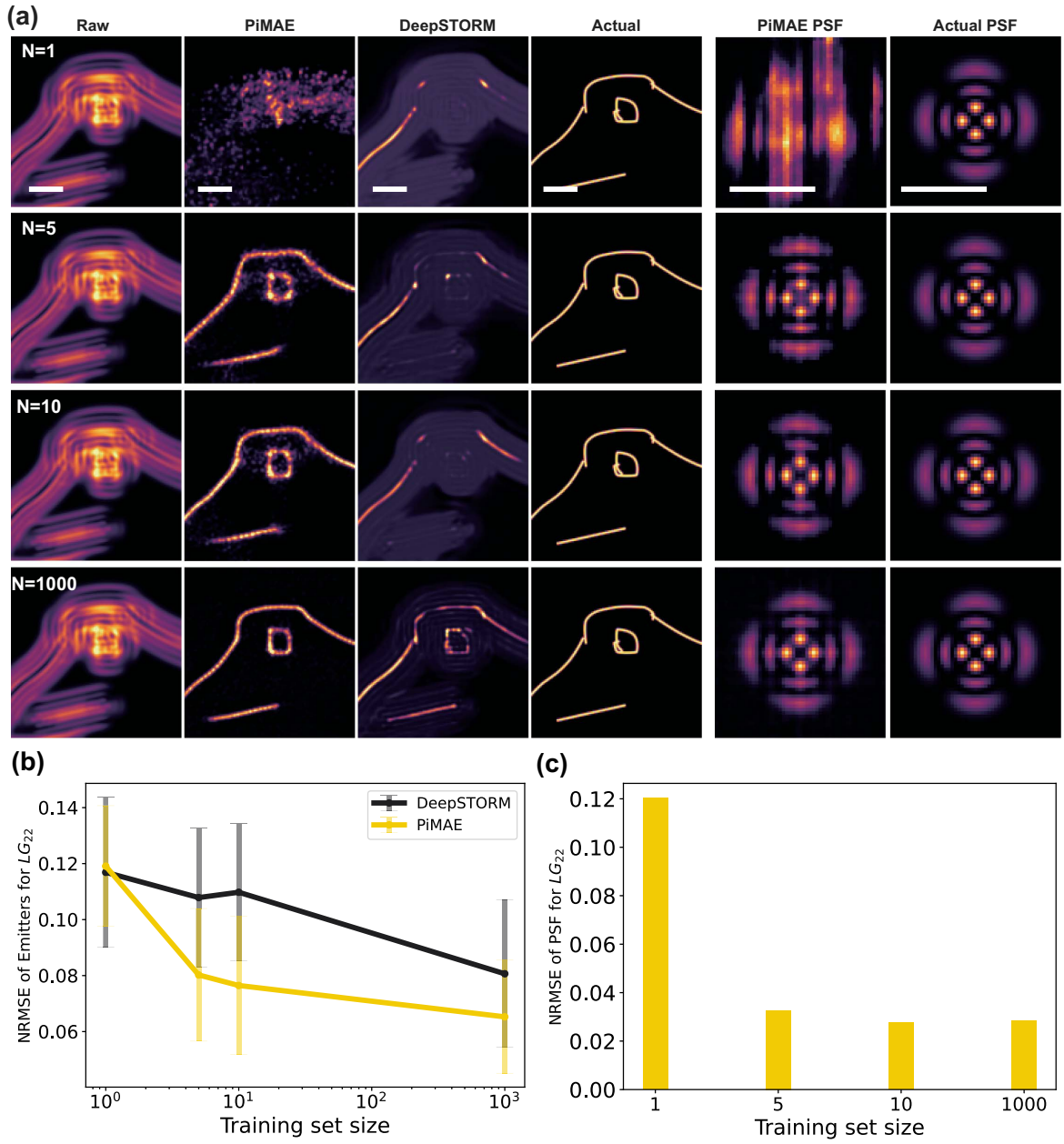

Fig. 15. Evaluating the effect of training set size. (a) Results of estimated PSF and emitters when the size N of the training set is 1, 5, 10, and 1000, and the size of the test set is 100; the scale bar is 1.0 μm. (b) NRMSE of the estimated emitters from synthetic data with different data set sizes; (c) NRMSE of PiMAE-estimated PSF from synthetic data with different data set sizes.

$$u_{nm}^{\mathrm{HG}}(x,y,z) = C_{nm}^{\mathrm{HG}}(1/w)\exp[-ik(x^2+y^2/2R)] \times \exp[-(x^2+y^2/w^2)]\exp[-i(-n+m+1)\psi] \times H_n\left(x\sqrt{2}/w\right)H_m\left(y\sqrt{2}/w\right), \quad \textbf{(D3)}$$

$$u_{nm}^{\mathrm{LG}}(r,\phi,z) = C_{nm}^{\mathrm{LG}}(1/w)\exp(-ikr^2/2R)\exp(-r^2/w^2) \times \exp[-i(n+m+1)\psi]\exp[-i(n-m)\phi] \times (-1)^{\min(n,m)}\left(r\sqrt{2}/w\right)^{|r-m|} \times L_{\min(n,m)}^{|n-m|}(2r^2/w^2), \quad \textbf{(D4)}$$

with $R(z) = (z_R^2+z^2)/z_R$, $\frac{1}{2}kw^2(z) = (z_R^2+z^2)/z_R$, and $\psi(z) = \arctan(z/z_R)$. $H_n(x)$ is the Hermite polynomial of order n, $L_p^l(x)$ is the generalized Laguerre polynomial, $k = \frac{2\pi}{\lambda}$ is the wavenumber, and z_R is the Rayleigh range of the mode. Here we set $w_0 = 2$ mm, wavelength $\lambda = 500$ nm, and $z = 0$.

APPENDIX E: TRAINING SET SIZE

We use LG_{22} as the PSF and a fixed test set size of 100 images with a shape of 512×512. The training set sizes for both PiMAE and DeepSTORM are 1, 5, 10, and 1000 images,

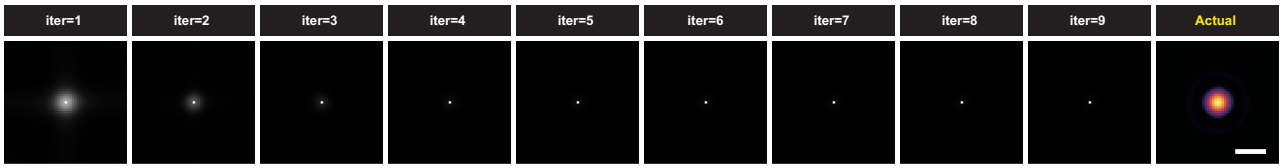

Fig. 16. Iterative optimization in Deconvblind. The PSF estimated by Deconvblind converges to a δ function. The scale bar is 0.5 μm.

respectively. As shown in Fig. 15, PiMAE performs well, even with a training set size as small as five images, whereas the performance of DeepSTORM decreases significantly.

APPENDIX F: ASSESSMENT METRICS

When evaluating the performance of emitter estimation, we use two metrics: the NRMSE and the MS-SSIM. NRMSE provides a quantitative measure of the difference between two images, while MS-SSIM is designed to assess the perceived similarity of images, taking into consideration the recognition of emitters by the human eye [52].

NRMSE is defined as

$$\mathrm{NRMSE} = \frac{\sqrt{\sum_{i,j}(\mathrm{Image}_{\mathrm{true}} - \mathrm{Image}_{\mathrm{test}})^2}}{\mathrm{Max}(\mathrm{Image}_{\mathrm{true}}) - \mathrm{Min}(\mathrm{Image}_{\mathrm{true}})}. \tag{F1}$$

MS-SSIM is defined as

$$\mathrm{MS\text{-}SSIM}(x,y) = [l_m(x,y)]^{\alpha_m} \cdot \prod_{j=1}^{M} [c_j(x,y)]^{\beta_j} [s_j(x,y)]^{\gamma_j}, \tag{F2}$$

where the exponents α_m, β_j, and γ_j are used to adjust the relative importance of different components. Here $\alpha_m = \beta_j = \gamma_j$ and values are 0.0448, 0.2856, 0.3001, 0.2363, 0.1333 for $j = 1, 2, 3, 4, 5$. The expressions of the exponents l_m, c_j, and s_j are the same as single-scale structural similarity at each scale j,

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \tag{F3}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \tag{F4}$$

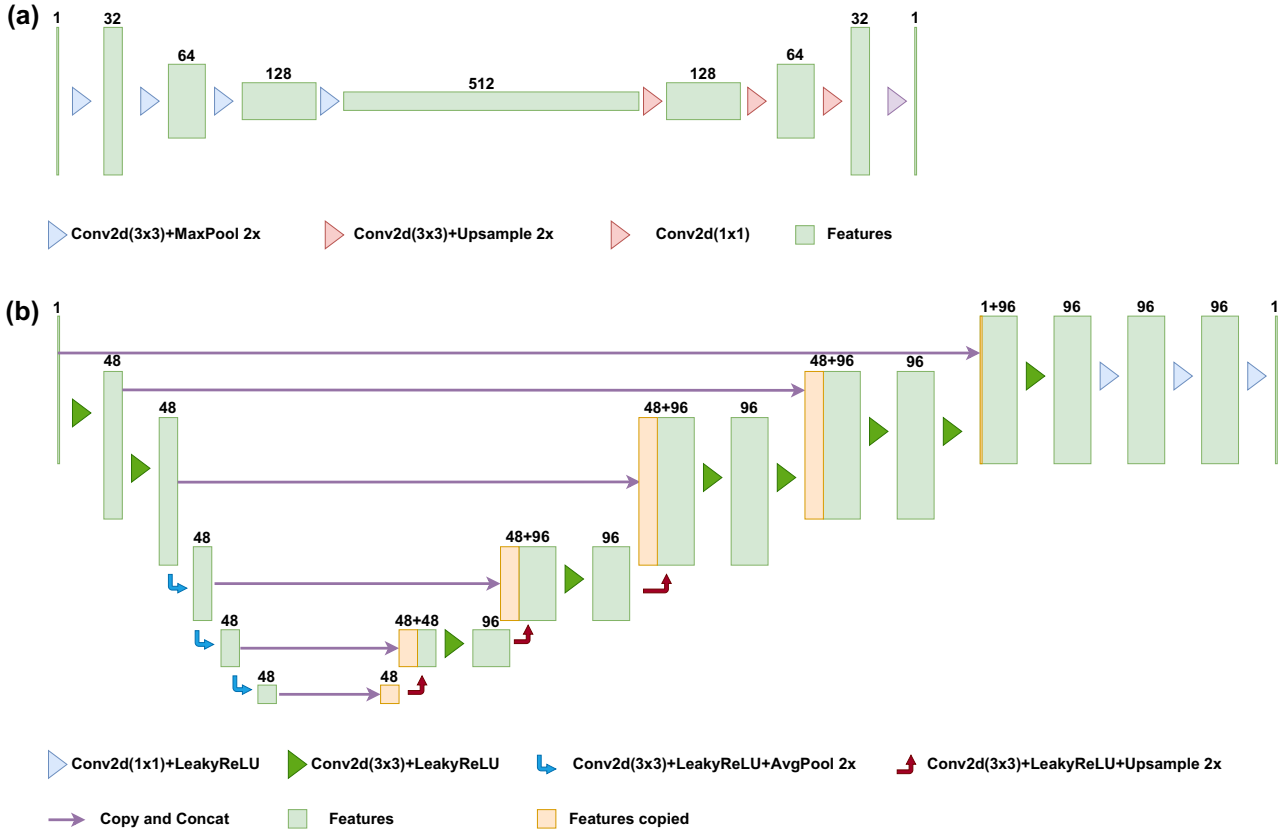

Fig. 17. Network architecture. (a) Original DeepSTORM architecture; (b) modified DeepSTORM architecture.

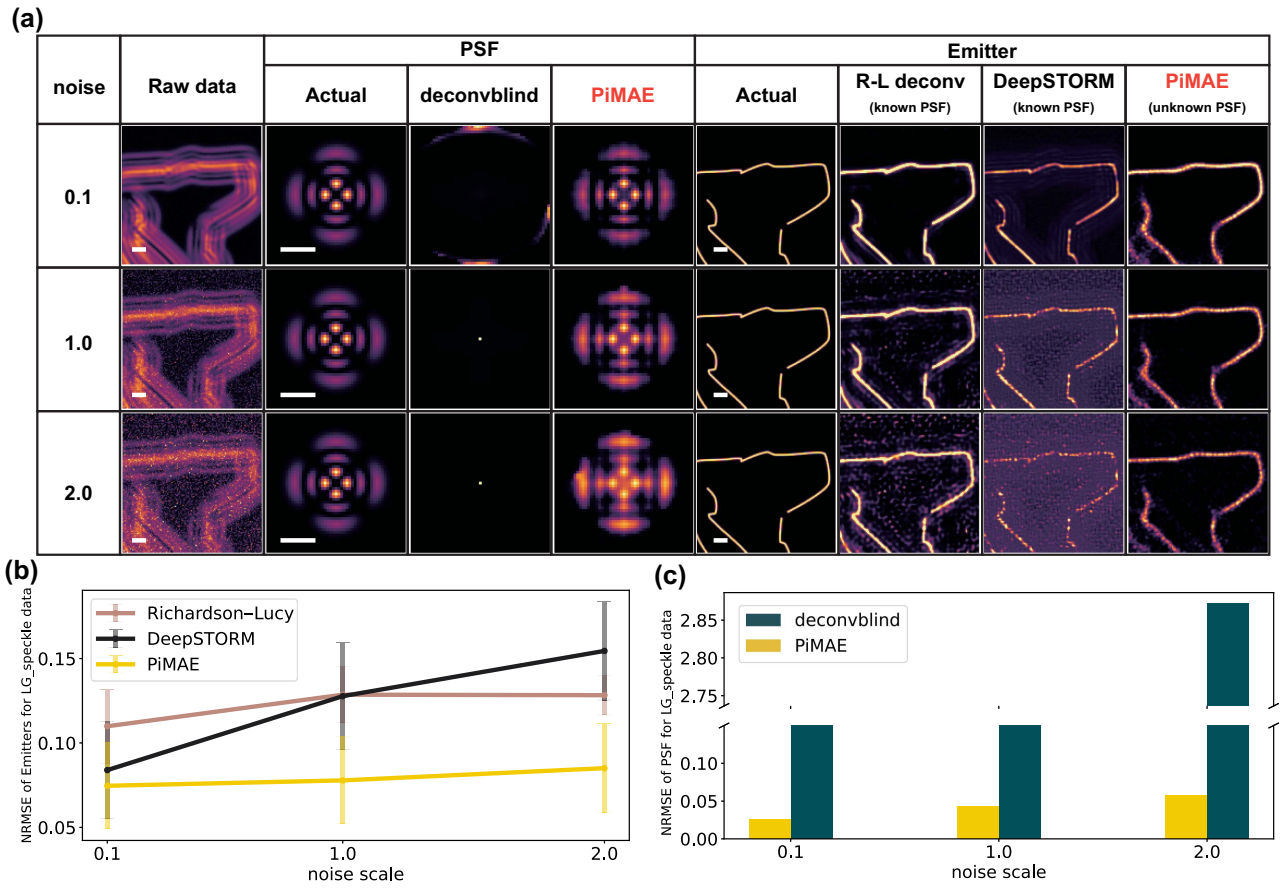

Fig. 18. Evaluation of speckle noise robustness. (a) The estimated PSF and emitters result from synthetic data with speckle noise. The scale bar is 0.5 μm. (b) NRMSE of estimated emitters from synthetic data with speckle noise; (c) NRMSE of estimated PSF from synthetic data with speckle noise.

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3}, \tag{F5}$$

where $C_1 = (K_1 L)^2$, $C_2 = (K_2 L)^2$, and $C_3 = C_2/2$; here $L = 255$, $C_1 = C_2 = 0$, $K_1 = 0.01$, and $K_2 = 0.03$. The sliding window size is 11.

In the assessment, we use the max-min normalization method to process each image as follows:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{F6}$$

where x_{norm} is the normalized image, x is the raw image, $x_{\min}$ is the minimum value in the image, and $x_{\max}$ is the maximum value in the image.

APPENDIX G: DECONVBLIND

The Deconvblind is one of the most popular methods for blind deconvolution, which iteratively updates the PSF and the estimated image. For each task, we used the training set consisting of 1000 images and applied the Deconvblind function in MATLAB [53] to estimate the PSF. These 1000 images were provided to Deconvblind in the form of a stack.

We demonstrate that the Deconvblind approach leads to a trivial solution, i.e., a δ function, for estimating the PSF. We evaluate the performance of Deconvblind and PiMAE on 1000 synthetic images generated from the Sketches data set, where the PSF is generated from a wide-field microscope in focus. As shown in Fig. 16, the PSF estimated by Deconvblind converges to a δ function, which is a trivial solution and results in the estimated emitter image being equal to the raw image. In contrast, the PiMAE-estimated PSF steadily approaches the actual PSF as the number of training epochs increases.

APPENDIX H: DEEPSTORM

We compare the performance of PiMAE with other deep-learning-based methods, such as DeepSTORM, DECODE, and those that train neural networks for predicting emitter locations using supervised learning. As a baseline for comparison, we reproduce the DeepSTORM method. The original DeepSTORM model is a fully convolutional neural network (FCN), which we upgrade to the U-net architecture [7,54,55], a powerful deep-learning architecture that has shown superior performance in various computer vision tasks (see Fig. 17). While incorporating this change, we ensure to adhere to the original DeepSTORM model's design and use the sum of MSE and L1 loss as the loss function.

During the training process, we use 1000 images containing randomly positioned emitters simulated using the ImageJ [56] ThunderSTORM [57] plugin. These images are convolved with the PSF of the task, normalized using the mean and averaged standard deviation, and then noise with an intensity of 10^{-5} is added to enhance robustness.

APPENDIX I: EVALUATION RESULTS OF ADDING SPECKLE NOISE TO SYNTHETIC DATA

Speckle noise is a type of granular noise texture that can degrade image quality in coherent imaging systems such as medical ultrasound, optical coherence tomography, as well as radar and synthetic aperture radar (SAR) systems. It is a multiplicative noise that is proportional to the image intensity. The probability density function of speckle noise can be described by an exponential distribution,

$$p(z) = \frac{1}{\sigma^2}\ \exp\left(-\frac{z}{\sigma^2}\right). \tag{I1}$$

Here, z represents the intensity, and σ^2 represents the speckle noise variance. To evaluate the impact of speckle noise on estimating PSF and emitters, we use LG_{22} as the PSF and Sketches as the emitters. We construct three sets of data with noise variances of 0.1, 1, and 2, respectively, each containing 1000 training images and 100 test images. We use the NRMSE metric to evaluate the results, as shown in Fig. 18.

APPENDIX J: THE RESULTS USING MS-SSIM AS THE METRIC

1. Results of Out-of-Focus Synthetic Data

In this section, we present the results of synthetic data with varying out-of-focus distances, assessed using the MS-SSIM metric. Gaussian noise with a standard deviation of $\text{noise}_{\text{std}}/\text{raw}_{\text{mean}} = 0.5$ is added to each synthetic data set. The results are displayed in Fig. 19.

2. Results of SPADE Synthetic Data

We present the results for synthetic data evaluated using the MS-SSIM metric for HG and LG modes. Gaussian noise with $\text{noise}_{\text{std}}/\text{raw}_{\text{mean}} = 0.5$ is added to each synthetic data set. The results are displayed in Fig. 20.

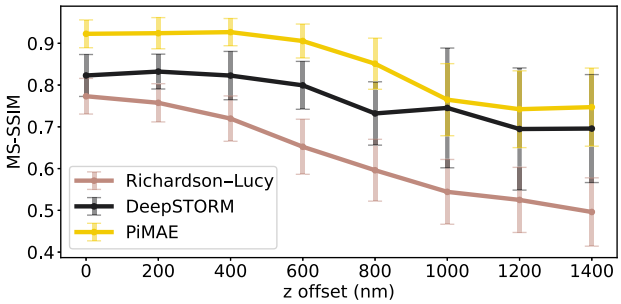

Fig. 19. MS-SSIM of the results of estimated emitters from out-of-focus synthetic data.

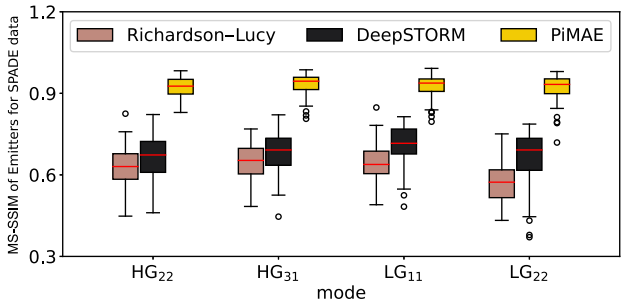

Fig. 20. MS-SSIM of the results of estimated emitters from the SPADE Sketches data set.

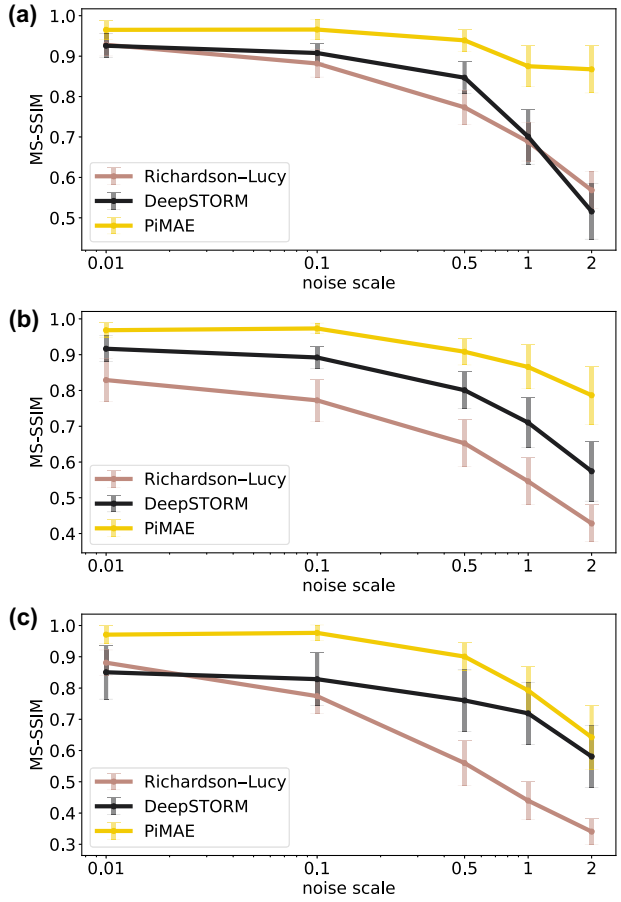

Fig. 21. Noise robustness. (a) MS-SSIM of the results of estimated emitters from the in-focus Sketches data set; (b) MS-SSIM of the results of estimated emitters from the 600 nm out-of-focus Sketches data set; (c) MS-SSIM of the results of estimated emitters from the LG_{22} mode Sketches data set.

3. Results of Noise Robustness

We present the results of synthetic data with different levels of noise measured using MS-SSIM as the metric. For each synthetic data set, Gaussian noise is added with levels of 0.01, 0.1, 0.5, 1, and 2, respectively. The results are depicted in Fig. 21.

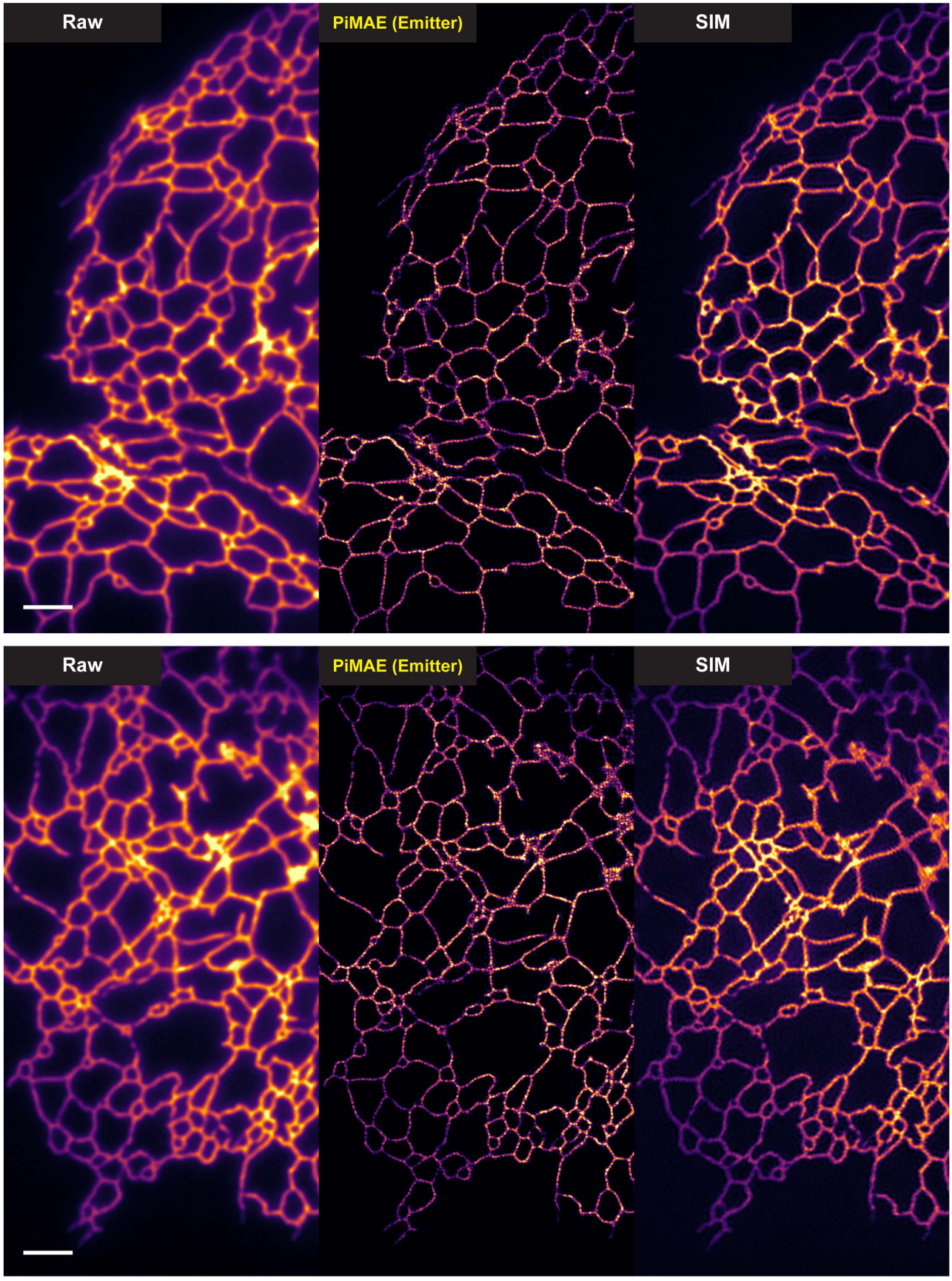

Fig. 22. Comparison of ER imaging results. Here we show some comparative results of wide-field microscopy, SIM, and PiMAE-resolved wide-field microscopy. The length of the scale bar is 2.50 μm. Data from BioSR data set [40].

APPENDIX K: RESULTS OF REAL-WORLD EXPERIMENTS

We evaluate PiMAE in two real-world experiments. First, we utilize the imaging results of ER structures obtained from both wide-field microscopy and SIM from the BioSR data set [40]. Second, we construct a custom-built wide-field microscope to image NV color centers in diamond. The ability of PiMAE to handle non-Gaussian PSFs is evaluated in both out-of-focus and aberrations scenarios.

1. Results of ER

Figure 22 shows the results of wide-field microscopy, SIM, and PiMAE-resolved wide-field microscopy of ER. Figure 23

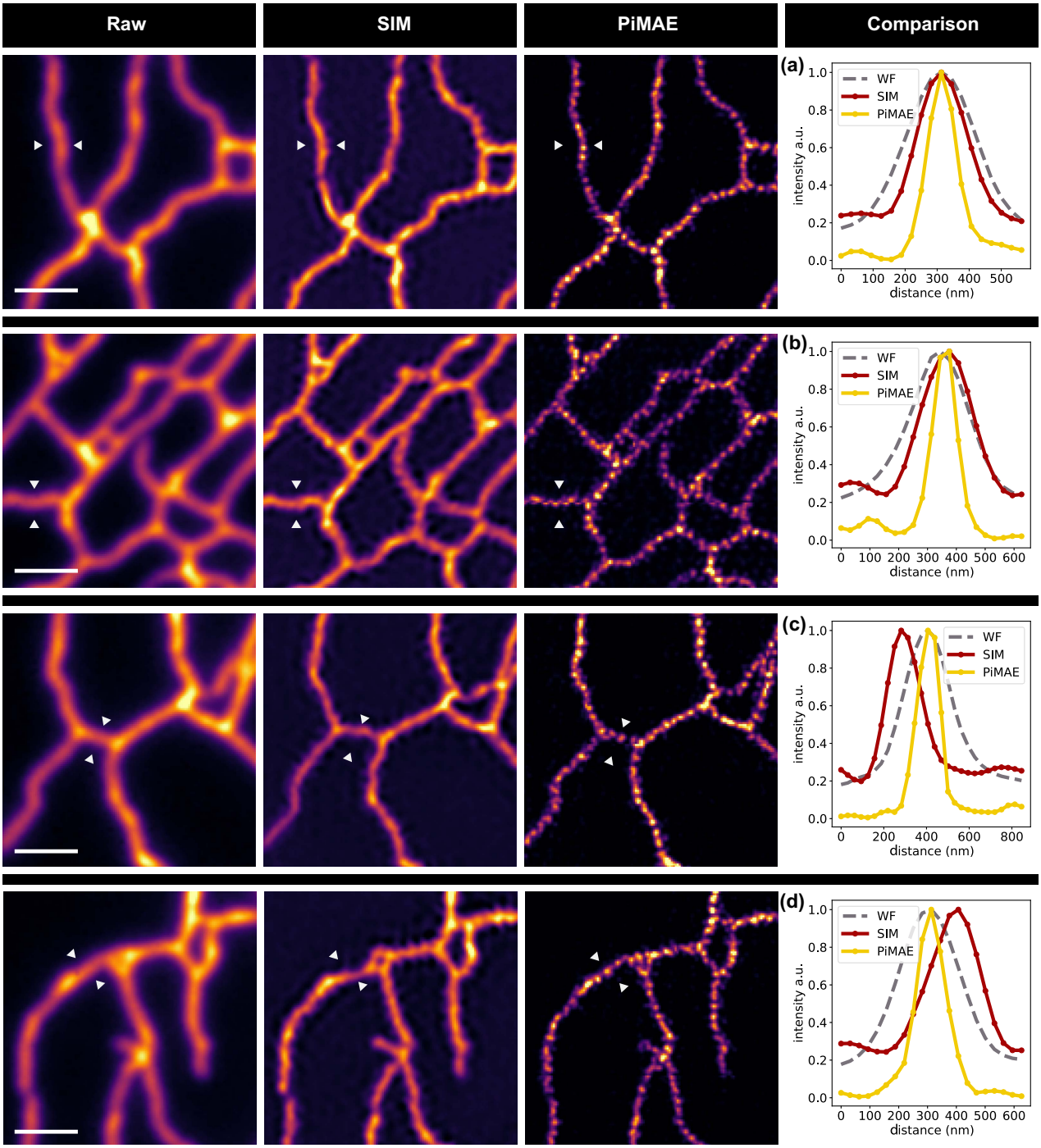

Fig. 23. Artifacts in superresolution images reconstructed using SIM. Reconstruction artifacts are a common issue in SIM-reconstructed images, as evidenced in (c) and (d), due to factors such as nonuniform fringe patterns or phase errors in the reconstruction process. In comparison, the PiMAE-estimated emitters do not exhibit these artifact problems. The scale bar is 1.00 μm.

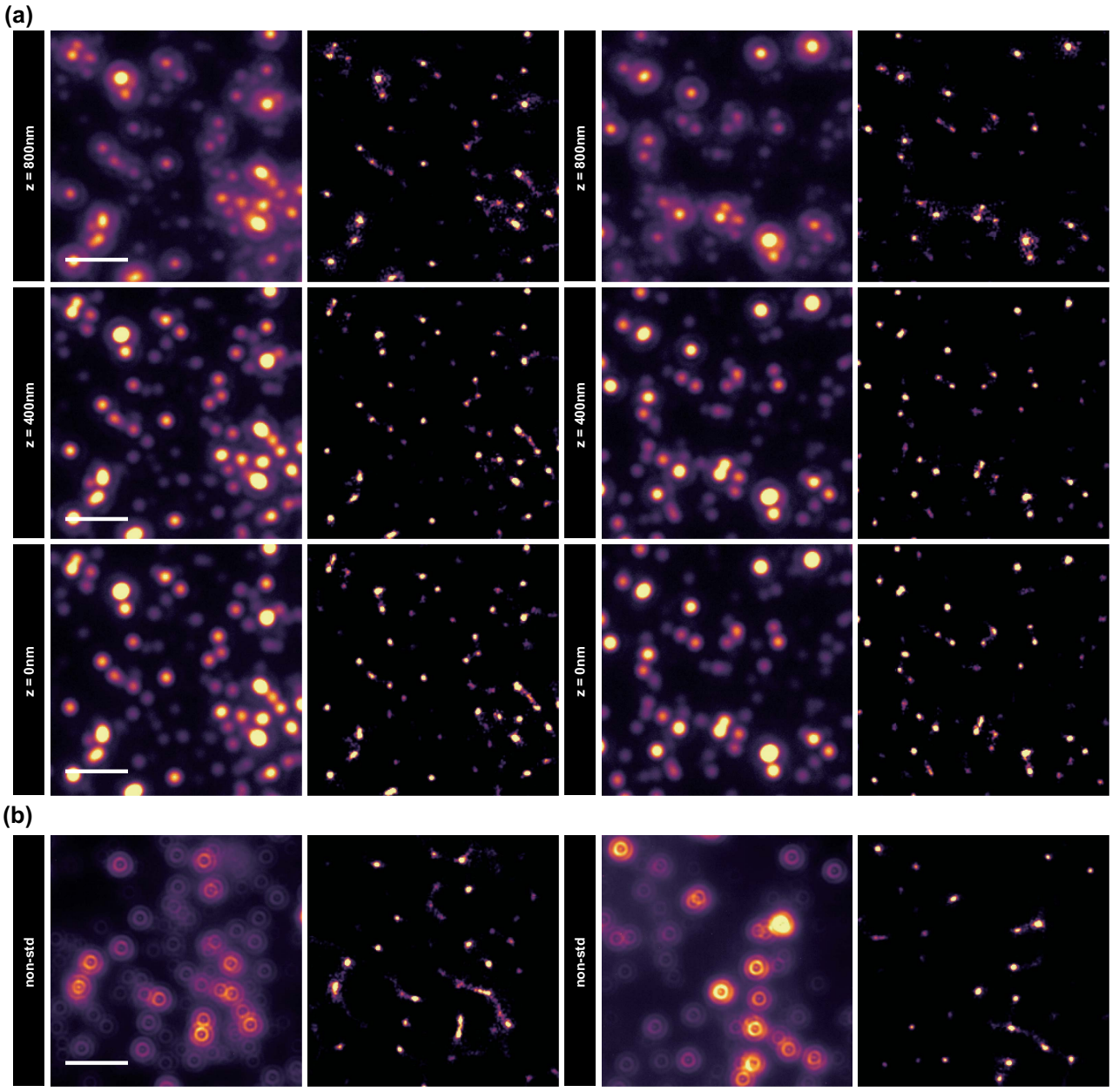

Fig. 24. Wide-field microscopy imaging of NV color centers. (a) Comparison of wide-field microscopy results and PiMAE estimated emitters results at different out-of-focus distances, with invariant field of view from top to bottom, and different field of view on the left and right, respectively; the scale bar is 2.50 μm. (b) Wide-field microscopy results and PiMAE estimated emitters of nonstandard PSF when the objective is mismatched to the coverslip; the scale bar is 6.40 μm.

demonstrates that PiMAE is capable of avoiding the artifact phenomenon seen in SIM.

2. Results of NV Center Imaging

The results of out-of-focus and aberrated wide-field microscopy imaging of NV color centers, as well as PiMAE-resolved results, are shown in Fig. 24. The aberrations were generated as follows: an objective lens with a phase aberration correction ring was first used to image a 50 nm nanodiamond on the opposite side of a coverslip with a thickness of 0.11–0.23 mm and a refractive index of 1.5. The correction ring of the objective lens was then rotated to match a coverslip with a thickness of 0.1 mm and a refractive index of 1.5, and this lens was used to observe nanodiamonds spin-coated on the opposite side of sapphire with a thickness of 0.15 mm and a refractive index of 1.72, thus artificially creating an aberration and resulting in a doughnut-shaped PSF. (Note: Olympus UPLXAPO40X objective lens was used.)

APPENDIX L: SUMMARY OF RESULTS

In this section, we summarize the results of all the synthetic tasks in Table 1.

Table 1. Summary of Synthetic Data Experiments[a]

Synthetic Data								
Task Info				NRMSE for Emitters			NRMSE for PSF	
Task	PSF	Emitters	Noise	PiMAE	DeepSTORM	RL	PiMAE	DB
1	1400 nm	Sketches	0.5	0.090	0.111	0.257	0.070	0.195
2	1200 nm	Sketches	0.5	0.090	0.106	0.238	0.075	0.144
3	1000 nm	Sketches	0.5	0.093	0.110	0.232	0.083	0.098
4	800 nm	Sketches	0.5	0.080	0.103	0.201	0.029	0.062
5	600 nm	Sketches	0.5	0.073	0.092	0.163	0.018	0.059
6	400 nm	Sketches	0.5	0.074	0.081	0.140	0.018	0.051
7	200 nm	Sketches	0.5	0.072	0.078	0.130	0.023	0.048
8	0 nm	Sketches	0.5	0.071	0.084	0.124	0.022	0.045
9	0 nm	Sketches	2	0.089	0.139	0.198	0.045	0.078
10	0 nm	Sketches	1	0.085	0.105	0.156	0.031	0.064
11	0 nm	Sketches	0.5	0.071	0.079	0.124	0.022	0.045
12	0 nm	Sketches	0.1	0.068	0.066	0.091	0.021	0.042
13	0 nm	Sketches	0.01	0.068	0.065	0.082	0.021	0.165
14	600 nm	Sketches	2	0.095	0.144	0.231	0.019	0.076
15	600 nm	Sketches	1	0.091	0.111	0.185	0.016	0.070
16	600 nm	Sketches	0.5	0.073	0.092	0.163	0.018	0.058
17	600 nm	Sketches	0.1	0.066	0.073	0.142	0.023	0.030
18	600 nm	Sketches	0.01	0.068	0.070	0.135	0.023	0.937
19	HG_{22}	Sketches	0.5	0.075	0.098	0.151	0.028	0.156
20	HG_{31}	Sketches	0.5	0.072	0.097	0.147	0.029	0.161
21	LG_{11}	Sketches	0.5	0.072	0.098	0.154	0.016	0.088
22	LG_{22}	Sketches	0.5	0.073	0.094	0.179	0.042	0.042
23	LG_{22}	Sketches	2	0.100	0.128	0.307	0.069	0.105
24	LG_{22}	Sketches	1	0.078	0.104	0.235	0.048	0.100
25	LG_{22}	Sketches	0.5	0.063	0.094	0.179	0.029	0.098
26	LG_{22}	Sketches	0.1	0.056	0.082	0.117	0.017	0.095
27	LG_{22}	Sketches	0.01	0.061	0.080	0.105	0.022	2.761
28	LG_{22}	Lines/$n = 10$	0.01	0.040	0.049	0.153	0.028	0.352
29	LG_{22}	Lines/$n = 20$	0.01	0.058	0.074	0.193	0.037	0.156
30	LG_{22}	Lines/$n = 50$	0.01	0.096	0.119	0.213	0.059	0.102
31	LG_{22}	Lines/$n = 100$	0.01	0.158	0.171	0.216	0.130	0.103
32	LG_{22}	Sketches/speckle noise	2	0.085	0.155	0.128	0.026	0.309
33	LG_{22}	Sketches/speckle noise	1	0.078	0.128	0.129	0.043	0.896
34	LG_{22}	Sketches/speckle noise	0.1	0.075	0.084	0.110	0.057	2.871
35	USTC	Sketches	0.01	0.086	0.114	0.160	0.135	0.187

[a]The training set consists of 1000 images, and the test set consists of 100 images.

Funding. Innovation Program for Quantum Science and Technology (2021ZD0303200); CAS Project for Young Scientists in Basic Research (YSBR-049); National Natural Science Foundation of China (62225506); Anhui Provincial Key Research and Development Plan (2022b13020006); USTC Center for Micro and Nanoscale Research and Fabrication.

Acknowledgment. The authors would like to thank Drs. Yu Zheng, Yang Dong, Ce Feng, and Shao-Chun Zhang for fruitful discussions.

Disclosures. The authors declare no conflicts of interest.

Data Availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

REFERENCES

1. S.-H. Lee, J. Y. Shin, and A. Lee, *et al.*, "Counting single photoactivatable fluorescent molecules by photoactivated localization microscopy (PALM)," Proc. Natl. Acad. Sci. USA **109**, 17436–17441 (2012).
2. M. J. Rust, M. Bates, and X. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)," Nat. Methods **3**, 793–796 (2006).
3. K. K. Bearne, Y. Zhou, and B. Braverman, *et al.*, "Confocal super-resolution microscopy based on a spatial mode sorter," Opt. Express **29**, 11784–11792 (2021).
4. S. W. Hell and J. Wichmann, "Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy," Opt. Lett. **19**, 780–782 (1994).
5. X. Chen, C. Zou, and Z. Gong, *et al.*, "Subdiffraction optical manipulation of the charge state of nitrogen vacancy center in diamond," Light Sci. Appl. **4**, e230 (2015).
6. E. Nehme, L. E. Weiss, and T. Michaeli, *et al.*, "Deep-STORM: super-resolution single-molecule microscopy by deep learning," Optica **5**, 458–464 (2018).

7. A. Speiser, L.-R. Müller, and P. Hoess, *et al.*, "Deep learning enables fast and dense single-molecule localization with high accuracy," Nat. Methods **18**, 1082–1090 (2021).
8. D. S. Biggs and M. Andrews, "Acceleration of iterative image restoration algorithms," Appl. Opt. **36**, 1766–1775 (1997).
9. T. F. Chan and C.-K. Wong, "Total variation blind deconvolution," IEEE Trans. Med. Imaging **7**, 370–375 (1998).
10. D. Krishnan, T. Tay, and R. Fergus, "Blind deconvolution using a normalized sparsity measure," in *Conference on Computer Vision and Pattern Recognition (CVPR)* (2011), pp. 233–240.
11. G. Liu, S. Chang, and Y. Ma, "Blind image deblurring using spectral properties of convolution operators," IEEE Trans. Med. Imaging **23**, 5047–5056 (2014).
12. T. Michaeli and M. Irani, "Blind deblurring using internal patch recurrence," in *European Conference on Computer Vision* (2014), pp. 783–798.
13. J. Pan, D. Sun, and H. Pfister, *et al.*, "Deblurring images via dark channel prior," IEEE Trans. Pattern Anal. Mach. Intell. **40**, 2315–2328 (2017).
14. J. Pan, Z. Hu, and Z. Su, *et al.*, "L_0-regularized intensity and gradient prior for deblurring text images and beyond," IEEE Trans. Pattern Anal. Mach. Intell. **39**, 342–355 (2016).
15. W. Ren, X. Cao, and J. Pan, *et al.*, "Image deblurring via enhanced low-rank prior," IEEE Trans. Med. Imaging **25**, 3426–3437 (2016).
16. L. Sun, S. Cho, and J. Wang, *et al.*, "Edge-based blur kernel estimation using patch priors," in *IEEE International Conference on Computational Photography (ICCP)* (2013), pp. 1–8.
17. Y. Yan, W. Ren, and Y. Guo, *et al.*, "Image deblurring via extreme channels prior," in *IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 4003–4011.
18. W. Zuo, D. Ren, and D. Zhang, *et al.*, "Learning iteration-wise generalized shrinkage–thresholding operators for blind deconvolution," IEEE Trans. Med. Imaging **25**, 1751–1764 (2016).
19. A. Shajkofci and M. Liebling, "Spatially-variant CNN-based point spread function estimation for blind deconvolution and depth estimation in optical microscopy," IEEE Trans. Med. Imaging **29**, 5848–5861 (2020).
20. L. B. Lucy, "An iterative technique for the rectification of observed distributions," Astrophys. J. **79**, 745 (1974).
21. A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," arXiv, arXiv:1807.03748 (2018).
22. Z. Wu, Y. Xiong, and S. X. Yu, *et al.*, "Unsupervised feature learning via non-parametric instance discrimination," in *IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 3733–3742.
23. K. He, H. Fan, and Y. Wu, *et al.*, "Momentum contrast for unsupervised visual representation learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 9729–9738.
24. T. Chen, S. Kornblith, and M. Norouzi, *et al.*, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (PMLR)* (2020), pp. 1597–1607.
25. C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *IEEE International Conference on Computer Vision* (2015), pp. 1422–1430.
26. A. Dosovitskiy, J. T. Springenberg, and M. Riedmiller, *et al.*, "Discriminative unsupervised feature learning with convolutional neural networks," arXiv, arXiv:1406.6909 (2014).
27. J. Devlin, M.-W. Chang, and K. Lee, *et al.*, "BERT: pre-training of deep bidirectional transformers for language understanding," arXiv, arXiv:1810.04805 (2018).
28. T. Chen, S. Liu, and S. Chang, *et al.*, "Adversarial robustness: from self-supervised pre-training to fine-tuning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 699–708.
29. X. Chen, W. Chen, and T. Chen, *et al.*, "Self-PU: self boosted and calibrated positive-unlabeled training," in *International Conference on Machine Learning (PMLR)* (2020), pp. 1510–1519.
30. M. Chen, A. Radford, and R. Child, *et al.*, "Generative pretraining from pixels," in *International Conference on Machine Learning (PMLR)* (2020), pp. 1691–1703.
31. O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *International Conference on Machine Learning (PMLR)* (2020), pp. 4182–4192.
32. D. Pathak, P. Krahenbuhl, and J. Donahue, *et al.*, "Context encoders: feature learning by inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 2536–2544.
33. T. H. Trinh, M.-T. Luong, and Q. V. Le, "Selfie: self-supervised pretraining for image embedding," arXiv, arXiv:1906.02940 (2019).
34. K. He, X. Chen, and S. Xie, *et al.*, "Masked autoencoders are scalable vision learners," arXiv, arXiv:2111.06377 (2021).
35. A. Dosovitskiy, L. Beyer, and A. Kolesnikov, *et al.*, "An image is worth 16×16 words: transformers for image recognition at scale," arXiv, arXiv:2010.11929 (2020).
36. D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 9446–9454.
37. T.-Y. Lin, M. Maire, and S. Belongie, *et al.*, "Microsoft COCO: common objects in context," in *European Conference on Computer Vision* (2014), pp. 740–755.
38. L. Liu, H. Jiang, and P. He, *et al.*, "On the variance of the adaptive learning rate and beyond," in *8th International Conference on Learning Representations (ICLR)* (2020), pp. 1–13.
39. M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" ACM Trans. Graph. **31**, 44 (2012).
40. C. Qiao, D. Li, and Y. Guo, *et al.*, "Evaluation and development of deep neural networks for image super-resolution in optical microscopy," Nat. Methods **18**, 194–202 (2021).
41. A. Makandar, D. Mulimani, and M. Jevoor, "Comparative study of different noise models and effective filtering techniques," Int. J. Sci. Res. **3**, 458–464 (2013).
42. M. Tsang, R. Nair, and X.-M. Lu, "Quantum theory of superresolution for two incoherent optical point sources," Phys. Rev. X **6**, 031033 (2016).
43. K. Y. Han, K. I. Willig, and E. Rittweger, *et al.*, "Three-dimensional stimulated emission depletion microscopy of nitrogen-vacancy centers in diamond using continuous-wave light," Nano Lett. **9**, 3323–3329 (2009).
44. X.-D. Chen, E.-H. Wang, and L.-K. Shan, *et al.*, "Focusing the electromagnetic field to $10^{-6}\lambda$ for ultra-high enhancement of field-matter interaction," Nat. Commun. **12**, 6389 (2021).
45. C. L. Degen, F. Reinhard, and P. Cappellaro, "Quantum sensing," Rev. Mod. Phys. **89**, 035002 (2017).
46. Z.-H. Wang, "PiMAE," 2022, https://opticapublishing.figshare.com/s/f2a426e03ecac8ff8417.
47. Y. Zhang, D. Zhou, and S. Chen, *et al.*, "Single-image crowd counting via multi-column convolutional neural network," in *IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 589–597.
48. T. Xiao, P. Dollar, and M. Singh, *et al.*, "Early convolutions help transformers see better," arXiv, arXiv:2106.14881 (2021).
49. H. Zhao, O. Gallo, and I. Frosio, *et al.*, "Loss functions for image restoration with neural networks," IEEE Trans. Image Process. **3**, 47–57 (2016).
50. B. Zhang, J. Zerubia, and J.-C. Olivo-Marin, "Gaussian approximations of fluorescence microscope point-spread function models," Appl. Opt. **46**, 1819–1829 (2007).
51. M. W. Beijersbergen, L. Allen, and H. Van der Veen, *et al.*, "Astigmatic laser mode converters and transfer of orbital angular momentum," Opt. Commun. **96**, 123–132 (1993).
52. Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *37th Asilomar Conference on Signals, Systems & Computers* (2003), Vol. **2**, pp. 1398–1402.
53. https://www.mathworks.com/products/matlab.html.
54. O. Ronneberger, P. Fischer, and T. Brox, "U-NET: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention* (2015), pp. 234–241.
55. S. K. Gaire, E. Flowerday, and J. Frederick, *et al.*, "Deep learning-based spectroscopic single-molecule localization microscopy for simultaneous multicolor imaging," in *Computational Optical Sensing and Imaging* (2022), paper CTu5F-4.
56. T. J. Collins, "Imagej for microscopy," Biotechniques **43**, S25–S30 (2007).
57. M. Ovesný, P. Křžek, and J. Borkovec, *et al.*, "ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging," Bioinformatics **30**, 2389–2390 (2014).