Huang *et al. PhotoniX* 2023, **4**(1):2
https://doi.org/10.1186/s43074-022-00077-x

PhotoniX

RESEARCH Open Access

Enhancing image resolution of confocal fluorescence microscopy with deep learning

Boyi Huang[1†], Jia Li[1†], Bowen Yao[1], Zhigang Yang[1], Edmund Y. Lam[2], Jia Zhang[1*], Wei Yan[1*] and Junle Qu[1*]

[†]Boyi Huang and Jia Li contributed equally to this work.

*Correspondence: julyzhang2021@163.com; weiyan@szu.edu.cn; jlqu@szu.edu.cn

[1] Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen 518060, China
[2] Department of Electrical and Electronic Engineering, University of Hong Kong, Pokfulam, Hong Kong, SAR, China

Abstract

Super-resolution optical imaging is crucial to the study of cellular processes. Current super-resolution fluorescence microscopy is restricted by the need of special fluorophores or sophisticated optical systems, or long acquisition and computational times. In this work, we present a deep-learning-based super-resolution technique of confocal microscopy. We devise a two-channel attention network (TCAN), which takes advantage of both spatial representations and frequency contents to learn a more precise mapping from low-resolution images to high-resolution ones. This scheme is robust against changes in the pixel size and the imaging setup, enabling the optimal model to generalize to different fluorescence microscopy modalities unseen in the training set. Our algorithm is validated on diverse biological structures and dual-color confocal images of actin-microtubules, improving the resolution from ~ 230 nm to ~ 110 nm. Last but not least, we demonstrate live-cell super-resolution imaging by revealing the detailed structures and dynamic instability of microtubules.

Keywords: Super-resolution fluorescence microscopy, Image resolution enhancement, Deep learning, Generative adversarial network

Introduction

Since super-resolution fluorescence microscopy can resolve biological structures at the nanometer scale, effectively overcoming the limitation in resolution capacity of diffraction-limited optical microscopy, it has been considered a powerful technique to allow us to probe into many fundamental processes of life, such as the inner workings of cells and the ultrastructures of organelle dynamics [1–4].

Among a variety of super-resolution methods, stimulated emission depletion (STED) microscopy is the most promising due to its high spatial resolution and temporal resolution [5, 6]. STED microscopy employs a second laser beam (STED beam) in the confocal setup to rapidly deplete the excited state of fluorophores back to their ground state, which requires the stimulated emission to win the competition with spontaneous fluorescence emission within a few nanoseconds [7, 8]. The STED beam has a doughnut-shaped focal intensity distribution with zero intensity at the center, and its overlap with the excitation laser beam results in a smaller residual fluorescence spot, thereby sharpening the point spread function (PSF) and improving resolution. Nevertheless, the strong power of the depletion laser is prone to cause photodamage in the biological

samples and photobleaching of the fluorophores, which prevents the wide adoption of STED for practical long-term live-cell imaging.

To address this issue, in addition to developing adequately photostable and live-cell compatible highly fluorescent labelling reagent, researchers also explore computational algorithms that can transform a captured low-resolution image into a high-resolution one without the need for directly applying super-resolution fluorescence microscopy to live-cell imaging. There has been research presenting deep-learning-based algorithms where they build a generative adversarial network (GAN) or deep Fourier channel attention network (DFCAN) to achieve super-resolution and cross-modality image transformations [9, 10]. The networks do not require modeling of the image-formation process or manually tuning of the parameters. Although these algorithms can enhance the resolution of diffraction-limited low-resolution images to match those obtained by super-resolution microscopy, they are not applicable to confocal images of microtubules and microfilaments, particularly with live-cell imaging, and the resolution of their network output images needs to be further improved. To overcome this limitation, we propose an efficient resolution enhancement algorithm based on deep learning. We note that automatic feature extraction is a remarkable advantage that deep learning has over conventional machine learning algorithms, and deep learning has more complex ways of connecting layers together with a larger amount of computing power than previous networks [11, 12]. All these advances have kindled a lot of interest in this approach. Recent applications of deep learning to image processing have been implemented successfully in a variety of research fields [13–18].

In our approach, we achieve super-resolution by building a two-channel attention network (TCAN) architecture and training the network to learn representations of information in both the spatial domain and the frequency domain. This enables the network to precisely map the diffraction-limited input images into super-resolved ones. TCAN framework requires neither special instrumentation nor special fluorophores, and does not constrain the pixel sizes or imaging modalities of the input images. Even if they are different from those in the training data, our network is still capable of super-resolving the low-resolution input images. This also promotes the application of our TCAN model to various fields of view (FOV) of input images. More importantly, we use this model, trained with only the static images, to achieve a long-term live-cell imaging, capturing the dynamic microtubules with finer structures and a higher resolution. This circumvents the need to acquire long time-lapse STED images of microtubules dynamics, which suffers from photobleaching/phototoxicity and remains challenging. Compared with STED, we demonstrate a superior algorithmic performance by inferring super-resolution images of diverse biological structures in terms of higher signal-to-noise ratio (SNR) and better image quality. We also improve the resolution of dual-color confocal images of microtubules and filaments, and their relative positions and crosstalk are better revealed by our method.

Methods

In theory, deep learning can be considered as using algorithms for acquiring structural descriptions from training data examples. A model is built to contain the structural information extracted from the training data, and then those structures or the

model can be employed to predict unknown data. In reality, we use low-resolution (confocal)—high-resolution (STED) image pairs of the same view as the training data examples, and build TCAN model to learn the mapping from low-resolution image to its corresponding high-resolution image by capturing the feature representations of these training data.

TCAN model

Inspired by U-net [19] and deep Fourier channel attention network (DFCAN) [10], we construct the TCAN architecture based on the conditional generative adversarial network (cGAN) framework, as depicted in Fig. 1. It can be divided into two parts, a generative model and a discriminative model. The confocal image is firstly fed into the generative model that generates a high-resolution image. This generated high-resolution image, together with the STED image, is then fed into the discriminative model that compares these two images and estimates the probability of the generated high-resolution image being the STED image. The above process is repeated in the training stage till the discriminative model cannot distinguish the generated high-resolution image from the STED image. The generative model finds optimal parameters and is forced to efficiently generate high-resolution image similar to the STED training image. In other words, the generative model achieves the modeling of resolution enhancement in a way of deep learning (i.e., convolution and other operations). The discriminative model plays a role in evaluating whether the generated high-resolution image and the ground truth are as close as possible. The training enables TCAN model to learn the ability of mapping such that it can directly infer high-resolution image when applied to new low-resolution image.

Generative model

The generator in our TCAN is composed of U-net and DFCAN, enabling the network to learn representations of information in both the spatial domain and the Fourier domain. The former is proposed to learn to suppress irrelevant regions while highlighting salient structures of varying shapes and sizes, yielding improved prediction performance across diverse datasets [19]. The latter focuses on learning hierarchical representations of high-frequency information and more precise mappings form low-resolution images to high-resolution images. Figure 1 illustrates the structure of the generator used in this work. The input image is firstly fed into a convolutional block, and then the outputs of U-net and DFCAN are summed and go through another convolutional block to form the network output. Both convolutional blocks perform the following operation:

$$x_o = \mathrm{LReLU}[\mathrm{Conv}(\mathrm{x_i})], \tag{1}$$

in which the output and input of the convolutional block are represented with x_o and x_i, respectively. Conv() is the convolution operation, and LReLU[] is the leaky rectified linear unit activation function [20] with a slope of $\alpha = 0.1$, defined as

$$\mathrm{LReLU}(\mathrm{x};\alpha) = \max(0,\mathrm{x}) - \alpha \times \max(0,-\mathrm{x}). \tag{2}$$

The architecture of U-net used in this work is illustrated in Fig. 2, which consists of four downsampling blocks and four upsampling blocks, and they are connected. Let d_k

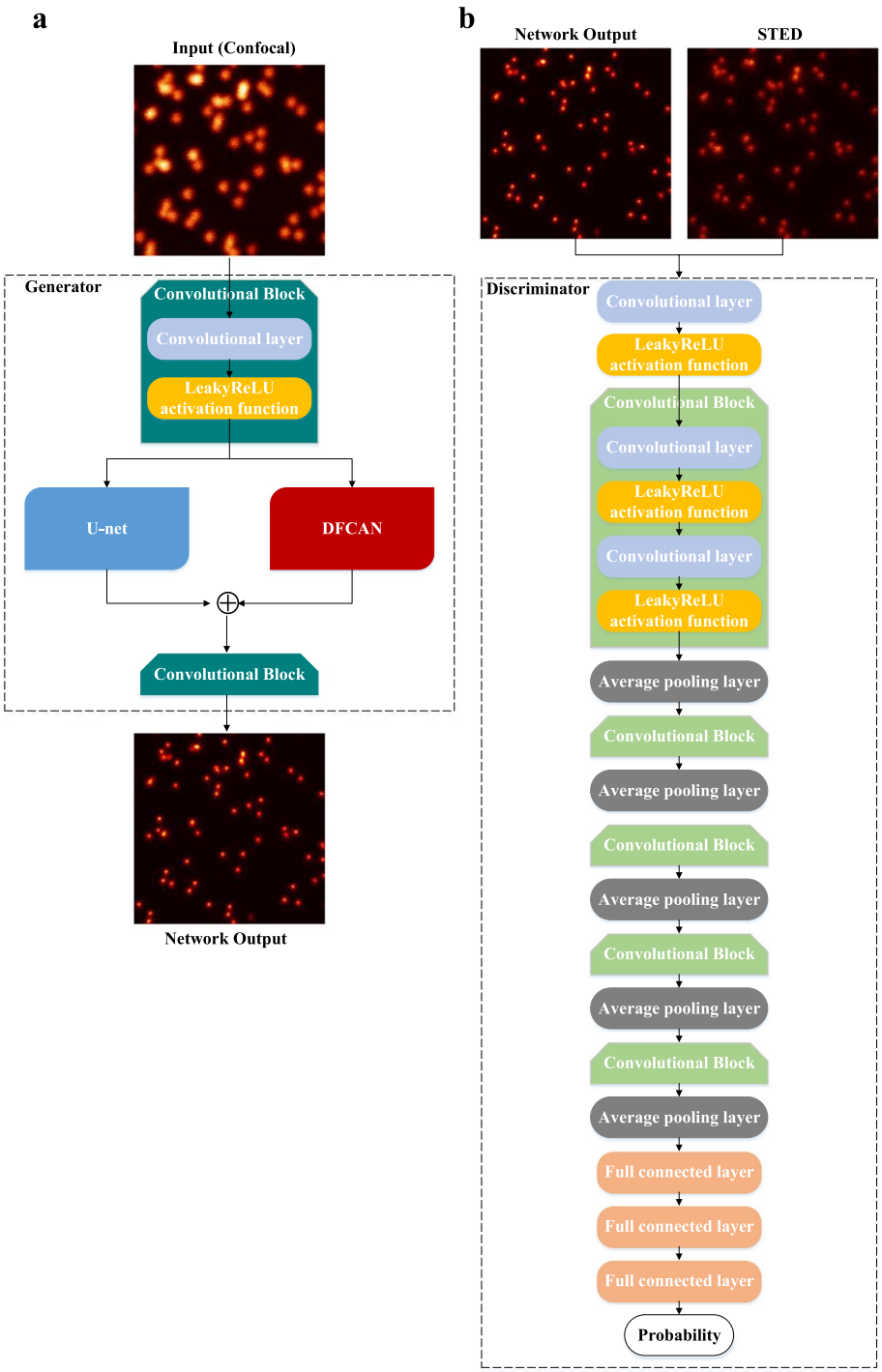

Fig. 1 The network architecture of TCAN. (**a**) The architecture of the generator of TCAN. (**b**) The architecture of the discriminator of TCAN

be the output of the kth downsampling block, and d_0 be the low-resolution input image. Each downsampling block includes three residual convolutional blocks, within which it performs

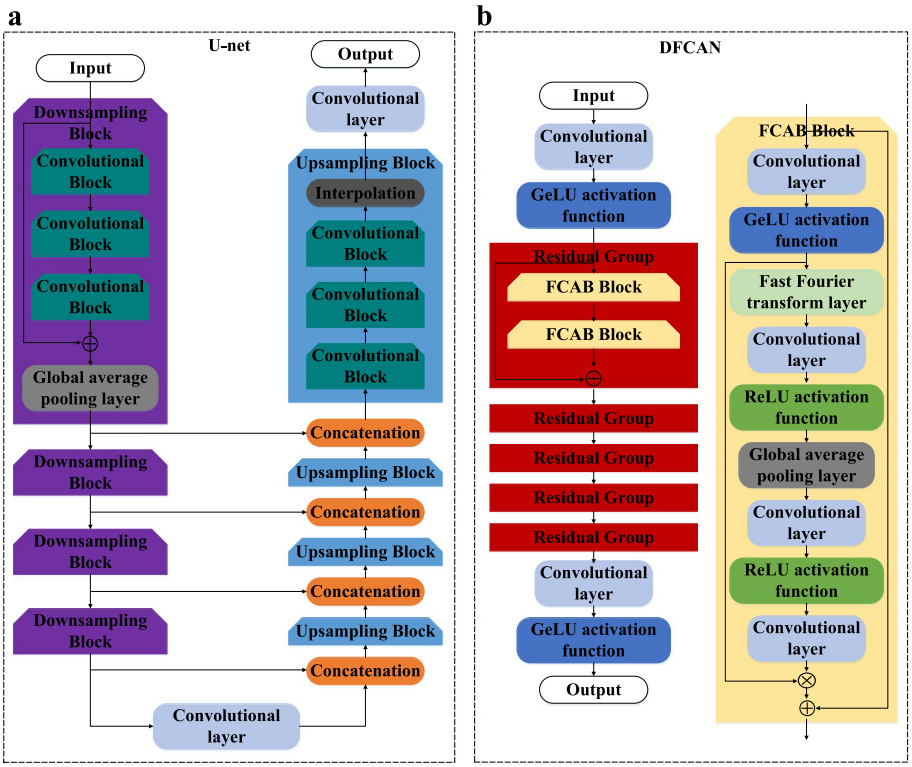

Fig. 2 The architecture of the generator of TCAN. (**a**) The architecture of U-net. (**b**) The architecture of DFCAN

$$d_k = d_{k-1} + \mathrm{LReLU}\left[\mathrm{Conv}\left(\mathrm{LReLU}\left[\mathrm{Conv}\left(\mathrm{LReLU}\left[\mathrm{Conv}\left(\mathrm{d_{k-1}}\right)\right]\right)\right]\right)\right]. \quad k = 1, 2, 3, 4 \tag{3}$$

Because the number of channels of d_{k-1} changes after passing through the convolutional blocks, the input of each downsampling block is zero-padded to ensure its direct addition to the result of three consecutive convolutional blocks. A global average pooling layer is inserted after the summation to achieve spatial downsampling.

Each upsampling block is also composed of three convolutional blocks, and we can derive its output as

$$u_k = \mathrm{LReLU}\left[\mathrm{Conv}\left(\mathrm{LReLU}\left[\mathrm{Conv}\left(\mathrm{LReLU}\left[\mathrm{Conv}\left(\mathrm{Concat}\left\{\mathrm{d_{5-k}}, \mathrm{u_{k-1}}\right\}\right)\right]\right)\right]\right], \quad k = 1, 2, 3, 4 \tag{4}$$

where u_k represents the output of the kth upsampling block and u_0 is the output of the convolution layer that lies at the bottom of this U-shape network. The downsampling block output and the upsampling block input is concatenated by Concat{,} which can strengthen feature propagation and improve efficiency [21]. A nearest neighbor interpolation is added in the upsampling block to achieve spatial upsampling. The last convolutional layer maps the 32 channels into 1 channel that corresponds to a monochrome grayscale high-resolution image.

We employ DFCAN in the generative model to enhance the learning ability of our model in the frequency domain, and its architecture is displayed in Fig. 2. A convolution layer is firstly used to generate the feature maps, and then a Gaussian error linear

unit (GELU) [22] is added for nonlinearity. The GELU activation function is formulated as

$$\mathrm{GELU[x]} = 0.5\mathrm{x}\left[1 + \mathrm{erf}\left(\frac{\mathrm{x}}{\sqrt{2}}\right)\right], \tag{5}$$

in which erf() denotes the error function, defined by

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\mathrm{dt}. \tag{6}$$

The output of GELU is fed into a residual group (RG), and five identical RGs are successively used in our DFCAN model. Each of them is composed of two Fourier channel attention blocks (FCAB) and a skip connection. We therefore have

$$\mathrm{RG(y)} = \mathrm{y} + \mathrm{FCAB}\big(\mathrm{FCAB(y)}\big), \tag{7}$$

where y represents the input feature maps of the RG. As in Ref. [10], the feature maps in each FCAB are rescaled in a channel-wise manner as:

$$\mathrm{FCAB(y)} = \mathrm{y} + \mathrm{y_o} \otimes \mathrm{Conv}\big(\mathrm{ReLU}\big[\mathrm{Conv}\big(\mathrm{ReLU}\big[\mathrm{Conv}\big(\mathrm{abs}\big\{\mathrm{FFT}\big(\mathrm{y_o}\big)\big\}^{\gamma}\big)\big]\big)\big]\big), \tag{8}$$

in which

$$y_o = \mathrm{GELU}\big[\mathrm{Conv}\big(\mathrm{GELU[Conv(y)]}\big)\big]. \tag{9}$$

In eq. (8), we use FFT() to represent the fast Fourier transform and put γ in the exponent to increase the contributions of the high-frequency components. The operator abs{} computes the absolute value, and ReLU[] denotes the rectified linear units (ReLU) [23]. Mathematically, $\mathrm{ReLU}[\cdot] = \max[\cdot, 0]$. A global average pooling layer is inserted between ReLU and the subsequent convolutional layer for spatial downsampling. The last RG is followed by a convolutional layer activated by the GELU activation function. The nearest neighbor interpolation is used to upsample the image to the same size as the ground truth to accommodate the inferred high-frequency information [10].

Discriminative model

Figure 1 describes the structure of our discriminator. It is a simple convolutional neural network (CNN) architecture that begins with a convolutional layer. Five convolutional blocks then follow, which are different from the convolutional blocks in the generative model, given by

$$\begin{aligned} &z_k = \mathrm{LReLU}\big[\mathrm{Conv}\big(\mathrm{LReLU}\big[\mathrm{Conv}\big(\mathrm{z_{k-1}}\big)\big]\big)\big], \\ &k = 1, 2, 3, 4, 5 \end{aligned} \tag{10}$$

where z_k denotes the output of the kth convolutional block, and z_0 is the input of the first convolutional block. We insert average pooling successive convolutional blocks to reduce the dimension, and it performs the downsampling operation by taking the spatial average of the feature maps in the corresponding 2×2 region while discarding redundant information. After that, there are 3 fully connected (FC) layers and the discriminator outputs the estimated probability. It is not necessary to add a sigmoid activation

function in the discriminative model, since it has been incorporated into the loss function code of BCEWithLogitsLoss() used in our work.

Loss function

We design the loss function of the generative model as a combination of MSE, binary cross-entropy (BCE) and the structural similarity (SSIM) index [24]. MSE loss ensures prediction accuracy by penalizing the difference between the network output and ground truth. BCE loss recovers the minute detail from the blurred images, and SSIM loss enhances the perceptual quality fidelity of the output. This leads to the following loss function

$$\mathcal{L}_{G|D}(X,Y) = \mathrm{MSE}(\mathrm{G(X)},\mathrm{Y}) + \mathrm{SSIM}(\mathrm{G(X)},\mathrm{Y}) + \mathrm{BCE}(\mathrm{D(G(X))} - \mathrm{D(Y)},\mathrm{Y_{label}}), \tag{11}$$

in which X and Y are input low-resolution image and high-resolution image used as ground truth, respectively. $G()$ is the generative model output, and $D()$ is the discriminative model prediction. Y_{label} is set as 1 in the process of training the generator.

The loss function of the discriminative model calculates the binary cross-entropy, i.e.

$$\mathcal{L}_{D|G}(X,Y) = \frac{1}{2}\{\mathrm{BCE}(\mathrm{D(G(X))} - \mathrm{D(Y)},\mathrm{Y_{label}}) + \mathrm{BCE}(\mathrm{D(Y)} - \mathrm{D(G(X))},1 - \mathrm{Y_{label}})\}, \tag{12}$$

when Y_{label} is set as 0 in the process of training the discriminator. Specific loss functions are given in Supporting Information.

Training

For each type of specimen and each imaging modality, we capture a total of ~80 groups of confocal (512 × 512 pixels) and STED (512 × 512 pixels) images. To prevent the model from being overfitting, we select ~60 groups of original data and perform random cropping, rotation transformation and horizontal/vertical flipping to further enrich the training dataset, which eventually generate ~3000 pairs of confocal images (256 × 256 pixels) and STED images (256 × 256 pixels). For the testing dataset, we select the remaining ~20 groups of original data to augment the dataset. Wide-field and SIM training data pairs are generated from BioSR dataset in Ref. [10], which is a high-quality image dataset covering four biology structures with nine signal levels and two upscaling-factors. We use 3000 pairs of linear low-and-high resolution images of the microtubules as training data, and their resolution is ~100 nm. The detail information of image acquisition is described in Supporting Information.

In order to accelerate the training speed and ensure the training efficiency, our patch size is set as 256 × 256, with a batch size of 2 due to the limitation of hardware. Note that TCAN works by alternating between training the generative model given the discriminative model, and updating the discriminator by keeping the generator unchanged. Both the generative model and the discriminative model are randomly initialized and optimized using the adaptive moment estimation (Adam) optimizer [25], with a starting learning rate of 0.0001 and 0.00005, respectively. This framework is implemented with Pytorch [26] framework version 1.7.1 and Python version 3.6.4 in the Microsoft Windows 10 operating system. The training is performed on a consumer-grade laptop

(Alienware-51r, Dell) equipped with a GeForce RTX2080 graphic card (NVIDIA) and a Core i9-9900K CPU @ 3.6 GHz (Intel). Our model is firstly trained with nano-beads, which takes ~12 h. After the transfer learning [27], the final models trained for cell nuclei and microtubules take ~24 h and ~26 h, respectively. A typical plot of the validation loss values during TCAN training is shown in Additional file 1 Fig. S1. In the competition process between the generator and the discriminator, the network gradually refines the learnt super-resolution image transformation and obtains better spatial details. We take the trained model at 200 epochs as the final testing model, which is sufficient for different images in our experiments. The iteration time is dependent to the patch and batch size.

Results and discussion

Resolution enhancement in confocal microscopy images of nano-beads and nucleus

We begin with evaluating the performance of our proposed TCAN model using 23 nm fluorescent beads. The nano-bead samples are imaged on a Leica TCS SP8 STED confocal microscope, and 1000 pairs of confocal-STED image patches with a size of 256 × 256 pixels are used as training data. Our network takes the confocal image in Fig. 3a as input, which is unseen by the network in the training stage, and outputs a super-resolved image in Fig. 3b. The result of the network is compared with the image (Fig. 3c) acquired using STED microscopy. It can be seen that some of the nano-beads in our samples are too close to be discerned in the raw confocal microscopy image and STED image, while our method is capable of reducing artifacts and blur and resolving these closely spaced nano-beads, as presented in Figs. 3d-f. This is also consistent with the intensity profiles (Fig. 3m) along the white dashed lines in Figs. 3d-f.

We further assess the impact of the proposed TCAN by two image-based criteria: one is image resolution, measured by the full width at half maximum (FWHM) of the PSF, and the other one is image quality, estimated by the signal-to-noise ratio. There are 20 isolated nano-beads selected randomly for the PSF measurement in the images of the confocal microscope and STED microscope, as well as the network output image. The attained FWHM of the confocal microscope PSF is 239 ± 25 nm, roughly corresponding to the lateral resolution of a diffraction-limited imaging system at an emission wavelength of 664 nm and numerical aperture of 1.4. The PSF distribution of the network output is even better than that of the STED system, with a FWHM of 58 ± 1 nm versus 83 ± 9 nm, respectively. Since our method also establishes a data-driven image transformation, similar to that discussed in Ref. [9], the learned PSF does not require any prior information on modeling of the image formation process or its parameters.

Next, we verify the practicality of the proposed TCAN by applying it to fixed HeLa cell nucleus. Figures 3g-i displays the input confocal microscopy image, the network output result and the STED image of the same field of view, respectively. We observe that our method succeeds in transforming a low-resolution confocal image into a super-resolution image. As exemplified by the magnified images of the green boxes in Figs. 3j-l, TCAN resolves the densely labeled nuclear pore complexes (NPCs) [28] better than STED image and reduces the background noise, reaching a compromise between retaining useful information and denoising. The rationale behind this

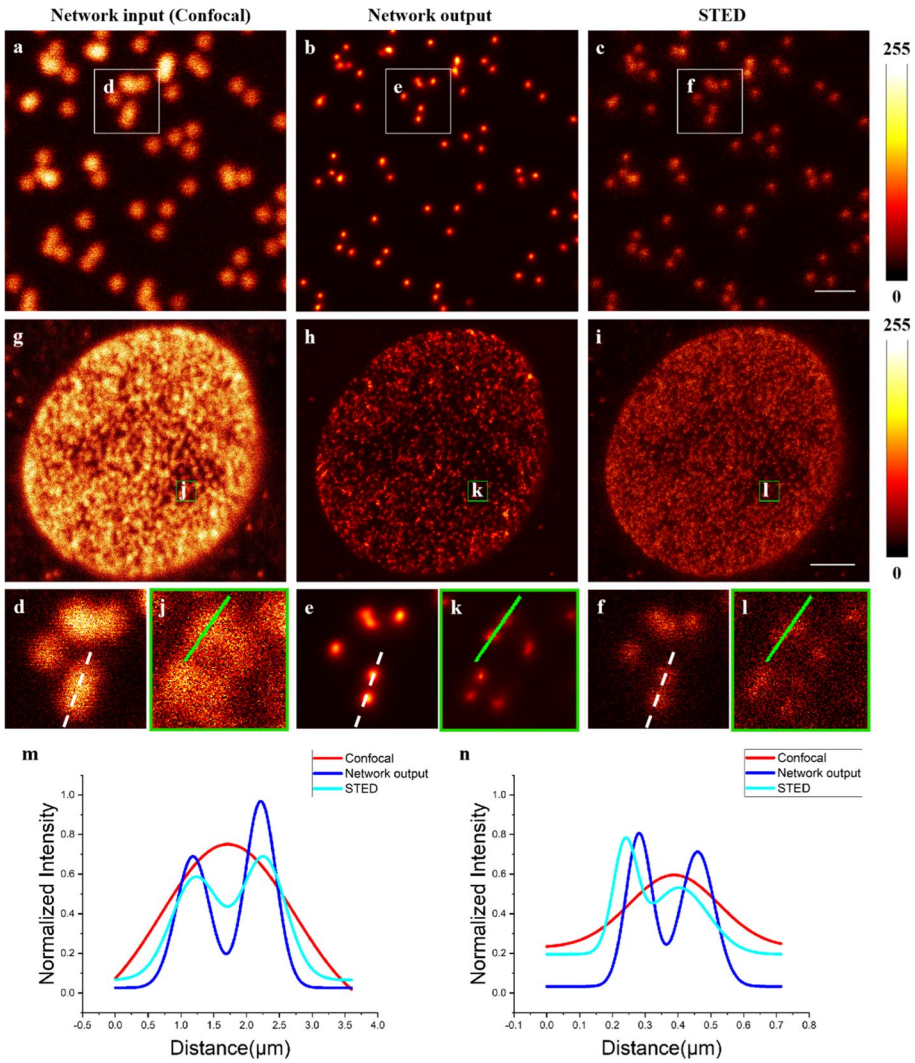

Fig. 3 Super-resolution imaging of the nano-beads and nucleus using TCAN. (**a**) A diffraction-limited confocal microscopy image of fluorescent nano-beads is used as input to the network. (**b**) The super-resolved output image. (**c**) The STED image of the same field of view. (**d**, **j**) Examples of closely spaced nano-beads or NPCs that cannot be resolved by confocal microscopy. (**e**, **k**) The trained network takes (**d**) and (**j**) as input and resolves the individual nano-beads or NPCs. (**f**, **l**) The STED microscopy image. (**g**) A diffraction-limited confocal microscopy image of HeLa cell nucleus is used as input to the network. (**h**) The super-resolved output image. (**i**) The STED image of the same field of view. (**m**, **n**) Intensity profiles along the white dashed lines and green solidlines in different images. Experiments are repeated with 20 images, achieving similar results. Scale bars, 4 μm

result is that the generator in our model benefits from both U-net and DFCAN, which simultaneously learns precise representation of the spatial structures and high-frequency information.

To verify the improvement of our network on image quality, we compare the SNR of the network output to the network input (confocal image), the STED images and the deconvolution of the STED image. SNR is calculated according to the following formula in Ref. [9],

$$\mathrm{SNR} = \left| \frac{\mathrm{s} - \overline{\mathrm{b}}}{\sigma_{\mathrm{b}}} \right|, \tag{13}$$

where s is the mean peak value of the signal calculated from a Gaussian fit to the particles, and $\overline{b}$ is the mean value of the background (e.g. randomly selected regions which do not contain any objects), and σ_b is the standard deviation of the background. The results listed in Table 1 demonstrate that our proposed method can suppress noise and improve the image quality by different types of samples.

Resolution enhancement in confocal microscopy images of microtubules

In case the confocal-STED training image pairs are not available, our network model trained with images captured by different imaging modalities is still able to infer super-resolution image. We employ 3000 pairs of wide-field and structured illumination microscopy (SIM) patches with a size of 256×256 pixels as training data, and apply the present framework to microtubules, a more complex structure. The results are compared against the STED image and deconvolution of the STED image, and the deconvolution is performed by using Huygens Software. Our TCAN model, as expected, reveals noticeably improved resolution in comparison with the input confocal images (Fig. 4a). It is worth noting that the resolution of the network output images (Fig. 4b) is indeed improved, especially that the regions of dense and complex microtubule structures are better resolved and appear sharper, compared with STED images in Fig. 4c, as exhibited by the magnified results of the green boxes. There are artifacts and noise between adjacent microtubules in the STED microscope images. For the comparison to the deconvolution of the STED images in Fig. 4d, it can be seen that there are obvious broken structures, and the discontinuity is more severe for sparsely distributed microtubules. Here we also employ transfer learning, which uses a learned network trained with nanobeads as the initial model, to speed up the training process for nucleus, microtubules and actin.

To quantitatively evaluate the overall performance of our method, we use three metrics, i.e., SNR, mean square error (MSE), and resolution, to measure the quality of the output super-resolved image. MSE numerically computes the pixel-level data fidelity by calculating the difference between the resulting image and the ground truth. Image resolution is performed by means of decorrelation analysis, which describes the highest frequency from the local maxima of the decorrelation functions instead of the theoretical resolution [29]. These results are illustrated in Figs. 4e-g, where generally larger SNR and smaller MSE of the network output indicate that the conventional STED images and the deconvolution of the STED image are inferior to our inference images. The measured resolution of input confocal image, network output,

Table 1 Quantification of SNR improvement

Types	Network input (Confocal)	Network output	STED
Nano-beads	6.0	13.2	14.1
Nucleus	3.6	12.0	8.2
Microtubules	9.6	10.9	10.4

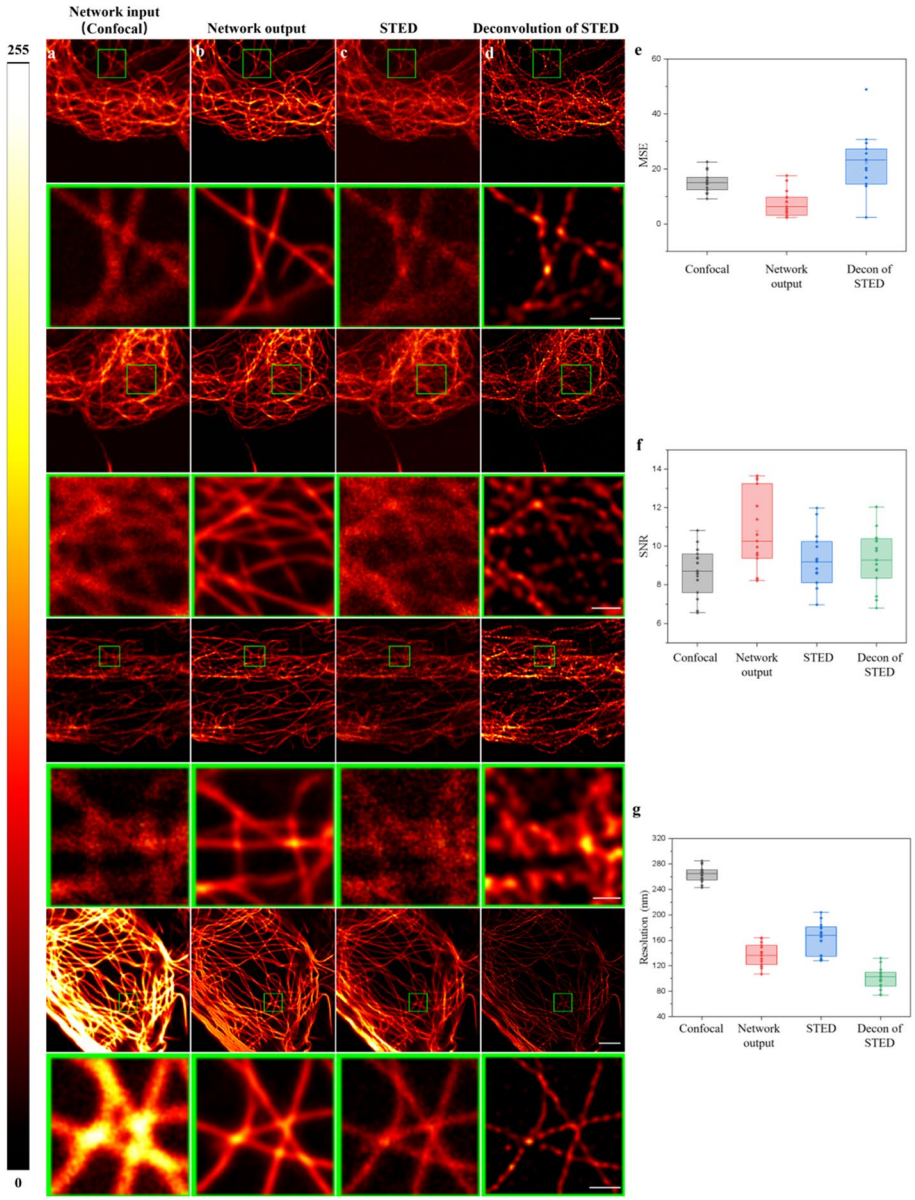

Fig. 4 Super-resolution imaging of the microtubules using TCAN. (**a**) The first column of images is diffraction-limited confocal microscopy images of the microtubules used as input to the network. (**b**) The second column of images is the super-resolved output images. (**c**) The third column of images is the STED images of the same field of view. (**d**) The fourth column of images is the deconvolution of the STED images of the same field of view. The green box regions are shown below at a magnified scale. (**e-g**) Statistical comparison of confocal images, output images, STED images and deconvolution of STED images in terms of MSE, SNR and resolution. Tukey box-and-whisker plots with outliers displayed as diamonds are shown. Experiments are repeated with 15 images, achieving similar results. Scale bars in the second row, the fourth row, the sixth row, the seventh row and the eighth row are 0.75 μm, 0.75 μm, 0.5 μm, 3 μm and 0.5 μm, respectively

STED and the deconvolution of STED image in the last row of Fig. 4 are 267 ± 12 nm, 136 ± 16 nm, 163 ± 23 nm and 101 ± 17 nm, respectively. The deconvolution of STED image achieves a higher resolution at the expense of obvious unstructured regions and even losing structural information.

Figures 4e-g are plotted in Tukey box-and-whisker format. The box extends from the 25th and 75th percentiles and the line in the middle of the box indicates the median. To define whiskers and outliers, the inter-quartile distance (IQR) is firstly calculated as the difference between the 25th and 75th percentiles. The upper whisker represents the larger value between the largest data point and the 75th percentile plus 1.5 times the IQR; the lower whisker represents the smaller value between the smallest data point and the 25th percentile minus 1.5 times the IQR. Data points larger than the upper whisker or smaller than the lower whisker are identified as outliers, which are displayed as black diamonds.

For deep learning methods, the training data determines what we want the neural network to learn. To achieve the best results, the imaging modality for the training data should in principle be precisely matched to that of the input images. However, we find that the image quality rather than the imaging modality of the training data is a critical factor affecting the image inference performance. This can be observed from Additional file 1 Fig. S2 in Supporting Information. Even though the input images and STED images are captured with the same imaging platform, the output images of the network trained by using deconvolution of the STED images are worse than the results of the network trained with high-quality SIM images. This is related to the fact that the input and output of the framework share a high degree of mutual information, and the quality of the information in the training examples has an effect on the pixel-to-pixel transformation and the resolution enhancement learned by the network. For the task of translating one possible representation of a scene into another, it is broadly referred to as image-to-image translation problem [30]. They share common process of predicting pixels from pixels, and the network architecture used for our training, i.e., conditional GANs [31] have been proven to be effective in learning such mapping. In this problem the input and output are renderings of the same underlying structures, and the training process can be viewed as utilizing this mutual information between the input and label images to restrict the network output. Accordingly, the network pays attention to the quality of structures in the training examples more than the imaging platform of the training data.

Additionally, if the pixel size is large, one microtubule distributes across fewer pixels; otherwise, more pixels are required to show the same structure. Hence the pixel size is another important parameter affecting the feature representation to be learned by the network and the ability of the network to distinguish adjacent microtubules as separate objects. For instance, direct application of a network that is trained with images with a pixel size of 50 nm would produce acceptable biological structures only when the input images have pixel size of 35 nm–70 nm. Therefore, if the pixel size of the input images and training images are different, we upsample/downsample the input images to match that of the training image pairs. After the upsampling/downsampling, the neural network successfully suppresses the artifacts and further improves the resolution of the confocal microscopy images. In Fig. 5, compared the network output images in the third column to the network output images in the fourth column, it is important to note that the effect of the pixel sizes can be compensated by upsampling/downsampling the input images to match the pixel sizes to that of the training data, thereby improving the quality of the inference images. Since the pixel size of our training data is 50 nm, we upsampe the pixel size of 75 nm of the input confocal images in the first row, while downsample

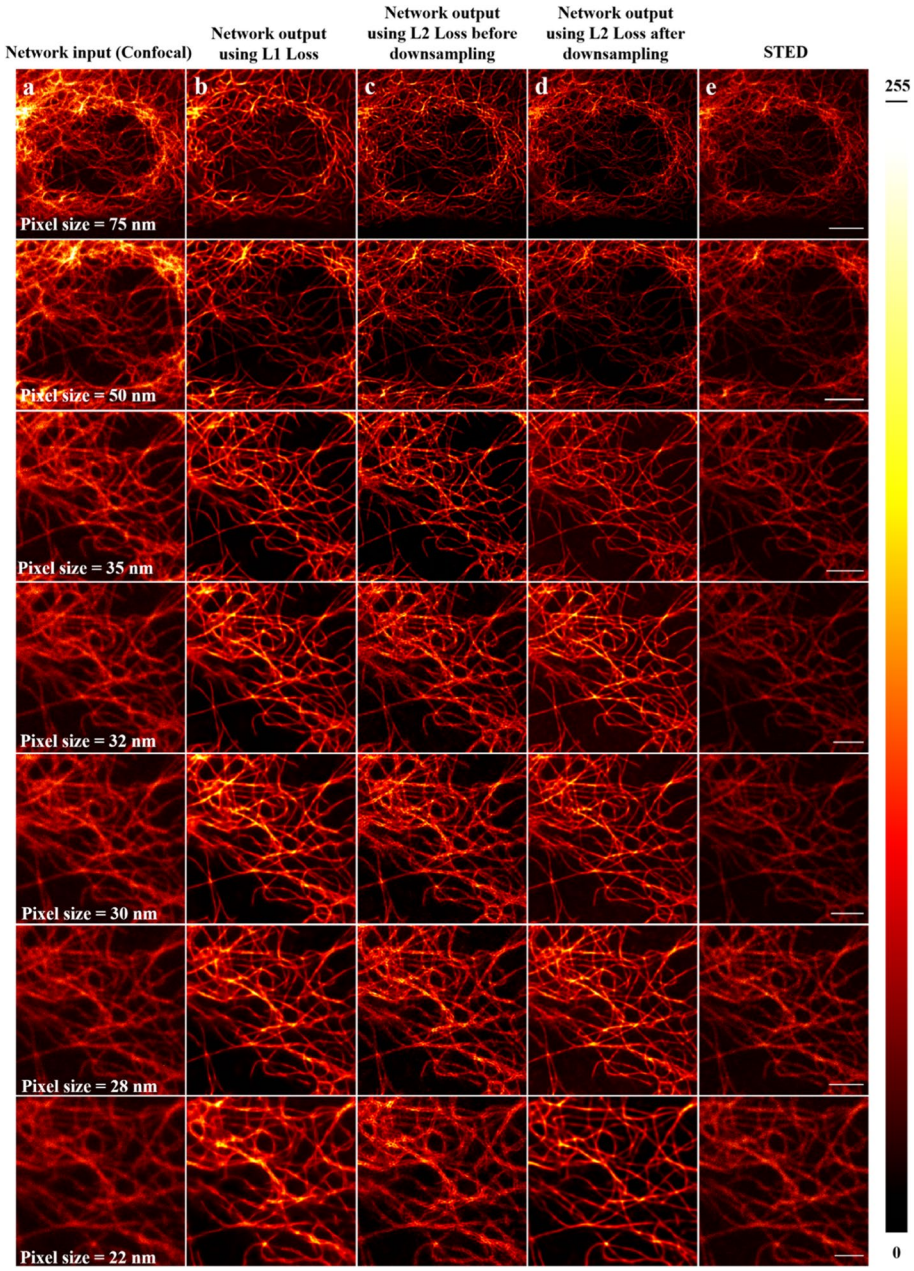

Fig. 5 (**a**) Diffraction-limited confocal microscopy images with different pixel sizes are used as input to the network. (**b**) The Super-resolved network inference images using L1 loss before upsampling/downsampling the input images to match the pixel size of the training data. (**c**) The Super-resolved network inference images using L2 loss before upsampling/downsampling the input images to match the pixel size of the training data. (**d**) The super-resolved network inference images using L2 loss after upsampling/downsampling the input images to match the pixel sizes of the training data. (**e**) The STED images of the same field of view. Scale bars in (**e**) are 8 μm, 6 μm, 4 μm, 3 μm, 3 μm, 3 μm and 2 μm, respectively from the first row to the seventh row

other pixel sizes of the input images in the third row to the seventh row. In addition, compared the network output images in the second column to the network output images in the third column, we notice that the model trained by L1 loss is more robust against the variations of pixel sizes than the model trained by L2 loss, although the latter

can obtain better inference images when the pixel size of the input images and the training data is the same (50 nm in our experiments). The result is related to the fact that L2 loss is more sensitive to outliers and gets stuck more easily in a local minimum [32, 33].

This also facilitates the application of the TCAN model to a large field of view of the confocal images. Figure 6 displays the results of applying our method to super-resolve confocal images of 45.88 μm × 45.88 μm (1024 × 1024 pixels) and 56.17 μm × 56.17 μm (2048 × 2048 pixels), respectively, revealing finer features of the microtubules. The above results demonstrate that the proposed framework is able to achieve favorable performance for various fields of view of input images.

When the input image is captured with a new experimental setup, our TCAN network model does not need to be trained again. We apply the network model trained with wide-field and SIM image pairs to directly super-resolve the images of microtubules captured with the Nikon A1R MP+ microscope. The confocal microscopy images are transformed into resolution-enhanced images, as shown in Fig. 7, exhibiting more sharp details of the microtubules. To provide further demonstration of the network's generalization, two large confocal image patches of 184.32 μm × 184.32 μm (3072 × 3072 pixels) and 71.68 μm × 71.68 μm (1024 × 1024 pixels), also acquired by the Nikon A1R MP+ microscope, are used as input, and Additional file 1 Fig. S3 in Supporting Information illustrates the advantage of the GAN-based super-resolution approach with upsampling/downsampling. It is possible to extend applications of our TCAN model to super-resolve low-resolution images captured with different imaging systems.

The generalization of our TCAN model includes improving resolution of images acquired with new imaging systems and improving image resolution on new types of samples that are not present in the training phase. As manifested in Fig. 7 and Additional file 1 Fig. S3, resolution enhancement of confocal images captured with the Nikon the A1R MP+ microscope are achieved by our network model trained with

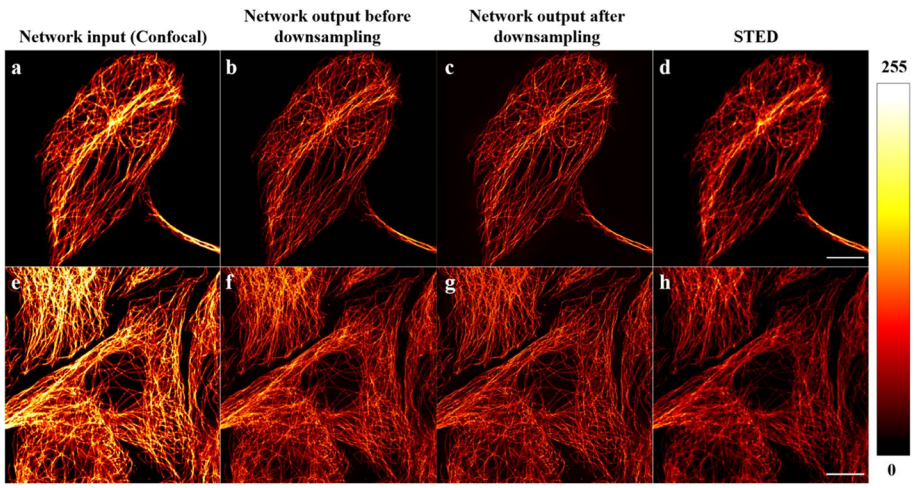

Fig. 6 Super-resolution imaging of large FOV of confocal images acquired on a Leica TCS SP8 STED confocal microscope using TCAN. (**a**, **e**) Diffraction-limited confocal microscopy images of the microtubules are used as input to the network. (**b**, **f**) The super-resolved network inference images before downsampling the input images to match the pixel size of the training data. (**c**, **g**) The super-resolved network inference images after downsampling the input images to match the pixel sizes of the training data. (**d**, **h**) The STED images of the same field of view. Scale bar in (**d**) is 8 μm, and scale bar in (**h**) is 10 μm

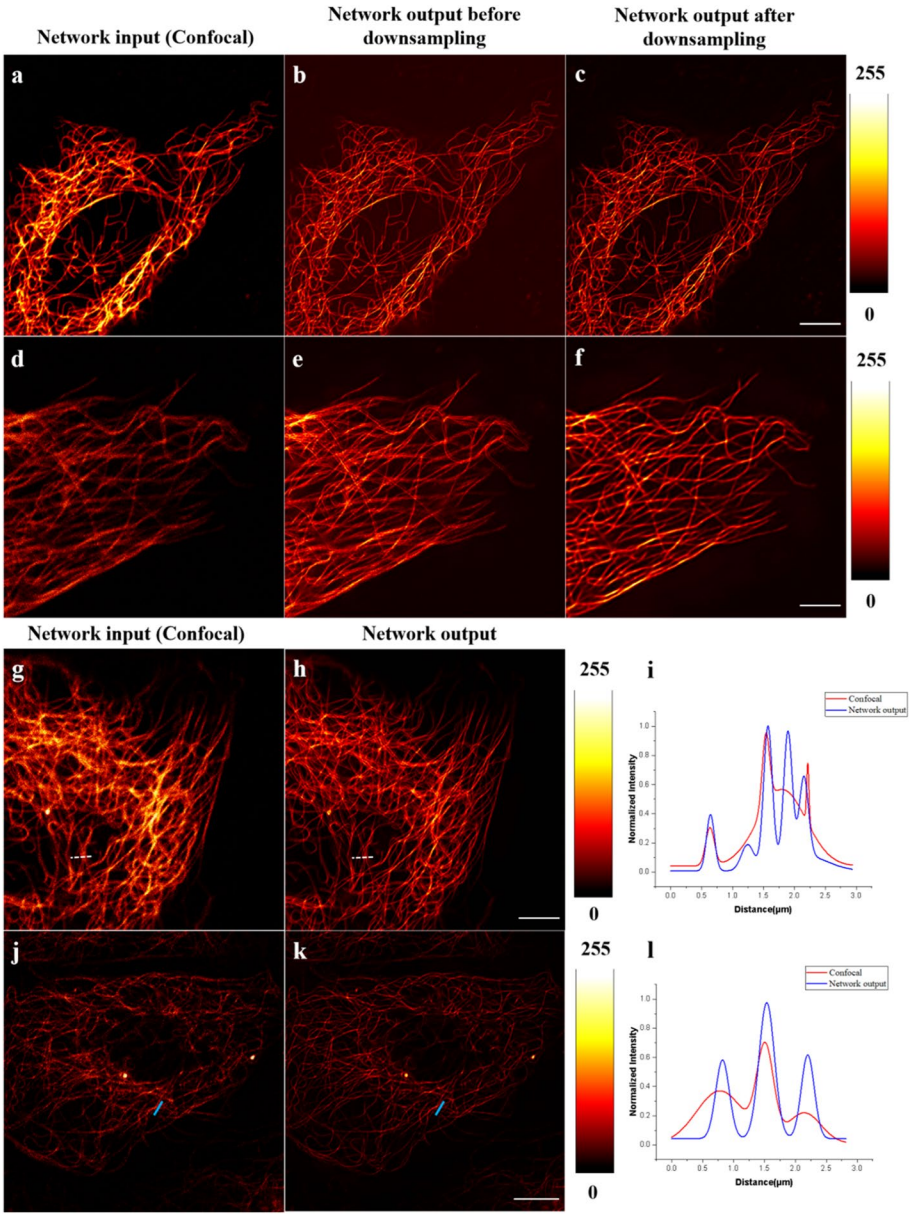

Fig. 7 Super-resolution imaging of confocal images captured with the Nikon A1R MP+ microscope using TCAN. (**a**, **d**, **g**, **j**) Diffraction-limited confocal microscopy images of the microtubules are used as input to the network. (**b**, **e**) The Super-resolved network inference images before downsampling the input images to match the pixel size of the training data. (**c**, **f**) The super-resolved network inference images after downsampling the input images to match the pixel sizes of the training data. (**h**, **k**) The Super-resolved network inference images without the need for downsampling the input images to match the pixel size of the training data. (**i**, **l**) Intensity profiles along the white dashed lines and blue solid lines in different images. Scale bar in (**c**), (**f**), (**h**) and (**k**) are 6 μm, 3 μm, 6 μm and 10 μm, respectively

wide-field and SIM image pairs. Another example of generation of our approach is supported by Fig. 8, where our TCAN model trained with only images of the microtubules is applied to super-resolve actins. Even though this new type of sample is unseen in the training dataset, our network is capable of inferring correctly their fine structures.

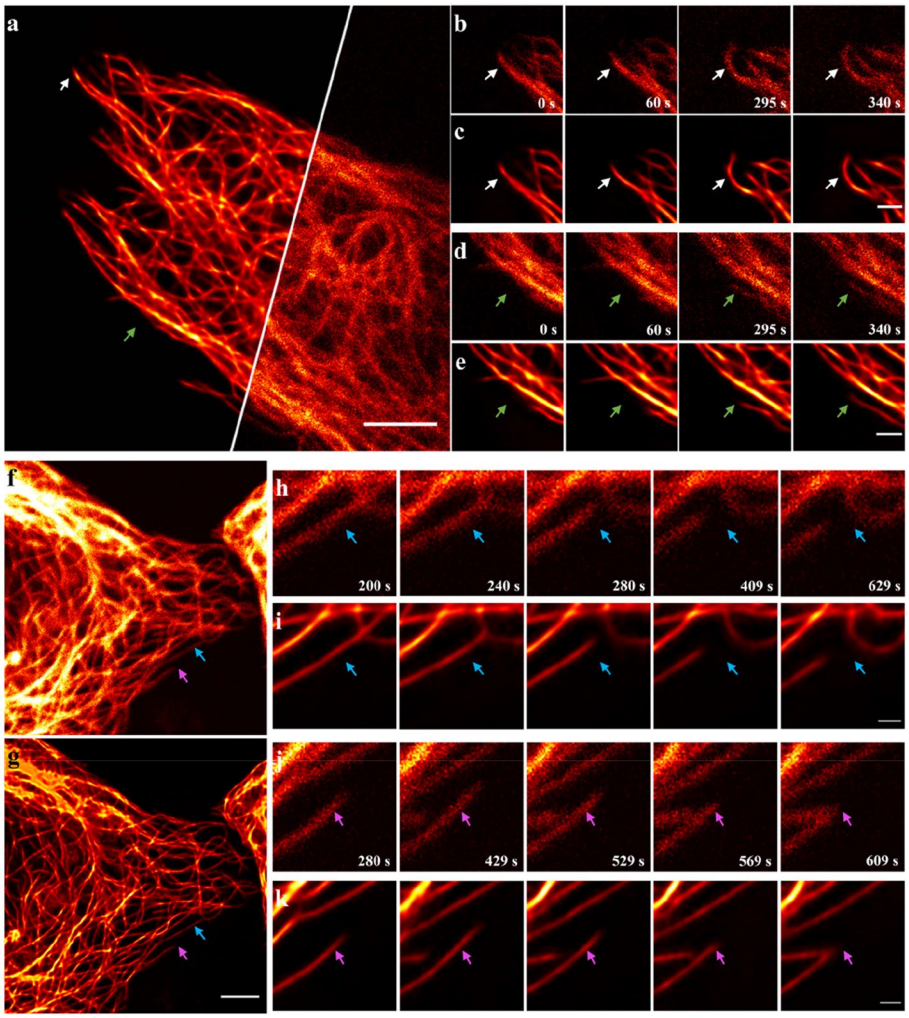

Fig. 8 Super-resolution imaging of live-cell microtubules using TCAN. (**a**) The super-resolved output image of live-cell microtubules. Bottom right: a fraction of the corresponding STED image (Leica TCS SP8 STED confocal microscope) used as comparison. Scale bar, 3 μm. (**b**), (**d**) Time-lapse STED images at four time points are shown at a magnified scale. (**c**, **e**) Time-lapse network output images at four time points are shown at a magnified scale. Scale bar, 1 μm. (**f**) A diffraction-limited confocal microscopy images (Nikon A1R MP+ microscope) used as input to the network. (**g**) The super-resolved output image of live-cell microtubules. Scale bar, 3 μm. (**h**, **j**) Time-lapse confocal images at five time points are shown at a magnified scale. (**i**, **k**) Time-lapse network output images at five time points are shown at a magnified scale. Scale bar, 0.5 μm

Resolution enhancement in confocal images of live-cell microtubules

To test whether TCAN is competent in live-cell imaging, we study the dynamic changes of microtubules by time-lapse imaging. The dynamic instability of the microtubules is important because of their involvement in delivering information, and it is a fast process demanding high spatiotemporal resolution imaging [34].

In this work, we employ the TCAN model trained with static microtubules images to transform low-resolution confocal images of live-cell microtubules into high-resolution ones. The raw images in both the confocal mode and STED mode are acquired for 10 frames at 45 s intervals (Fig. 9a). Figure 9a shows the resolution enhancement and superior image quality when comparing with STED images, and the resolution of our network output images is almost constant within at least 7 minutes (See Visualization 1).

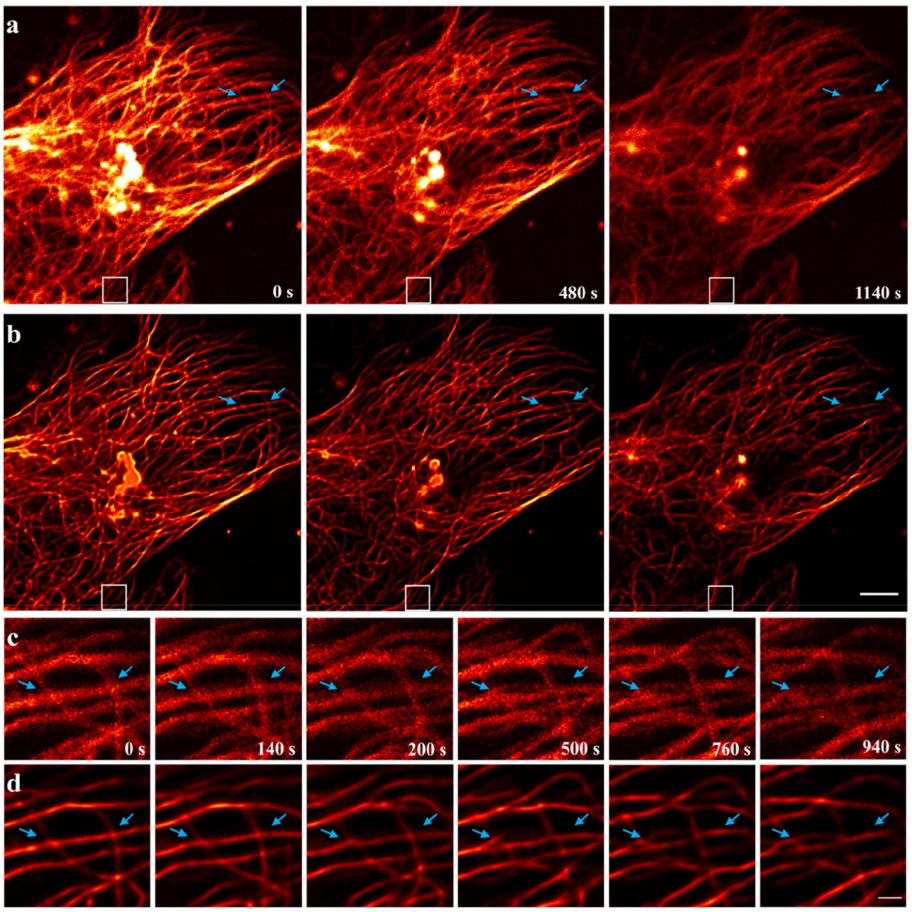

Fig. 9 Super-resolution imaging of live-cell microtubules captured with the Nikon A1R MP+ microscope. (**a**) Time-lapse confocal images at three time points used as input to the network. (**b**) Time-lapse network output images at three time points. Scale bar, 4 μm. (**c**) Time-lapse confocal images at six time points are shown at a magnified scale. (**d**) Time-lapse network output images at six time points are shown at a magnified scale. Scale bar, 1 μm

Then, the dynamic instability of microtubules is visualized, for example, as marked by arrows in Figs. 9b-e. The dynamic changes can be divided into two kinds, one is changing in the shape of microtubules (Figs. 9b-c), and the other is changing in the length of microtubules (Figs. 9d-e). For the first kind, we capture that microtubule varies distinctly, becoming curved from originally straight. This is consistent with the current model for microtubule assembly and dynamics, which postulates that microtubules grow by attachment of curved guanosine triphosphate (GTP)-tubulins to the ends of curved photofilaments [35]. For the second kind, the plus end of the microtubule grows due to assembly, and the quick transitions between microtubule growth and temporal pause even can be observed at a high temporal resolution in our experiments. The high spatial resolution of our TCAN model ensures the precision of microtubules dynamic characterization and detection of densely packed microtubules undetectable with other methods.

Similar improvement can be obtained when applying our method to super-resolve confocal images of live-cell microtubules acquired with the Nikon A1R MP+ microscope

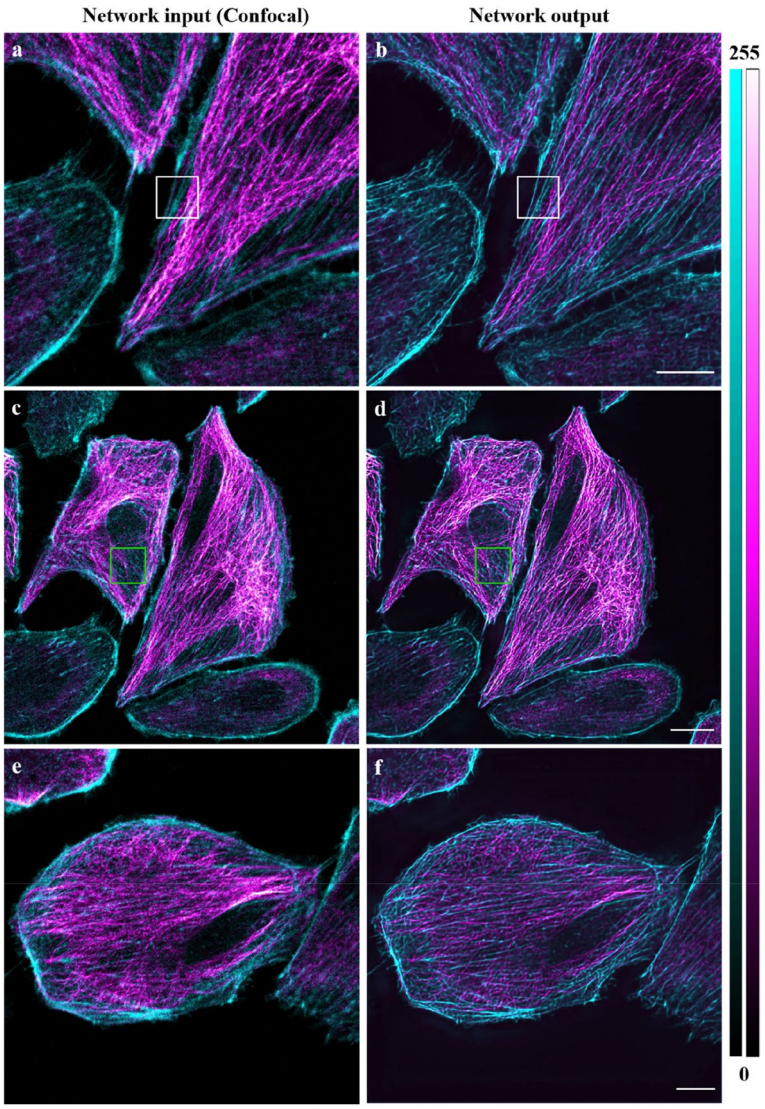

Fig. 10 Super-resolution imaging of dual-color actin-microtubules using TCAN. (**a**, **c**, **e**) Dual-color confocal microscopy images of the microtubules (magenta) and the filaments (cyan) are used as input to the network. (**b**, **d**, **f**) The super-resolved output images of the same field of view. Scale bars in (**b**), (**d**) and (**f**) are 10 μm, 10 μm and 15 μm, respectively

(See Visualization 2). We capture raw images for 31 frames at 20s intervals. This result discerns the dynamic changes at microtubule intersections, and we notice that the intersection, indicated by the blue arrows (Figs. 9h-i), gradually becomes separated because of the microtubule shrinkage. For the microtubule seen at the magenta arrow in Figs. 9j-k, it shrinks and the other microtubule grows over time until they are intersected.

The changes of the separation distance of the intersecting microtubules and microtubules shrinkage can also be viewed in Fig. 10, Visualization 3 and Visualization 4. We capture raw images for 61 frames at 20s intervals. As demonstrated in Ref. [36], lysosome transport has a strong correlation with the distance between the intersecting microtubules, and thus it is crucial to visualize the motion of the complex microtubule networks with a high-resolution. Moreover, the unchanged microtubules in the white

boxes in Fig. 10 signify that our region of interest is in the focus plane during the observation period, excluding the possibility that the dynamic changes of the microtubules are from defocusing. It should also be noted that the imaging time of live-cell microtubules in Visualization 3 and Visualization 4 is about 20 minutes. Since confocal microscope does not suffer from photobleaching and phototoxicity as severely as the STED microscope, our method is fit for long-term super-resolving confocal images of live-cell.

The above results give prominence to the feasibility and advantage of improving image resolution based on deep learning. In other words, the proposed TCAN model is conducive to resist photobleaching in the traditional STED technique by extending the maximum number of usable consecutive frames of time-lapse images [37].

Resolution enhancement in dual-color confocal images of actin-microtubules

As the components of cytoskeleton, actin-microtubule crosstalk is important for the core biological process [38]. Thus, we simultaneously image actin filaments (cyan) and microtubules (magenta) with the Nikon A1R MP+ microscope, and then improve the image resolution by our TCAN model trained with only the microtubules data. Raw confocal images in Fig. 8a, c and e exhibit spurious small structures outside of the filaments and large fluctuations in fluorescence along the actin filaments. In contrast, TCAN suppresses the artifacts and resolves successfully the densely packed structures of the microtubules and the fine branches of the actin filaments (Fig. 8b, d and f). The relative positions of the microtubules and filaments can also be observed in the super-resolved dual-color images. Typical means of crosstalk between the microtubules and actin can be found in our network output, for instance, actin-microtubule crosslinking (white box), actin barrier (green box), and mechanical cooperation (Figs. 8f) [38], while they are not clear in the confocal images due to poor resolution.

Conclusion

In this paper, a deep-learning-based algorithm is developed for the generation of super-resolution images directly from diffraction-limited confocal images without prior information about the image formation and imaging conditions. Quantitative comparison of the framework with STED indicates competitive and often superior performance of TCAN. We demonstrate this by taking confocal raw data as input, and then we can preserve more patterned information and finer structures when enhancing signals from the low-resolution samples, as reported in our Results. The resolution of raw confocal images can be improved from ~ 230 nm to ~ 110 nm of the final network output.

We devise the network architecture, which incorporates both spatial representations and frequency content difference across distinct features, enabling the network to learn more precise mapping from low-resolution images to super-resolution images. The strategy helps us improve the image SNR.

To reduce the effect of pixel sizes on the network output, we upsample/downsample the pixel sizes of the input images to match those of the training data. Accordingly, our algorithm offers the benefit of creating higher-resolution images under the conditions of various FOV. In fact, the image inference performance is more susceptible to pixel sizes and image quality of the training data.

As discussed in Results, we apply an existing trained model on new types of samples and new imaging systems unseen in the training process. Our method can achieve effective image resolution enhancement of the other microscopy modalities and different samples, showing comparable or better performance in comparison with super-resolution method of STED.

Furthermore, TCAN assists the investigation of dynamic instability of live-cell microtubules by capturing long-term time-lapse images. The model needs only the static images as the training data, potentially enabling new opportunity for live-cell imaging with reduced photobleaching and phototoxicity.

We achieve co-imaging of the microtubules and actin cytoskeleton at sufficient spatial resolution by applying our method to resolve dual-color confocal images. This is desirable for exploring how actin and microtubules co-regulate each other and exert their functions in different cellular processes such as cell migration and division.

All these results allow the proposed algorithm to be a prime candidate for computational microscopy and super-resolution imaging, especially with the increasing demand for highly accurate and fast live-cell imaging applications. TCAN also can be applied to improve other types of microscopic images, such as wide-field images and two-photon microscopic images. They have the following characteristics as confocal images, which makes them well suited for resolution enhancement with deep learning. From the optical standpoint, the PSF of their imaging system can be fitted by Gaussian function, and the feature representation of this type of imaging data can be extracted and processed by convolutional neural network which is part of generator in our method. From a deep learning standpoint, since we use supervised learning that requires a "target (ground truth)" in the training set, the network is able to know what the goal of its learning is in the training stage. As done in our experiments, we can select higher-resolution images as the ground truth for the confocal image, such as STED or SIM images, thus the network can learn from them and enhance the input image to match those high-resolution images.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43074-022-00077-x.

Additional file 1. Enhancing image resolution of confocal fluorescence microscopy with deep learning

Acknowledgements
We thank Dr. Min Zhang from Department of Biochemistry at Free University of Berlin and Dr. Iullia Golovynska from Shenzhen University for their helpful discussions.

Authors' contributions
The manuscript was written through contributions of all authors. WY conceives the research. JZ contributes to the experiments with the help of ZY, and BH develops the code with the help of BY. BH prepares the data and figures with the help of JL. The manuscript is written by JL, starting from a draft provided by BH. EYL revises the manuscript. WY provides insight into imaging and microscopy. JQ supervises the research. All authors have given approval to the final version of the manuscript.

Funding
The National Key R&D Program of China (2021YFF0502900); National Natural Science Foundation of China (61835009, 62127819, 61620106016, 62005171, 61975127); Natural Science Foundation of Guangdong Province (2020A1515010679); Key Project of Guangdong Provincial Department of Education (2021ZDZX2013); Shenzhen Science and Technology R&D and Innovation Foundation (JCYJ20220531102807017); Shenzhen International Cooperation Research Project (GJHZ20190822095420249).

Availability of data and materials
The datasets used and analysed during the current study are available from the corresponding author on reasonable request. The Structure of the generative model and the discriminative model, the training process, image acquisition, the effect of the training image quality on the network output, and enhanced images of live-cell microtubules (Videos) are in supplementary files.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 8 July 2022 Revised: 2 November 2022 Accepted: 14 November 2022
Published: 5 January 2023

References

1. Sage D, Kirshner H, Pengo T, Stuurman N, Min J, Manley S, et al. Quantitative evaluation of software package for single-molecule localization microscopy. Nat Methods. 2015;12(8):717–24.
2. Rust M, Bates M, Zhuang X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nat Methods. 2006;3(10):793–6.
3. Gustafsson MGL. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. J Microsc. 2000;198:82–7.
4. Agarwal K, Macháň R. Multiple signal classification algorithm for super-resolution fluorescence microscopy. Nat Commun. 2016;7:13752.
5. Hell SW, Wichmann J. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. Opt Lett. 1994;19:780–2.
6. Wang C, Taki M, Sato Y, Tamura Y, Yaginuma H, Okada Y, et al. A photostable fluorescent marker for the super-resolution live imaging of the dynamic structure of the mitochondrial cristae. Proc Natl Acad Sci U S A. 2019;116(32):15817–22.
7. Vicidomini G, Bianchini P, Diaspro A. STED super-resolved microscopy. Nat Methods. 2018;15(3):173–82.
8. Yang Z, Sharma A, Qi J, Peng X, Lee DY, Hu R, Lin D, Qu J. J Seung Kim, "Super-resolution fluorescent materials: an insight into design and bioimaging applications." Chem Soc Rev. 2016;45:4651–67.
9. Wang H, Rivenson Y, Jin Y, Wei Z, Gao R, Günaydın H, et al. Deep learning enables cross-modality super-resolution in fluorescence microscopy. Nat Methods. 2019;16:103–10.
10. Qiao C, Li D, Guo Y, Liu C, Jiang T, Dai Q, et al. Evaluation and development of deep neural networks for image super-resolution in optical microscopy. Nat Methods. 2021;18:194–202.
11. Patterson J. A Gibson, deep learning: a Practitioner's approach: O'Reilly Media; 2017.
12. LeCun Y, Bengio Y, Hinton G. Deep learning. Nat. 2015;521(7533):436–44.
13. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE; 2016. p. 770–8.
14. Zhang K, Zuo W, Chen Y, Meng D, Zhang L. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans Image Process. 2017;26(7):3142–55.
15. Ouyang W, Aristov A, Lelek M, Hao X, Zimmer C. Deep learning massively accelerates super-resolution localization microscopy. Nat Biotechnol. 2018;36:460–8.
16. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172:1122–31.
17. Wang N, Yan W, Qu Y, Ma S, Li SZ, Qiu M. Intelligent designs in nanophotonics: from optimization towards inverse creation. PhotoniX. 2021;2:22.
18. Wang K, Zhang MM, Tang J, Wang L, Hu L, Wu X, et al. Deep learning wavefront sensing and aberration correction in atmospheric turbulence. PhotoniX. 2021;2:8.
19. O Ronneberger, P Fischer, T Brox, "U-net: convolutional networks for biomedical image segmentation," arXiv: 1505.04597 (2015).
20. Maas AL, Hannun AY, Ng AY. Rectifier nonlinearities improve neural network acoustic model. In: 30th International Conference on Machine Learning (ICML). Atlanta: IMLS; 2013. p. 6–11.
21. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Hawaii: IEEE; 2017. p. 2261–9.
22. D Hendrycks, K Gimpel, "Gaussian error linear units (GELUs)," arXiv: 1606.08415 (2016).
23. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: 14th International Conference on Artificial Intelligence and Statistics (AISTATS). Fort Lauderdale: Society for Artificial Intelligence and Statistics; 2011.
24. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process. 2004;13:600–12.
25. DP Kingma, J Ba, "Adam: a method for stochastic optimization," arXiv: 1412.6980 (2014).

26. Paszke A. Automatic differentiation in PyTorch. In: in 31st Conference on Neural Information Processing Systems. Long Beach: NeurIPS Foundation; 2017.
27. Pan SJ, Yang Q. A survey on transfer learning. IEEE T Knowl Data En. 2010;22(10):1345–59.
28. Castello M, Tortarolo G, Buttafava M, Deguchi T, Villa F, Koho S, Pesce L, Oneto M, Pelicci S, Lanzanó L, Bianchini P, Sheppard CJR, Diaspro A, Tosi A, Vicidomini G. A robust and versatile platform for image scanning microscopy enabling super-resolution FLIM. Nat Methods. 2019;16:175–8.
29. Descloux A, Grußmayer KS, Radenovic A. Parameter-free image resolution estimation based on decorrelation analysis. Nat Methods. 2019;16:918–24.
30. Isola P, Zhu J-Y, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Hawaii: IEEE; 2017. p. 5967–76.
31. IJ Goodfellow, J Pouget-Abadie, M Mirza, B Xu, D Warde-Farley, S Ozair, A Courville, Y Bengio, "Generative adversarial networks," arXiv: 1406.2661 (2014).
32. Girshick R. Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE; 2015. p. 1440–8.
33. Zhao H, Gallo O, Frosio I, Kautz J. Loss functions for image restoration with neural networks. IEEE T Comput Imag. 2017;3(1):47–57.
34. Guo Y, Li D, Zhang S, Yang Y, Liu J-J, Wang X, Liu C, Milkie DE, Moore RP, Tulu US, Kiehart DP, Hu J, Schwartz JL, Betzig E, Li D. Visualizing intracellular organelle and cytoskeletal interactions at nanoscale resolution on millisecond timescales. Cell. 2018;175:1430–42.
35. Gudimchuk NB, McIntosh JR. Regulation of microtubule dynamics, mechanics and function through the growing tip. Nat. Rev. Mol. Cell Bio. 2021;22:777–95.
36. Bálint Š IV, Vilanova ÁSÁ, Lakadamyali M. Correlative live-cell and superresolution microscopy reveals cargo transport dynamics at microtubule intersections. Proc Natl Acad Sci U S A. 2013;110(9):3375–80.
37. Huang X, Fan J, Li L, Liu H, Wu R, Wu Y, et al. Fast, long-term, super-resolution imaging with hessian structured illumination microscopy. Nat Biotechnol. 2018;36(5):451–9.
38. Dogterom M, Koenderink GH. Actin-microtubule crosstalk in cell biology. Nat Rev Mol Cell Bio. 2019;20:38–54.

Publisher's Note