nature methods

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41592-018-0239-0

In the format provided by the authors and unedited.

Deep learning enables cross-modality super-resolution in fluorescence microscopy

Hongda Wang[1,2,3,9], Yair Rivenson[1,2,3,9], Yiyin Jin[1], Zhensong Wei[1], Ronald Gao[4], Harun Günaydın[1], Laurent A. Bentolila[3,5], Comert Kural[6,7] and Aydogan Ozcan[1,2,3,8]*

[1]Electrical and Computer Engineering Department, University of California, Los Angeles, CA, USA. [2]Bioengineering Department, University of California, Los Angeles, CA, USA. [3]California NanoSystems Institute, University of California, Los Angeles, CA, USA. [4]Computer Science Department, University of California, Los Angeles, CA, USA. [5]Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, USA. [6]Department of Physics, Ohio State University, Columbus, OH, USA. [7]Biophysics Graduate Program, Ohio State University, Columbus, OH, USA. [8]Department of Surgery, David Geffen School of Medicine, University of California, Los Angeles, CA, USA. [9]These authors contributed equally: Hongda Wang and Yair Rivenson.
*e-mail: ozcan@ucla.edu

Table of Contents

Supplementary Notes:

Supplementary Figures:

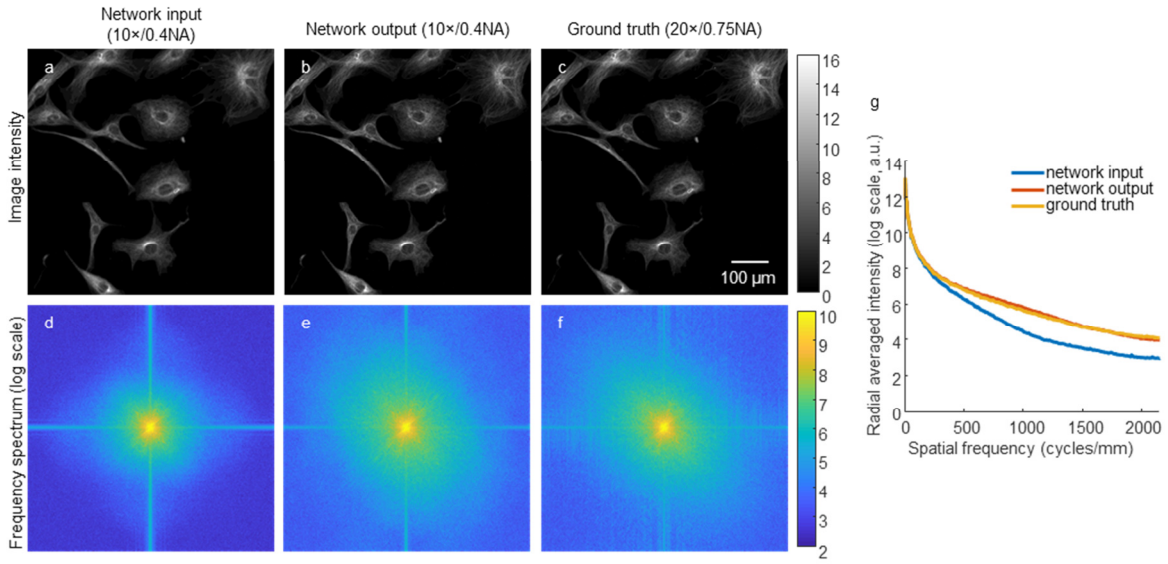
Supplementary Table:

Supplementary Note 1: Generalization of neural network

We tested the generalization of our trained network model in improving image resolution on new types of samples that were not present in the training phase for the wide-field microscopy experiments. Supplementary Fig. 6 demonstrates the resolution enhancement when applying our network model trained with F-actin (Supplementary Figs. 6(a-c)) to super-resolve images of mitochondria in BPAEC (Supplementary Figs. 6(d-f)), blood vessels in a mouse brain tumor (Supplementary Figs. 6(g-i)), and actin in a mouse kidney tissue (Supplementary Figs. 6(j-l)). Even though these new types of objects were not part of the network's training set, the deep network was able to correctly infer their fine structures through blind inference. Another example of this generalization behavior of our approach is shown in Supplementary Fig. 7, where the F-actin in BPAEC (Supplementary Figs. 7(d-f)), melanoma cells in a mouse brain tumor (Supplementary Figs. 7(g-i)), and glomeruli and convoluted tubules in a mouse kidney tissue (Supplementary Figs. 7(j-l)) are super-resolved by a neural network that was trained with *only* the images of microtubules captured with FITC filter set. In these experiments both the training and the blind testing images were taken with the same fluorescence filter set. Supplementary Figs. 6(e,f,h,i,k,l), 7(e,f,h,i,k,l) further support the enhanced DOF of the network output images for various types of samples when compared to the ground truth, higher NA images.

Supplementary Note 2: Quantification of the deep network results using spatial frequency spectrum analysis

For wide-field microscopy images, we performed quantification of the deep network results using spatial frequency spectrum analysis: in Supplementary Fig. SN2.1 we compare the spatial frequency spectrum of the network output images (for BPAEC structures) with respect to the network input images to demonstrate the frequency extrapolation nature of our deep learning framework. The cross-section of the radially-averaged power spectrum confirms the success of the network output, matching the extended spatial frequency spectrum that is expected from a higher-resolution imaging system (as illustrated with the overlap of the red and orange curves in Supplementary Fig. SN2.1(g)).



Supplementary Figure SN2.1

Illustration of the spatial frequency extrapolation achieved by deep learning. The deep learning model takes (a) an input image of microtubules in BAPEC obtained using a 10×/0.4NA objective lens and super-resolves it as shown in (b), to match the resolution of (c) the ground truth image which is acquired with a 20x/0.75NA objective lens. (d-f) show the spatial frequency spectra in log scale, corresponding to (a-c), respectively. (g) shows the radially-averaged intensity of each one of the spatial frequency spectra shown in (d,e,f). Analysis was performed on a randomly selected image from a group of 94 images with similar results.

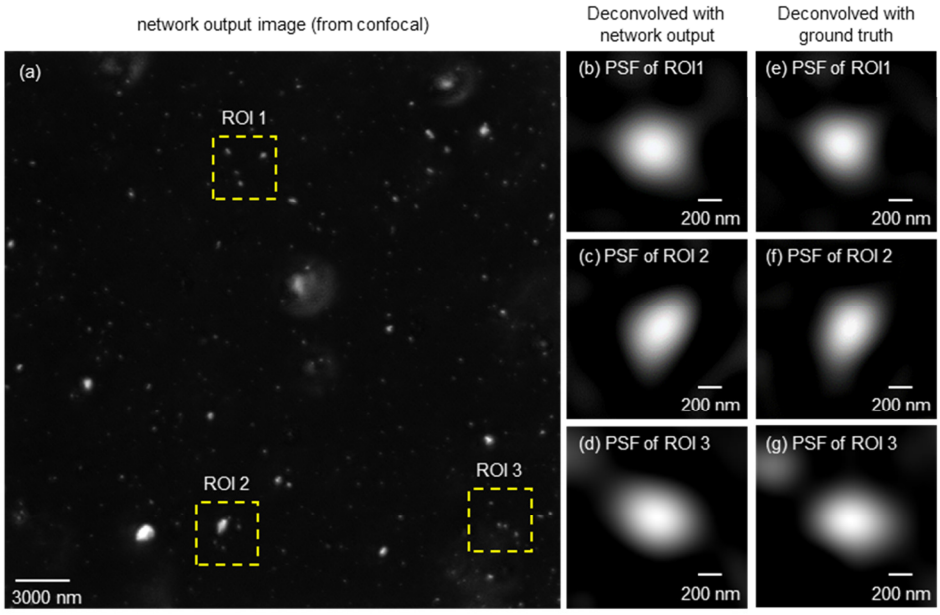
Supplementary Note 3: Quantification of resolution enhancement in wide-field images using PSF analysis

We further quantified the resolution improvement achieved in wide-field images using our approach by imaging 20 nm fluorescent beads at an emission wavelength of 645 nm (see the **Methods** section) and used the images acquired with a 10×/0.4NA objective lens as input to our deep network model, which was trained *only* with F-actin (as demonstrated in Figures 1 and 2). The super-resolution results of the deep network are summarized in Supplementary Fig. 8. To quantify the resolution improvement in these results, we measured the PSFs arising from the images of single/isolated nano-beads across the imaging FOV[1]; this was repeated for >100 individual particles that were tracked in the network input and output images, as well as the ground truth images (acquired using a 20×/0.75NA objective lens). The full-width at half-maximum (FWHM) of the 10× input image PSF is centered at ~1.25μm, corresponding to a sampling rate limited by an effective pixel size of ~0.65 μm. Despite the fact that the fluorescent signal from 20 nm beads is rather weak, our deep neural network (trained only with BPAEC samples) successfully picked up the signal from individual nano-beads and blindly improved the resolution to match that of the ground truth, as shown in the PSF comparison reported in Supplementary Fig. 8(d). These results further highlight the robustness of our deep learning method to low SNR (signal-to-noise ratio) as well as its generalizability to different spatial structures of the object. The broadening of the PSF distribution in 20×/0.75NA images (see Supplementary Fig. 8(d)) can be attributed to the smaller DOF of the high NA objective lens, where the nano-beads at slightly different depths are not in perfect focus and therefore result in varying PSF widths. The deep network results, on the other hand, once again demonstrate the enhanced DOF of our network output image, showing uniform focusing with improved resolution at the network output image.

Supplementary Note 4: Calculation of the learned PSFs of the confocal-STED cross-modality transformation network

Since we establish a data-driven image transformation (from lower resolution to higher resolution images, after the network converges), we can estimate the effective local PSF of the lower-resolution imaging system *with respect to* the ground truth modality used in the training phase. This can also be useful to shed more light onto the inner workings of the deep neural network and help us better understand its inference success. For this, we used the confocal-to-STED transformation results that we presented earlier to calculate the "learned" PSFs of our deep neural network by locally deconvolving the network output with the network input, through sub-regions of 20 nm particle images. As shown in Supplementary Fig. SN4.1, the results reveal a significant variation in the inferred PSF as a function of the FOV, which highlights another advantage of our framework in comparison to standard deconvolution methods that assume a shift-invariant PSF. Thus, in our presented approach the spatially-varying PSF information is indirectly learned at the end of the training phase through image data, without any prior assumptions about the image formation process or related aberrations.

We calculated the above discussed local PSFs with a pair of network input (confocal) and network output images, by deconvolving the same local regions of the input images with the corresponding output images using the regularized inverse filter (RIF), with regularization parameter defined as the inverse of the noise variance so that the RIF becomes equivalent to Wiener filtering.[2] This algorithm is performed using Fiji[3] plugin DeconvolutionLab2[4], while setting the input local region as the image to be deconvolved. The resulting deconvolved image from this process can be regarded as the local PSF (*with respect to the ground truth modality* used in the training phase) that is learned by the neural network.



Supplementary Figure SN4.1

Spatially-varying PSFs that our confocal-to-STED transformation neural net has converged to. These spatially-varying PSFs are locally calculated within our imaging FOV, by deconvolving the network output with the network input (Figures 3(a) and (b) of the main text). These results clearly demonstrate

the power of our presented framework that achieves both cross-modality image transformations (i.e., confocal-to-STED in this case) and blind measurement of the inherent spatially-varying PSF (with respect to STED), statistically inferred through deep learning using image data. The corresponding match between (b, c, d) and (e, f, g), respectively, is another validation for the success of our super-resolution framework. Analysis was performed on a randomly selected image from a group of 75 images with similar results.

Supplementary Note 5: Discussion on data-driven cross-modality transformation framework enabled by GANs

The generalized point spread function of an imaging system, which accounts for the finite aperture of the optical system, as well as its aberrations, noise and optical diffraction, can be considered as a probability density function, $p(\zeta,\eta)$, where ζ, η denote the spatial coordinates. $p(\zeta,\eta)$ represents the probability of photons emitted from an ideal point source on the sample to arrive at a certain displacement on the detector plane. Therefore, the super-resolution task that the presented deep learning framework has been learning is to transform the input data distribution $X\left(p_{\mathrm{LR}}(\zeta,\eta)\right)$ into a high-resolution output, $Y\left(p_{\mathrm{HR}}(\zeta,\eta)\right)$, where the former is created by a lower resolution (LR) imaging system and the latter represents a higher resolution (HR) imaging system. The network architecture that we have used for training, i.e., GANs[5] have been proven to be extremely effective in learning such distribution transformations ($X \rightarrow Y$) *without* any prior information on or modelling of the image formation process or its parameters.[6,7]

It is important to note that GANs were originally proposed to create new data points that appear as if they were drawn from a desired distribution, and the seeds of such a generative network were random noise vectors[5]. Unlike these earlier studies, in this work the input and output distributions share a high degree of mutual information, and the output probability distribution is conditional upon the input data distribution. This extension, where the GAN output is conditioned upon some external information, is broadly referred to as "conditional GAN"[8,9]. Stated differently, the conditional GAN framework uses prior information that restricts the possible generator outputs. In fact, through a rigorous image alignment and registration process (precisely matching the fields-of-view of the input images with the ground truth labels, as described in the **Methods** Section), the training process of a GAN establishes this mutual information between the input and label images as a strong constraint and regularization term for the network to perform super-resolution on input images. In addition, we also restrict the possible generator outputs using the structural similarity and mean-squared-error terms (see the **Methods** Section, Equation 1), calculated with respect to the high-resolution label images. Stated differently, the high resemblance between the low-resolution input images and the high-resolution labels is a powerful regularization established during the training process that strongly suppresses artefacts. Moreover, in case feature hallucinations are observed in e.g., the images of new types of samples that were not introduced to the network during the training, these can be penalized in the loss function as they are discovered, and the network can be further regularized to avoid such artifacts from repeating, e.g., by changing the weights of the loss function terms, or adding other loss terms (see the **Methods** Section).

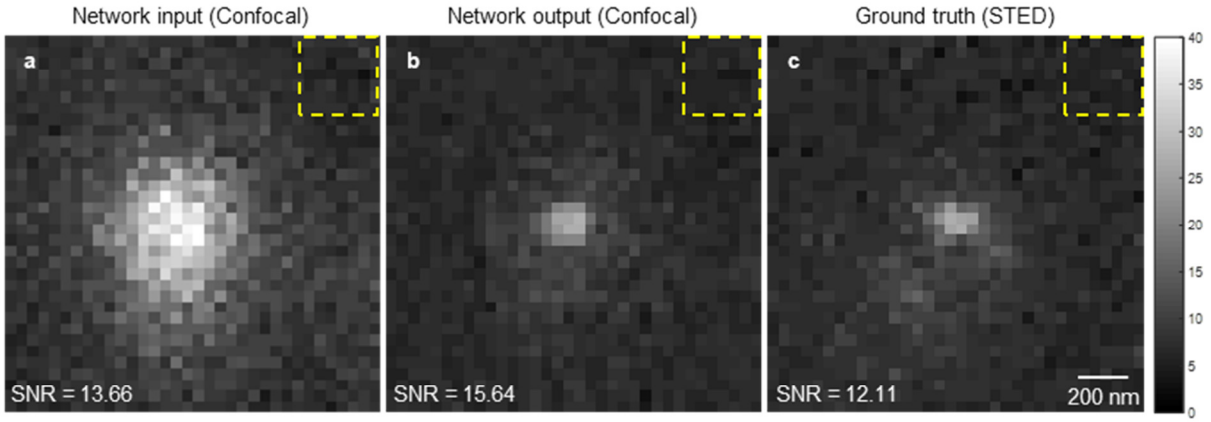
Unlike other statistical super-resolution methods, the presented approach is data-driven, and the deep network aims to find a distribution generated by real microscopic imaging systems that it was trained with. This feature makes the network much more robust to poor image SNR or aberrations of the imaging system, also eliminating the need for prior information on e.g., the PSF[10] and sensor-specific noise patterns, which are required for any standard deconvolution and localization method[11]. A similar resilience to spatial and spectral aberrations of an imaging system has also been demonstrated for bright-field microscopic imaging using a neural network.[12]

Supplementary Note 6: Quantification of SNR improvement

To further highlight the SNR improvement achieved by our deep learning-based super-resolution approach, we conducted an additional analysis using our confocal-to-STED network results (see Supplementary Fig. SN6.1). For this analysis, we selected a small FOV containing a single 20 nm bead and calculated the SNR for the network input (confocal image), the network output and the ground truth image (STED). The SNR is defined as:

$$SNR = \left| \frac{s - \bar{b}}{\sigma_b} \right| \tag{1}$$

where s is the peak value of the signal calculated from a Gaussian fit to the particle (see the **Methods** section), $\bar{b}$ is the mean value of the background (e.g. the regions defined with the yellow dashed lines in Supplementary Fig. SN6.1), σ_b is the standard deviation of the background. The results shown in Supplementary Fig. SN6.1 reveal that the deep neural network suppresses noise and improves the SNR compared to the input image as well as the ground truth image (STED).



Supplementary Figure SN6.1

Quantification of the SNR improvement achieved by our confocal-to-STED transformation network. (a) Input image SNR= 13.66. (b) Network output image SNR=15.64. (c) STED image SNR= 12.11. The yellow dashed line regions are used to calculate the background mean and variation. Analysis was performed on a randomly selected particle from a group of 75 images with similar results.

Supplementary Note 7: Quantification of super-resolution artifacts using NanoJ-Squirrel

We quantified the level of artifacts in the network output images using the Fiji[3] software plugin NanoJ-Squirrel[13]. The plugin iteratively estimates a resolution scaling function (RSF) from the low-resolution (LR) image to the high-resolution (HR) image, convolves the HR image with this RSF and calculates its pixel-wise absolute difference from the LR image. The plugin also provides two globally averaged scores: Resolution Scaled Error (RSE) and Resolution Scaled Pearson coefficient (RSP), defined as:

$$\begin{aligned} \mathrm{RSE}(f,g) &= \sqrt{\frac{\sum_{x,y}\left(f(x,y)-g(x,y)\right)^2}{n}} \\ \mathrm{RSP}(f,g) &= \frac{\sum_{x,y}\left(f(x,y)-\overline{f}\right)\left(g(x,y)-\overline{g}\right)}{\sqrt{\sum_{x,y}\left(f(x,y)-\overline{f}\right)}\sqrt{\sum_{x,y}\left(g(x,y)-\overline{g}\right)}} \end{aligned} \tag{2}$$

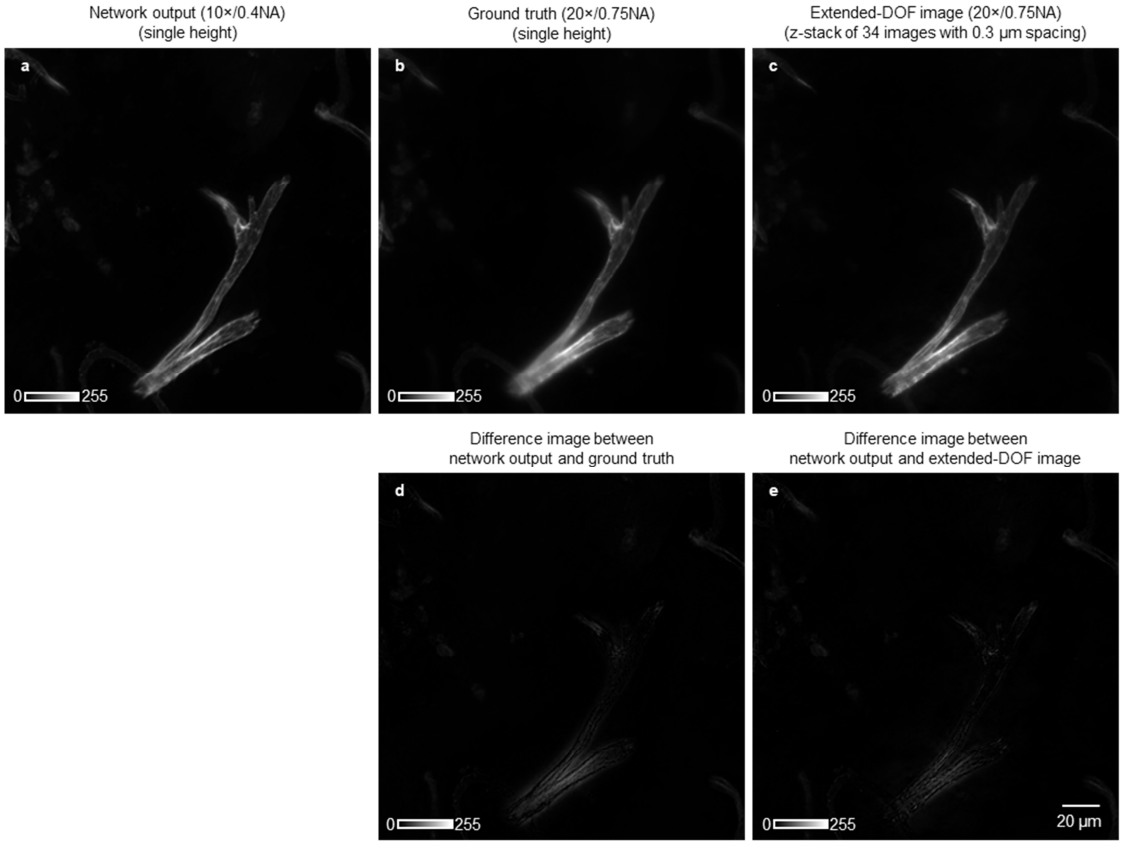
where, f and g are the LR and simulated LR images, respectively, and $(\overline{\ })$ refers to the two-dimensional mean operator. Generally, the RSE is more sensitive to brightness and contrast differences, while the RSP helps to assess the image qualities across modalities, by quantifying their correlation.

In our implementation using this plugin, the "Reference image" was set to the LR input image, the "Super-resolution reconstruction" was set to the network output image. "RSF Estimate Image" was set to "RSF unknown, estimate via optimization" with "Max. Mag. in Optimization" set to 5. The error map of the network's output image with respect to the network's input (LR image) is shown in Supplementary Fig. 1(e), resulting in RSE = 0.912 and RSP = 0.999.

We then repeated the same operations detailed above, estimating the error map between the low-resolution input image and the ground truth (HR) image, as shown in Supplementary Fig. 1(f), which resulted in RSE = 1.509 and RSP = 0.998. These results show that the network output image does not generate noticeable super-resolution related artifacts and in fact has the same level of spatial mismatch error that the ground truth HR image has with respect to the LR input image (with a correlation of ~1 and an absolute error ~1 out of 255). This conclusion is further confirmed by Supplementary Fig. 1(d), which overlays the network output image and the ground truth image using different colors, revealing no obvious feature mismatch between the two.

Supplementary Note 8: The differences between the network output images and the corresponding ground truth images

In Supplementary Figs. 10 and 11, we show the differences between the network output and ground truth images for all the modalities used in this manuscript to demonstrate the high degree of agreement between the network output images and the corresponding ground truth labels. The minor differences between the network output and ground truth images are partially due to the extended-DOF of our output images, as detailed earlier. In fact, in Supplementary Fig. SN8.1 we also demonstrate that the synthesized extended-DOF image (which used a z-stack of 34 axially-separated images to digitally extend the DOF of the ground truth) shows smaller difference with respect to our network output; this clearly explains that part of the difference between the network output and a single ground truth image is due to the limited DOF of the latter, not because of network hallucination.



Supplementary Figure SN8.1

The extended-DOF image provides an improved ground truth image. (a) Network output image from a single height input image. (b) High-resolution image captured with a 20×/0.75NA objective lens. (c) extended-DOF image synthesized from 34 high resolution images, separated axially by 0.3 µm. (d) Difference image between (a) and (b). (e)Difference image between (a) and (c). As shown in the comparison of (d) and (e), extended-DOF image shows smaller difference than a single height ground truth image when compared against the network output image. This confirms that part of the difference between the network output and a single ground truth image is due to the limited DOF of the latter. Analysis was performed on the same images reported in Supplementary Fig. 9.

Supplementary Note 9: Calculation of the image shift from normalized cross-correlations

Given two images to be registered, the first step is to calculate the normalized cross-correlation map, which is defined as:

$$\mathrm{nCCM} = \frac{\mathrm{CCM} - \min(\mathrm{CCM})}{\max(\mathrm{CCM}) - \min(\mathrm{CCM})} \cdot \left(\mathrm{PPMCC}_{\max} - \mathrm{PPMCC}_{\min}\right) + \mathrm{PPMCC}_{\min} \tag{3}$$

where CCM is the cross-correlation map[14] defined as:

$$\mathrm{CCM}(u,v) = \sum_{x,y}\left[f(x,y) - \overline{f}\right]\left[g(x-u, y-v) - \overline{g}\right], \tag{4}$$

where, f and g represent two images, and $(\overline{\ })$ refers to the two-dimensional mean operator. The locations of the maximum and minimum values of CCM indicate the most likely and the most unlikely (respectively) relative shifts of the input image pair. PPMCC is the Pearson product-moment correlation coefficient, defined as:

$$\mathrm{PPMCC}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{5}$$

where cov is the covariance function, σ_X is the standard deviation of X, and σ_Y is the standard deviation of Y. The values of $\mathrm{PPMCC}_{\max}$ and $\mathrm{PPMCC}_{\min}$ refer to the Pearson product-moment correlation coefficients calculated when applying the most likely and the most unlikely shifts to the input images, respectively. The normalized cross-correlation map (nCCM) is then fit to a 2D Gaussian function, which is defined as:

$$G(x,y) = A\exp\left(-\left(\frac{(x-x_o)^2}{2\sigma_x^2} + \frac{(y-y_o)^2}{2\sigma_y^2}\right)\right). \tag{6}$$

where x_o and y_o refer to the refined sub-pixel shift amount in x and y direction, respectively, between the input image pairs, and A refers to the similarity of the two images.

Supplementary Note 10: Neural network training procedures

1. Prepare images for neural network training

    a. Gather low- and high-resolution image pairs of the same samples for network training.

    b. If the effective pixel size of low- and high- resolution images are different (e.g., images captured with different objective-lenses or microscopes), upsample the ones with larger pixel size to match the scales of each image pair.

    c. Match the fields-of-view (FOVs) of the low-and high-resolution images so that each training pair shows roughly the same region of the sample.

    d. **Critical** Apply *rigid image registration* algorithms to each image pair to correct lateral shifts. In this manuscript, we calculated the normalized cross-correlation map to calculate the shift amount between the two image sets, as described in **Supplementary Note 9**. Our method detects sub-pixel shifts (as small as ~0.1 pixel) and corrects them by linear interpolation.

    e. **Critical** Apply the *elastic image registration* algorithm, which can be especially important if there are spatial and/or spectral aberrations between the low and high-resolution image pairs (see the **Methods** section: Image pre-processing and **Supplementary Fig. 12**).

    f. Crop the borders of the post-registered images to avoid any registration-step related artifacts. The cropped amount should be larger than the shift amount. The typical cropping size is ~20 pixels on each side.

    g. Save the image pairs to the format that can be accessed by your training program. In this manuscript, we converted all the images to single precision floating data type, scaled to 0~255 dynamic range, and saved as NumPy (.npy) files.

2. Training a neural network model

    a. Randomly initialize the parameters of the generative and discriminative models.

    b. Set the loss function to be:

$$\mathcal{L}(G;D) = -\log D\big(G(x)\big) + \lambda \times \mathrm{MSE}\big(G(x), y\big) - \nu \times \log\Big[\big(1 + \mathrm{SSIM}\big(G(x), y\big)\big)/2\Big]$$

$$\mathcal{L}(D;G) = -\log D(y) - \log\Big[1 - D\big(G(x)\big)\Big]$$

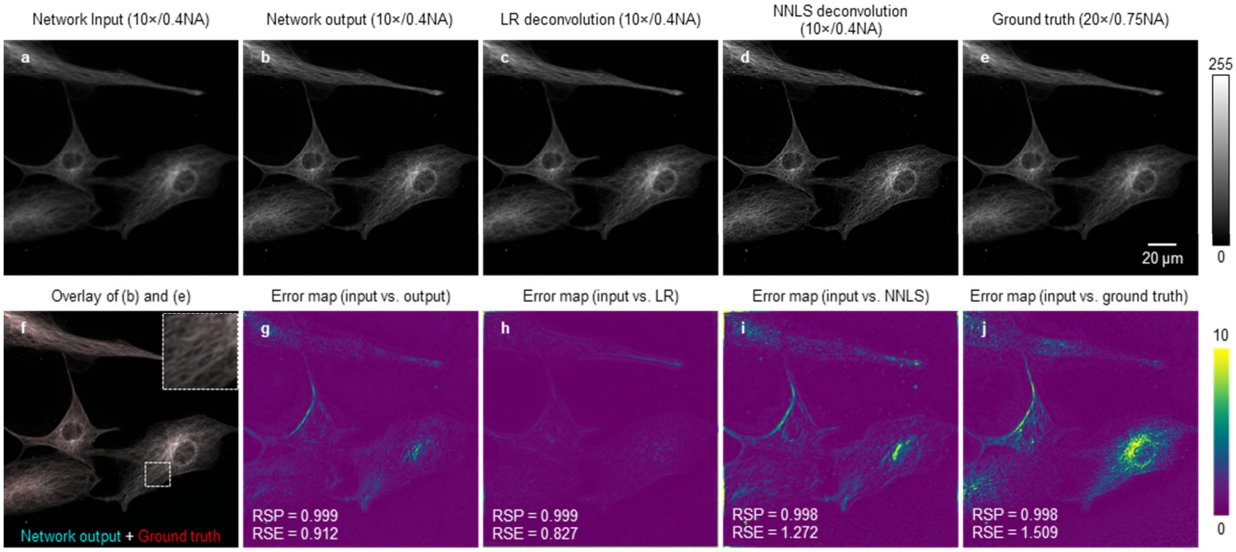
    as described in the **Methods** section of the main text. Take a few trials and set λ and ν to accommodate the MSE loss and the SSIM loss to be ~1-10% of the combined generative model loss $\mathcal{L}(G;D)$.

    c. Use the Adam optimizer with a starting learning rate of e.g., 1×10^{-4} and 1×10^{-5} for the generative and discriminative models, respectively.

    d. **Critical** Adjust the learning rates, as needed, to balance the adversarial training process: if the discriminative model always wins, increase the generative model learning rate and decrease the discriminative model learning rate, and vice versa if the if the generative model always wins.

e. **Critical** If the adversarial training process cannot be balanced by tuning the learning rates, adjust the ratio of the optimization steps for the generative and discriminative models. For example, when discriminative model is overwhelmingly strong, optimize generative model a few times while only optimizing discriminative model once in each iteration.

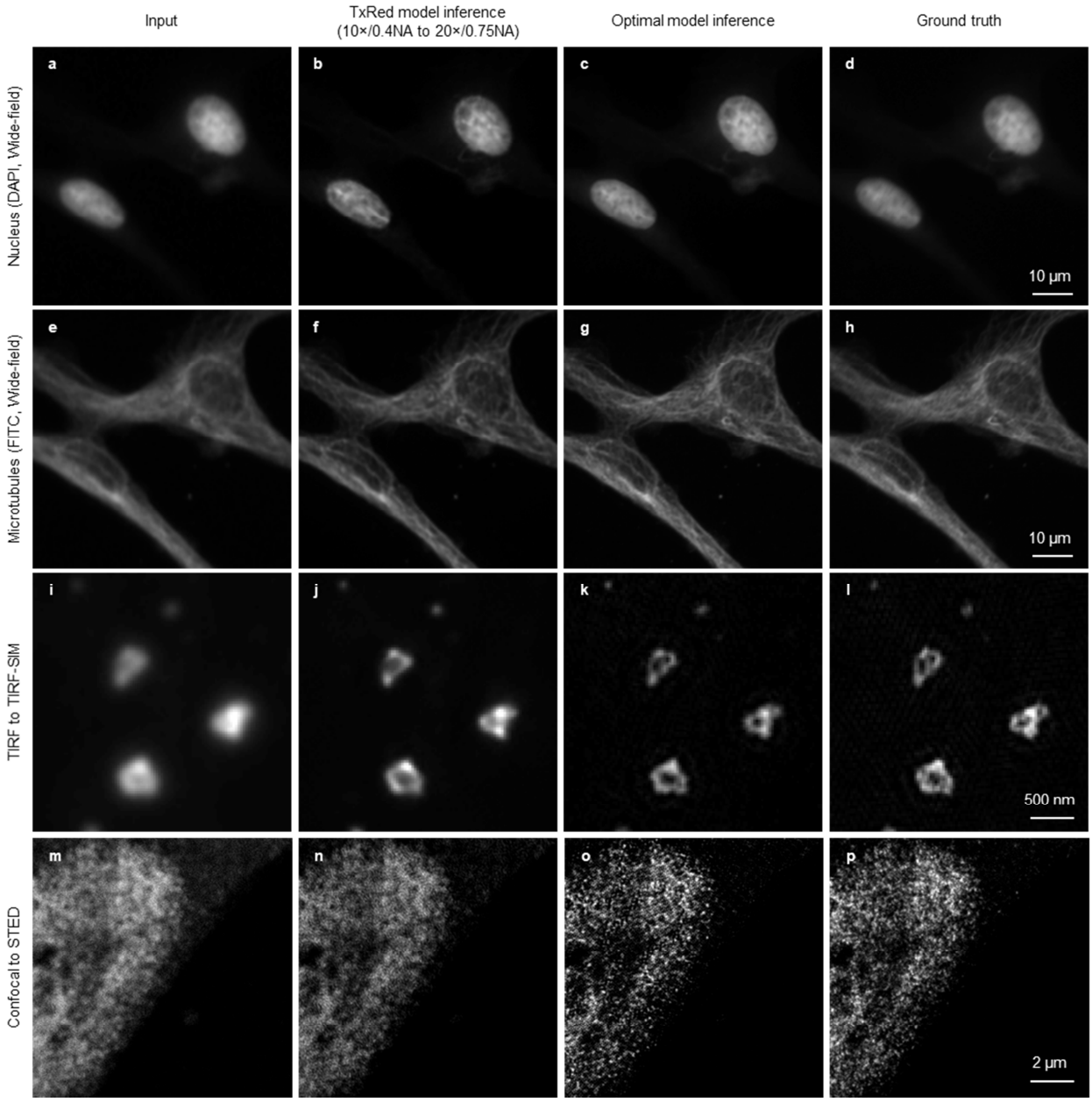
f. Validate the models with an *independent* validation dataset and select the best generative model with the lowest validation loss.



Supplementary Figure 1

Quantification of super-resolution artifacts using the NanoJ-Squirrel Plugin.[13]

(a) Network input, (b) network output, (c) Lucy-Richardson (LR) deconvolution, (d) non-negative least square (NNLS) deconvolution, and (e) ground truth images of the microtubule structure inside a BPAEC. (f) Overlay image of (b) in cyan and (e) in red shows sharp features without red or cyan color blocks, which means there is no obvious feature mismatch between the network output image and the ground truth image. (g-j) Error maps of the network input image vs. the network output image (g), LR deconvolution image (h), NNLS deconvolution image (j), and the ground truth image (j), calculated by NanoJ-Squirrel. All the maps (g-j) show high RSP (resolution scale Pearson-correlation) scores that are almost 1, and low resolution scaled error (RSE) scores of ~1, out of 255 (see **Supplementary Note 7**). Note that the network output image has better agreement in RSE than the ground truth image. This can be partially explained by the larger depth-of-field of a low NA objective lens that is used for acquiring the network input image (also see the **Results** section of the main text). Analysis was performed on a randomly selected image from a group of 94 testing images with similar results.

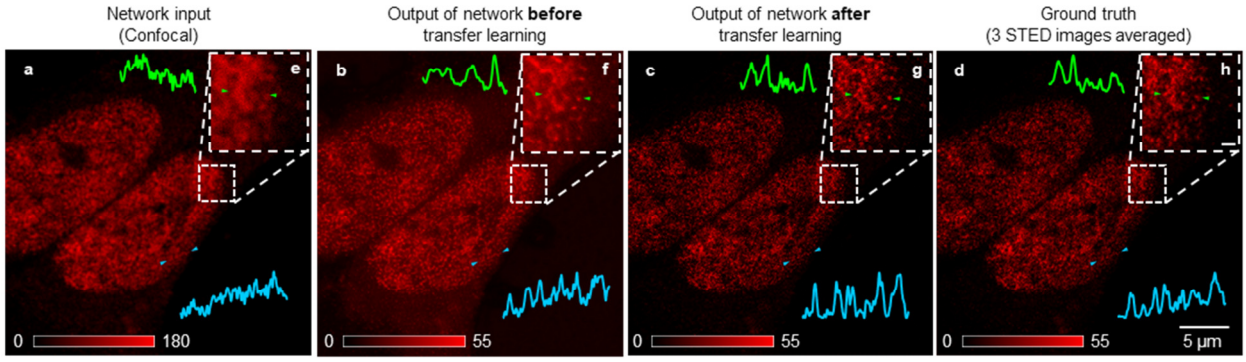


Supplementary Figure 2.

A deep neural network, trained on a specific image dataset and hardware (e.g., super-resolution model from 10×/0.4NA objective to 20×/0.75NA objective, using a Texas Red excitation/emission filter cube, trained with *only* the images of F-actin) was blindly applied on image datasets that originated from different types of objects/samples and imaging hardware.

(a-d) Wide-field images of BPAEC nuclei acquired with DAPI filter set; the input image is acquired using a 10×/0.4NA objective lens, and the 2nd and 3rd columns refer to the corresponding network output images for this input. The ground truth image is acquired using a 20×/0.75NA. (e-h) Wide-field images of BPAEC acquired with FITC filter set; the input image is acquired using a 10×/0.4NA objective lens, and the 2nd and 3rd columns refer to the corresponding network output images for this input. The ground truth

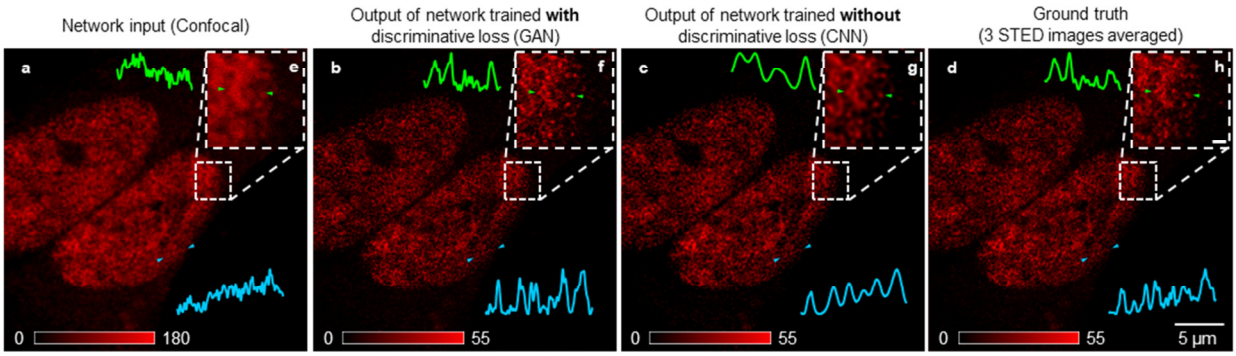
image is acquired using a 20×/0.75NA. (i-l) TIRF and TIRF-SIM images (same object as in Fig. 6). The input image is acquired using a TIRF microscope, and the ground truth image is acquired using a TIRF-SIM. (m-p) Confocal and STED images of a HeLa cell nucleus. The input image is acquired using a confocal microscope, and the ground truth image is acquired using a STED microscope. The optimal model inference (3rd column) refers to the results of the correct network model trained on the same imaging hardware as the input image. Each experiment was performed with a randomly selected image from a group of similar quality images; see the main text for further details.



Supplementary Figure 3

A neural network model trained with nano-bead images exhibits significantly improved performance in blindly inferring Histone 3 distributions within fixed HeLa cell nuclei after applying transfer learning with similar images.

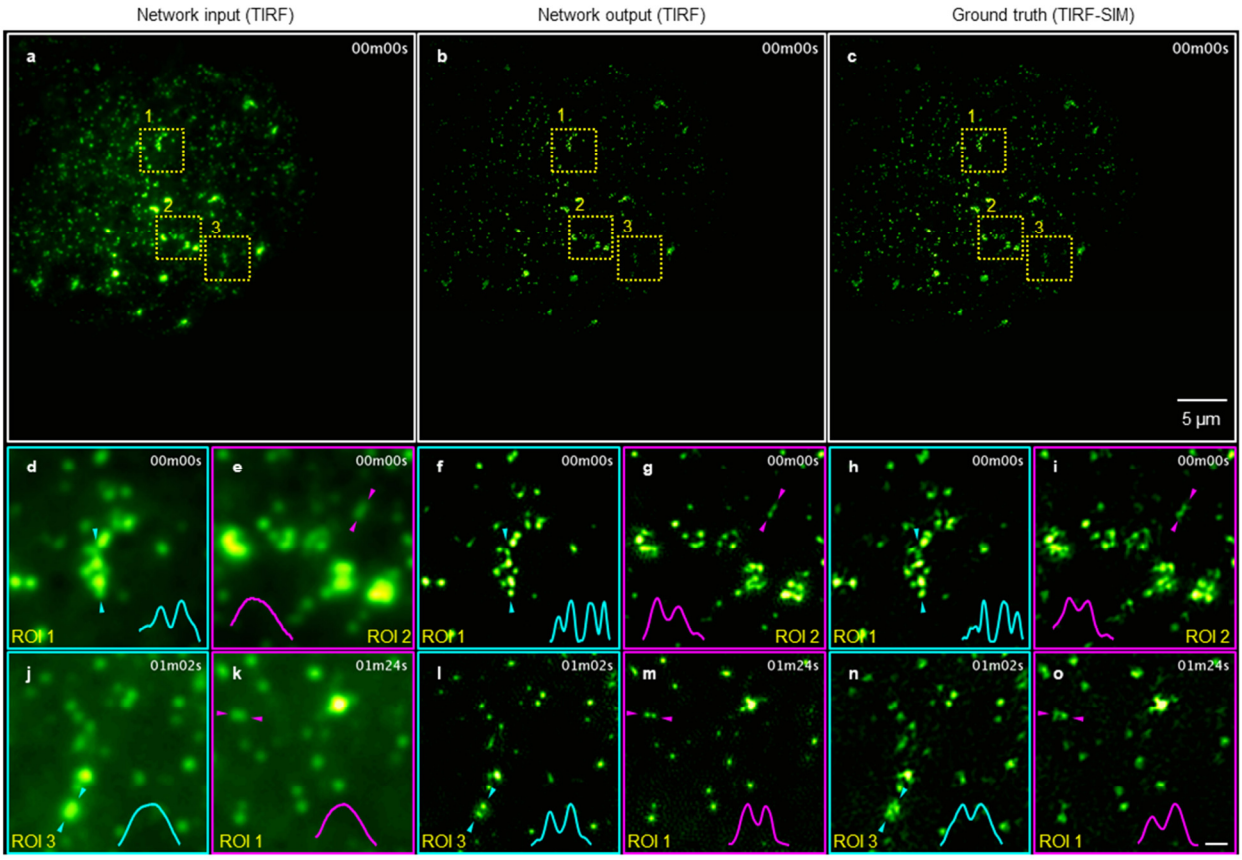
(a) Network input image captured with a confocal microscope. (b) Network inference image by a model pre-trained only with fluorescent particle images. (c) Network inference image by a model pre-trained only with fluorescent particle images and then transfer learnt with cell nuclei images. (d) The ground truth image captured with a STED microscope. (e-h) Zoomed-in regions (a-d). Scale bar in (h) is 500 nm. Arrows in each image point to the line of the shown cross-section. Also see Figure 5 of the main text. Experiment was performed on the same image as in Fig. 5, which was randomly selected from a group of 30 images with similar visual image quality.



Supplementary Figure 4

Discriminative loss is critical to the training of a generative network.

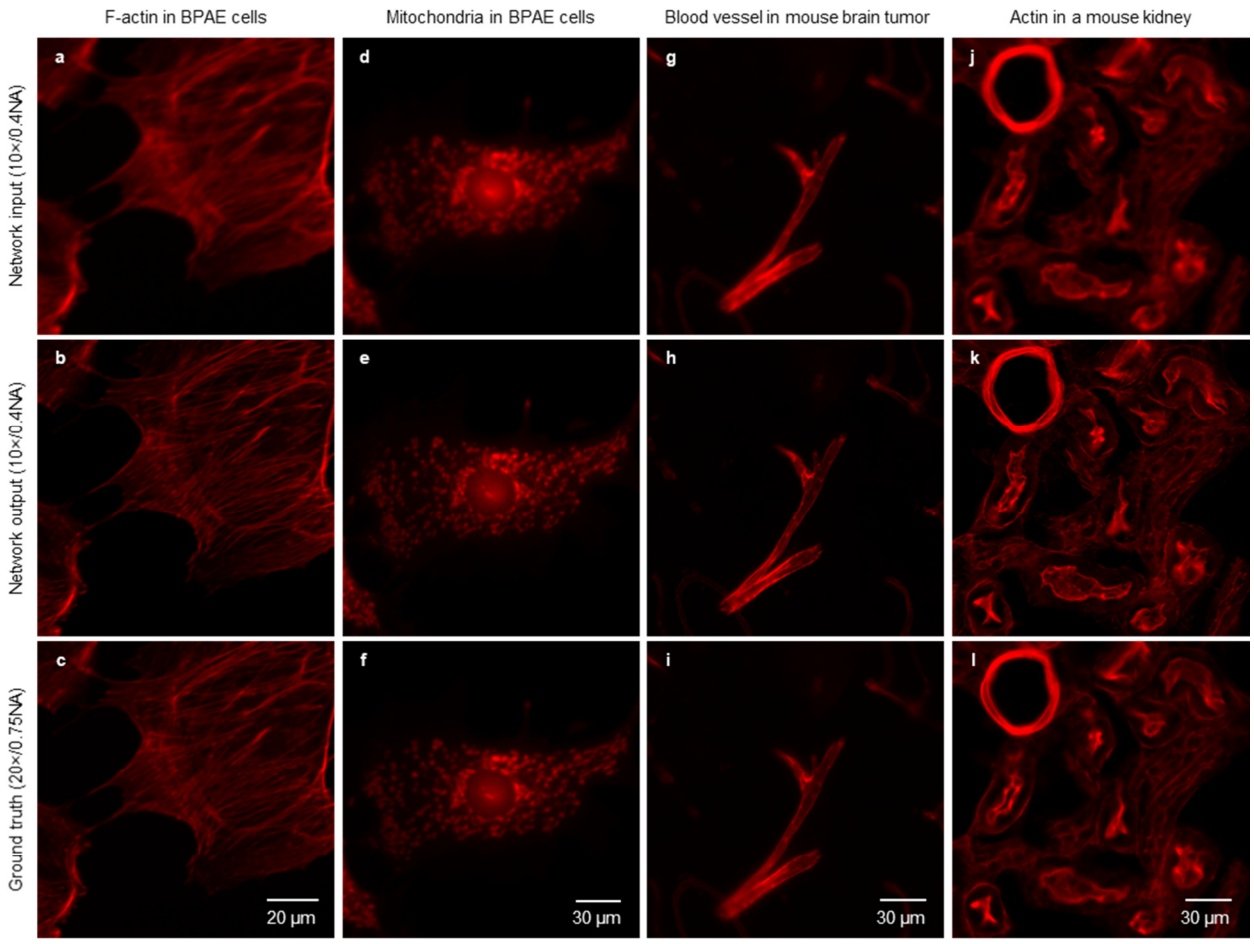
(a) Network input image captured with a confocal microscope. (b) Network inference image by the same generative model as in Fig. 5 (main text), trained *with* the discriminative loss, i.e., the GAN framework. (c) Network inference image by the same generative model as in Fig. 5 (main text), trained *without* the discriminative loss, shows compromised performance compared to (b). (d) The ground truth image captured with a STED microscope. (e-h) Zoomed-in regions (a-d). (c) and (g) show over-smoothed structures and missing details. Scale bar in (h) is 500 nm. Arrows in each image refer to the line of the shown cross-section. Also see Figure 5 of the main text. Analysis was performed on a randomly selected image from a group of 30 images with similar visual image quality.



Supplementary Figure 5

Super-resolution imaging of amnioserosa tissues of a Drosophila embryo expressing Clathrin-mEmerald using the TIRF to TIRF-SIM transformation network that was trained only with AP2 images (see Figure 6 of the main text).

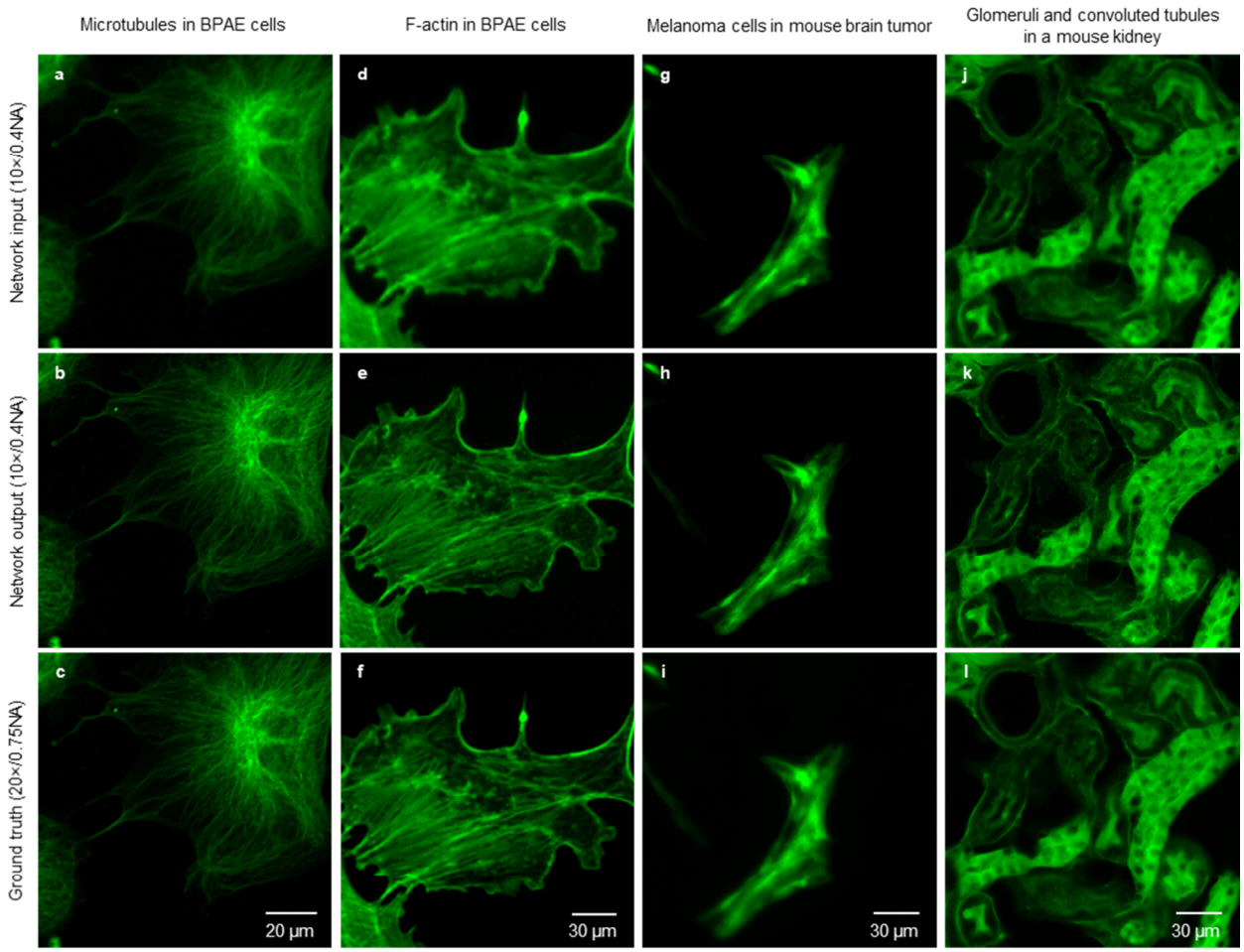
(a) Diffraction-limited TIRF image as network input. (b) Network inference image by a model-pretrained only with AP2 images. (c) Ground truth image by SIM remonstration. (d-o) Comparison of enlarged ROIs at different time points shows super-resolved details of the amnioserosa tissues. The capturing time point is labeled on the upper-right corner of each image. These results provide additional examples of the generalization of our network's inference to new sample types that it has never seen before. To position the apical surface of amnioserosa cells within the evanescent excitation field of our TIRF system, we gently pressed the dechorionated embryo against the coverglass. We attribute relatively high levels of reconstruction artifacts observed in the TIRF-SIM images to the autofluorescence of the vitelline membrane (surrounding the entire embryo) as well as the excitation/emission light scattering within amnioserosa cells that undergo rapid morphological changes during development, which negatively impacts the structured illumination/emission profiles. Scale bar in (o) is 500 nm. Arrows in each image refer to the line of the shown cross-section. Also see Figure 6 of the main text. Experiments were repeated with >1000 images with similar results.



Supplementary Figure 6

Generalization of a neural network model trained with F-actin to new types of structures that it was not trained for.

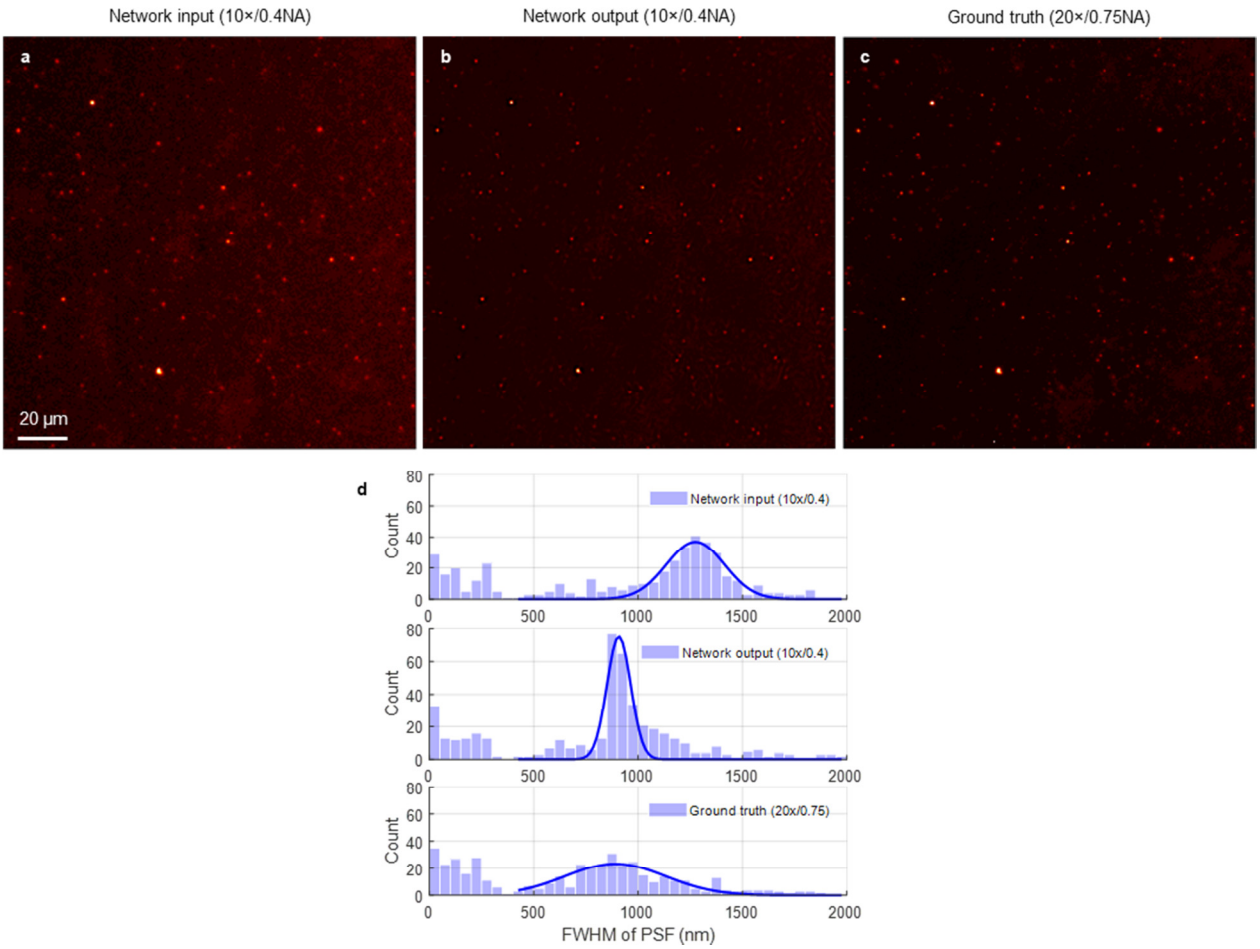
Network input, output, and ground truth images corresponding to (a-c) F-actin inside a BPAEC (image not in the training dataset), (d-f) mitochondria inside a BPAEC, (g-i) blood vessel in mouse brain tumor, and (j-l) actin in a mouse kidney section demonstrate that all these structures can be blindly super-resolved by a neural network that was trained with only F-actin images. Experiments were repeated with >50 images with similar results.



Supplementary Figure 7

Generalization of a neural network model trained with microtubules to new types of structures that it was not trained for.

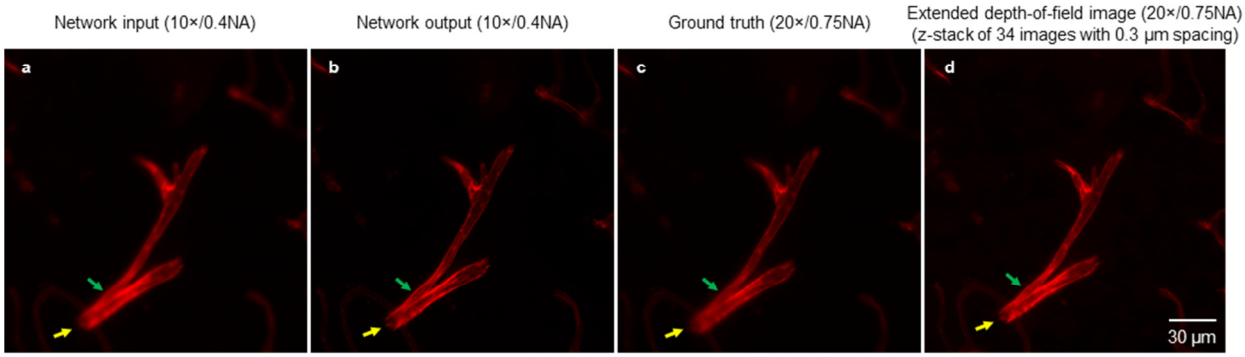
Network input, output, and ground truth images corresponding to (a-c) microtubules inside a BPAEC (image not in the training dataset), (d-f) F-actin inside a BPAEC, (g-i) melanoma cells in mouse brain tumor, and (j-l) glomeruli and convoluted tubules in a mouse kidney section demonstrate that all these structures can be blindly super-resolved by a neural network that was trained with only microtubule images. Experiments were repeated with >50 images with similar results.



Supplementary Figure 8

PSF characterization of wide-field images.

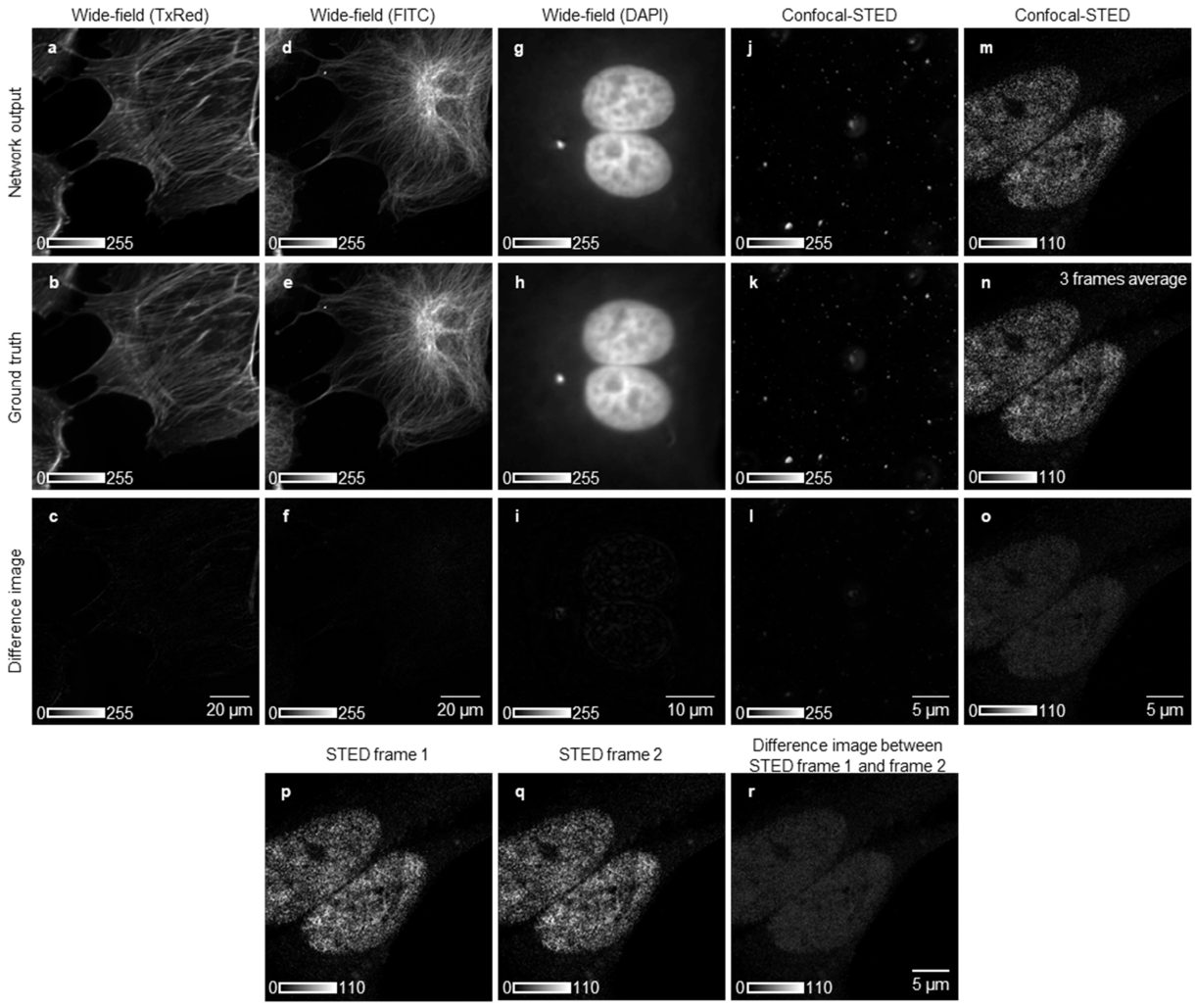
(a) An example image of 20 nm fluorescent particles captured with a 10×/0.4NA objective lens as the neural network input. (b) The network inference image with a model pre-trained with only F-actin images. (c) The ground truth image captured with a 20×/0.75NA objective lens. (d) PSF characterization, before and after the network inference, and its comparison to the ground truth image. We extracted more than 200 bright spots from the same locations of the network input (10×/0.4NA), network output (10×/0.4NA), and the corresponding ground truth (20×/0.75NA) images. Each one of these spots was fit to a 2D Gaussian function and the corresponding FWHM distributions are shown in each histogram. Analysis was performed over 3 different images randomly selected from the same nanobead sample.



Supplementary Figure 9

Demonstration of extended depth-of-focus (DOF) of our network with a mouse brain blood vessel sample.

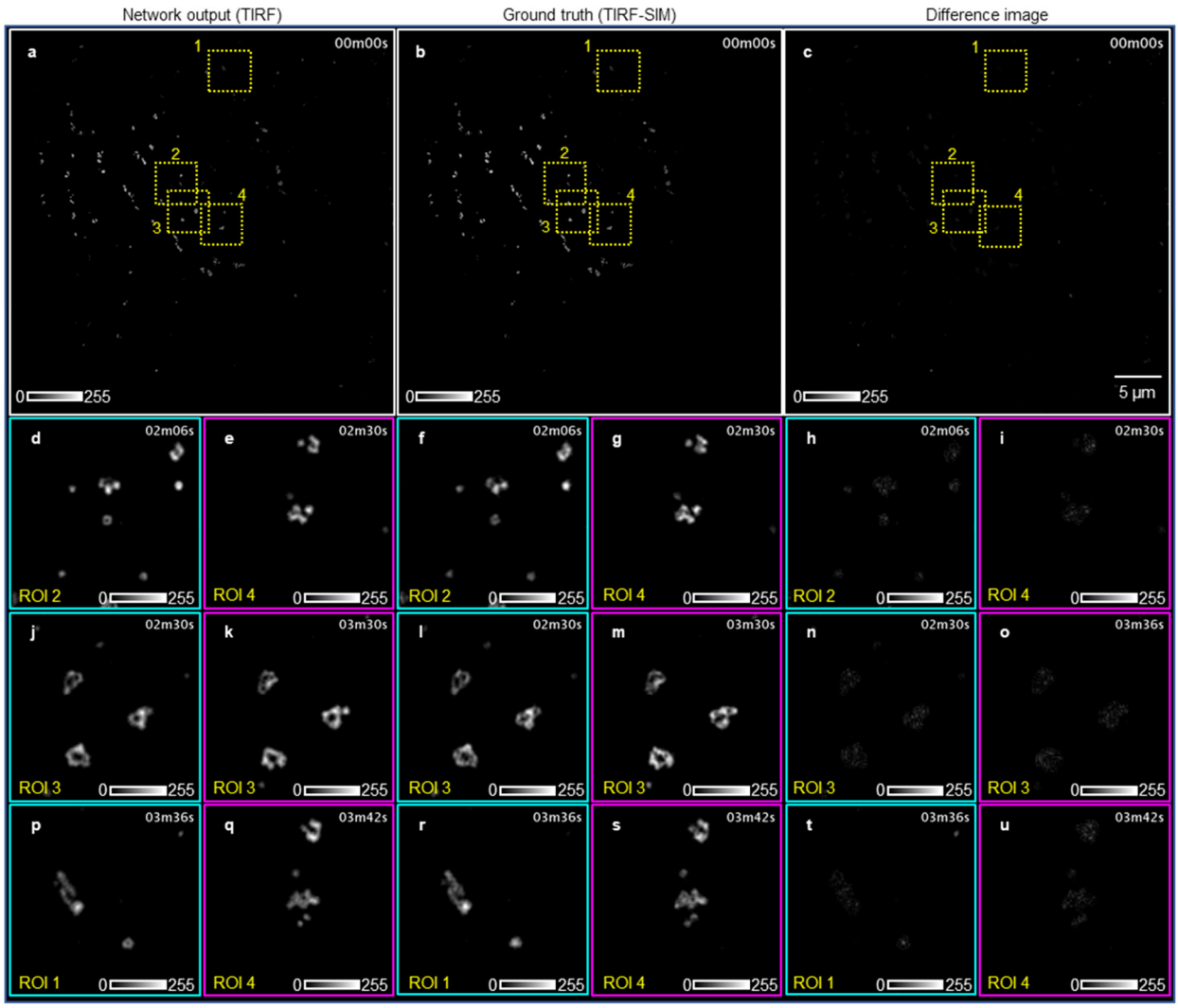
(a) The network input image captured with a 10×/0.4NA objective lens. (b) The network output image with a model pre-trained with only F-actin images. (c) High-resolution image captured with a 20×/0.75NA objective lens. The focusing plane is automatically selected by the microscope software using an auto-focusing algorithm. (d) An extended-DOF image synthesized from a z-stack of 34 high resolution images (separated axially by 0.3 µm) using ImageJ Plugin EDF[15]. The output image from a single input image is demonstrated and compared to extended-DOF image. This experiment was only performed once, as this is a unique 3D sample that better reveals the extended-DOF feature of our network. Also see **Supplementary Note 8**.



Supplementary Figure 10

The differences between the network output images and the corresponding ground truth images are shown for various imaging modalities and network models used in our manuscript (also see Figures 2, 3 and 5 of the main text).

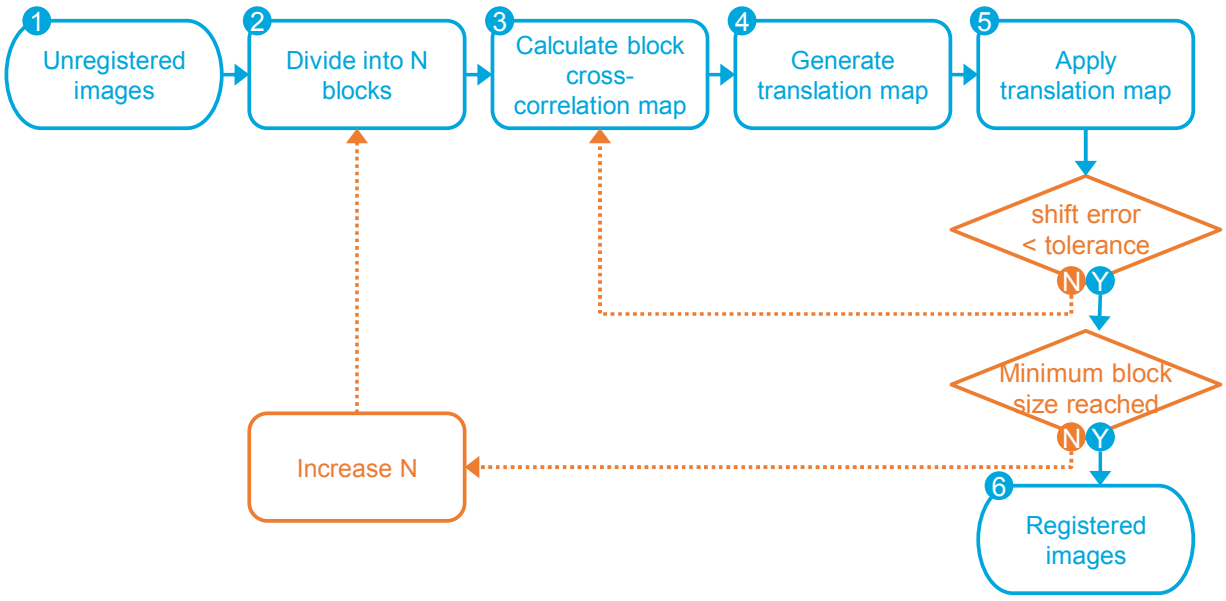
The combinations of network inferred image and ground truth images are shown as: (a-c) wide-field images of F-actin captured with TxRed filter cube set, (d-f) wide-field images of microtubules captured with FITC filter cube set, (g-i) wide-field images of HeLa cell nuclei captured with DAPI filter cube set, (j-l) confocal and STED images of 20 nm particles, and (m-o) confocal and STED images of Histone 3 sites of HeLa cell nuclei. In (p-r) we also provide the difference between two successive STED images of the same sample (HeLa cell nuclei) to show the variation in the ground truth label. The background is normalized to the same level before the two images are subtracted from one another. Our analysis was performed on the same images reported in the main text, see Figs. 2, 3, 5.



Supplementary Figure 11

The differences between the network output images (TIRF) and the corresponding ground truth images (TIRF-SIM).

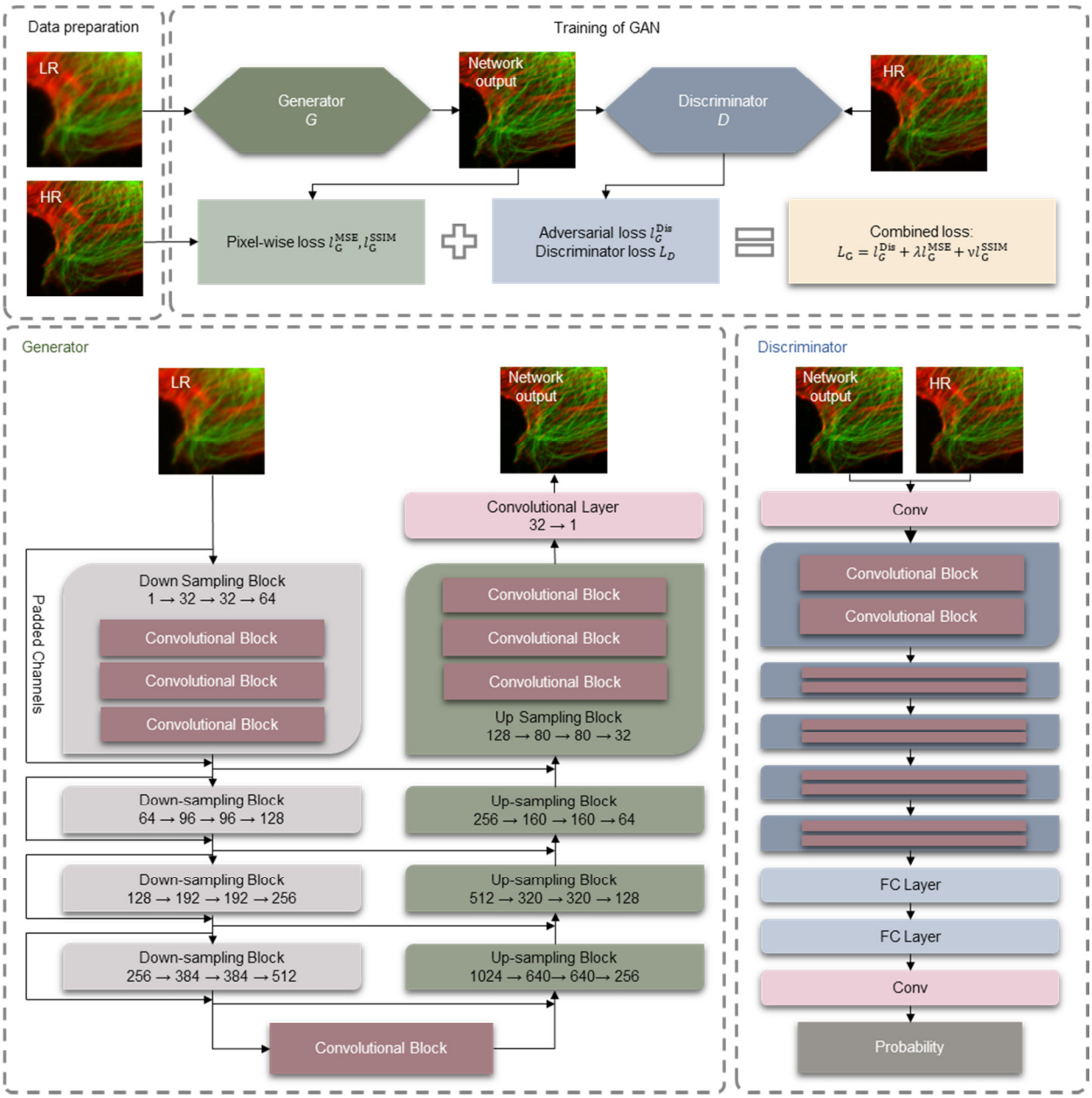
(a) Network inference image by a model-pretrained only with AP2 images. (b) Ground truth image by SIM reconstruction. (c) Difference image of (a) and (b). (d-u) Zoom-in regions of (a-c) at the labeled ROIs and time points. (also see Figure 6 of the main text). The background is normalized to the same level before the two images are subtracted from one another. Our analysis was performed on the same images reported in main text, see Fig. 6.



Supplementary Figure 12

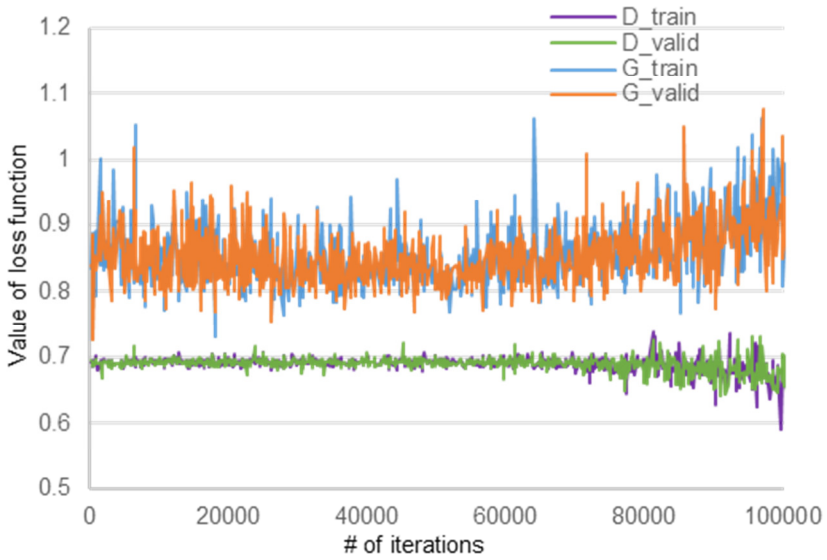
Pyramidal elastic registration workflow.

(1) The registration starts with roughly registered image pairs (e.g., 1024×1024 pixels). (2) The images are divided in to N blocks (e.g., $N = 4$). (3) The 2D cross-correlation map of each block pair from the corresponding two input images is calculated. (4) The shift of each block is calculated by fitting a 2D Gaussian function to the peak of the cross-correlation map (see **Supplementary Note 9**). This shift map (N×N) is interpolated to the image size (e.g., 1024×1024 pixels) as a translation map. (5) Apply the translation map to the image to be registered by linear interpolation. If the maximum value of the translation map is greater than the tolerance value (e.g. 0.2 pixels), repeat steps (3-5). Else if the block size is larger than the minimum block size (e.g. 64×64), increase N and shrink the block size (e.g., 1.2 times), and repeat steps (2-5). (6) When the shift error is below the tolerance, and the block size has reached the minimum set value, the input image pairs have been finely registered to each other with sub-pixel level of accuracy.



Supplementary Figure 13

The training process and the architecture of the generative adversarial network (GAN) that we used for image super-resolution.



Supplementary Figure 14

A typical plot of the loss functions of the generative and discriminative models during the GAN training.

The loss functions for the generator (G) and the discriminator (D) quickly converge to an equilibrium stage. The discriminator loss keeps stable while the generator loss slightly decreases, which means the MSE and SSIM losses that take a very small portion of the total generator loss are decreasing. In this competition process between G and D, the network gradually refines the learnt super-resolution image transformation and recovers better spatial details. After ~60,000 iterations, the discriminator takes advantage and the generator loss begins to increase, which will lead to a mode collapse in the GAN network.[16] Therefore, we take the trained model at ~50,000 iterations as the final testing model.

Super-resolution network	Number of training image pairs	Number of validation image pairs	Number of testing image pairs
Wide-field (TxRed)	1945	680	94
Wide-field (FITC)	1945	680	94
Wide-field (DAPI)	1945	680	94
Confocal-STED (nanobeads)	607	75	75
Confocal-STED (transfer learning)	1100	100	30
TIRF-SIM	3003	370	1100

Supplementary Table 1

Number of experimental image datasets used for each network. Each image has 1024×1024 pixels.

1. Farahani, J. N., Schibler, M. J. & Bentolila, L. A. Stimulated emission depletion (STED) microscopy: from theory to practice. *Microscopy: science, technology, applications and education* **2**, 1539–1547 (2010).
2. Wiener, N. *Extrapolation, interpolation, and smoothing of stationary time series, with engineering applications.* (Technology Press of the Massachusetts Institute ofTechnology, 1950).
3. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nature Methods* **9**, 676–682 (2012).
4. Sage, D. *et al.* DeconvolutionLab2: An open-source software for deconvolution microscopy. *Methods* **115**, 28–41 (2017).
5. Goodfellow, I. J. *et al.* Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]* (2014).
6. Rivenson, Y. *et al.* Deep learning-based virtual histology staining using auto-fluorescence of label-free tissue. *arXiv:1803.11293 [physics]* (2018).
7. Sønderby, C. K., Caballero, J., Theis, L., Shi, W. & Huszár, F. Amortised MAP Inference for Image Super-resolution. *arXiv:1610.04490 [cs, stat]* (2016).
8. Conditional generative adversarial nets for convolutional face generation - Semantic Scholar. Available at: https://www.semanticscholar.org/paper/Conditional-generative-adversarial-nets-for-face-Gauthier/42f6f5454dda99d8989f9814989efd50fe807ee8. (Accessed: 13th October 2018)
9. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-Image Translation with Conditional Adversarial Networks. in 5967–5976 (IEEE, 2017). doi:10.1109/CVPR.2017.632
10. Culley, S. *et al.* Quantitative mapping and minimization of super-resolution optical imaging artifacts. *Nature Methods* **15**, 263–266 (2018).
11. Li, Y. *et al.* Real-time 3D single-molecule localization using experimental point spread functions. *Nature Methods* (2018). doi:10.1038/nmeth.4661
12. Rivenson, Y. *et al.* Deep Learning Enhanced Mobile-Phone Microscopy. *ACS Photonics* **5**, 2354–2364 (2018).

13. Culley, S. *et al.* Quantitative mapping and minimization of super-resolution optical imaging artifacts. *Nature Methods* **15**, 263–266 (2018).

14. Lewis, J. P. Fast normalized cross-correlation. in **10**, 120–123 (1995).

15. Forster, B., Van De Ville, D., Berent, J., Sage, D. & Unser, M. Complex wavelets for extended depth-of-field: a new method for the fusion of multichannel microscopy images. *Microsc. Res. Tech.* **65**, 33–42 (2004).

16. Metz, L., Poole, B., Pfau, D. & Sohl-Dickstein, J. Unrolled Generative Adversarial Networks. *arXiv:1611.02163 [cs, stat]* (2016).