SUPPLEMENTARY INFORMATION

Fast super resolution optical fluctuation imaging using transformer-optimized neural network

Zitong Ye,[a] Yuran Huang,[a] Hanchu Ye,[a] Enxing He,[a] Yile Sun,[a] Haoyu Zhou,[a] Xin Luo,[a] Yubing Han,[a,c,*] Cuifang Kuang,[a,b,d,*] Xu Liu[a,b]

[a] State Key Laboratory of Extreme Photonics and Instrumentation, College of Optical Science and Engineering, Zhejiang University, Hangzhou, Zhejiang 310027, China

[b] ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou 311200, China

[c] Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China

[d] Collaborative Innovation Center of Extreme Optics, Shanxi University, Taiyuan, Shanxi 030006, China

[*] *Correspondence: cfkuang@zju.edu.cn; hanyubing@zju.edu.cn.*

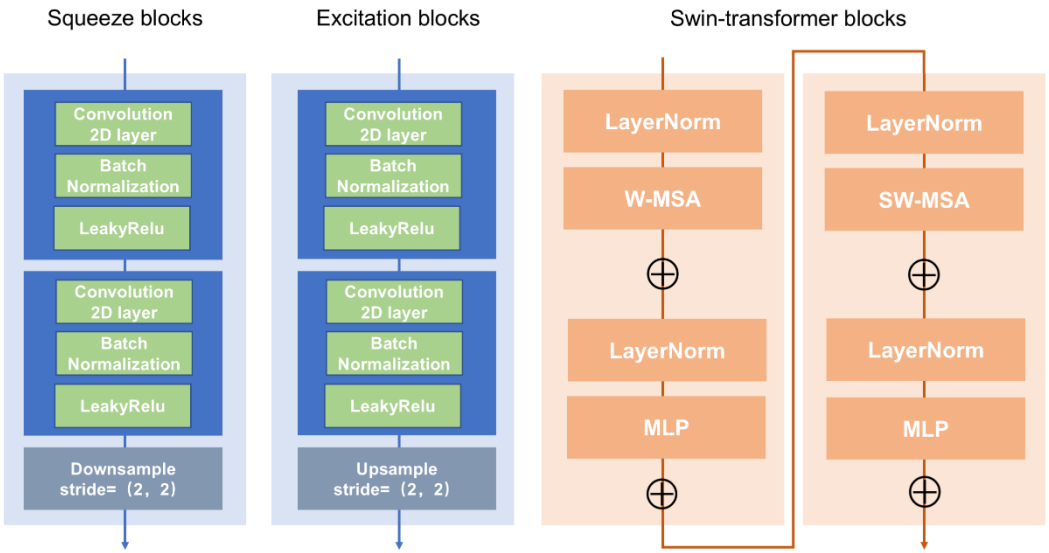


Fig. S1. The proposed network architecture integrates three principal components: Squeeze-and-Excitation blocks for channel-wise attention, Swin Transformer blocks for spatial feature learning, and a hierarchical convolutional backbone.The convolutional base progressively increases feature channels through consecutive layers with dimensions [64, 128, 256, 512]. The Swin Transformer module employs shifted window partitioning with 11×11 window dimensions, implementing an efficient self-attention mechanism through 8 parallel computation heads.

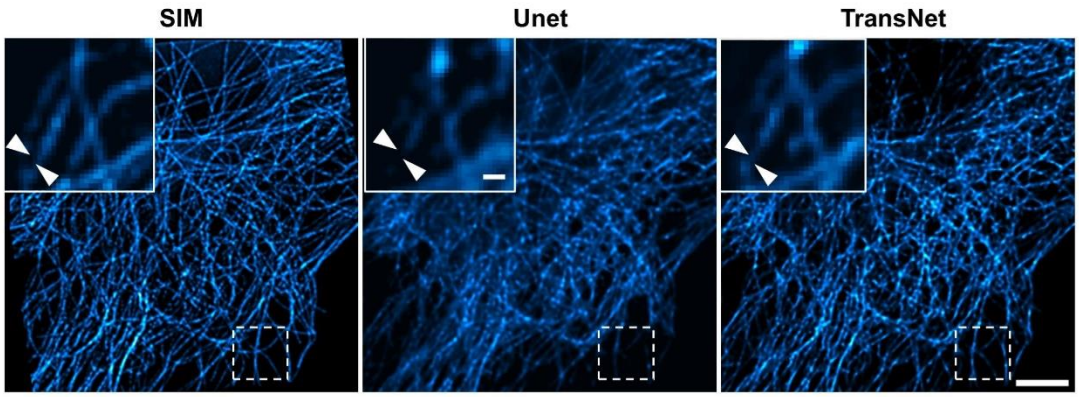


Fig. S2. Comparision of TRUS performance using different network structure. SIM served as the reference. Magnified regions of interest (ROIs), delineated by white dashed boxes, are displayed in the upper-left corner, with white arrows highlighting areas where the U-Net architecture failed to reconstruct fine anatomical details. Scale bars: 500 nm (middle top), 5 μm (right).

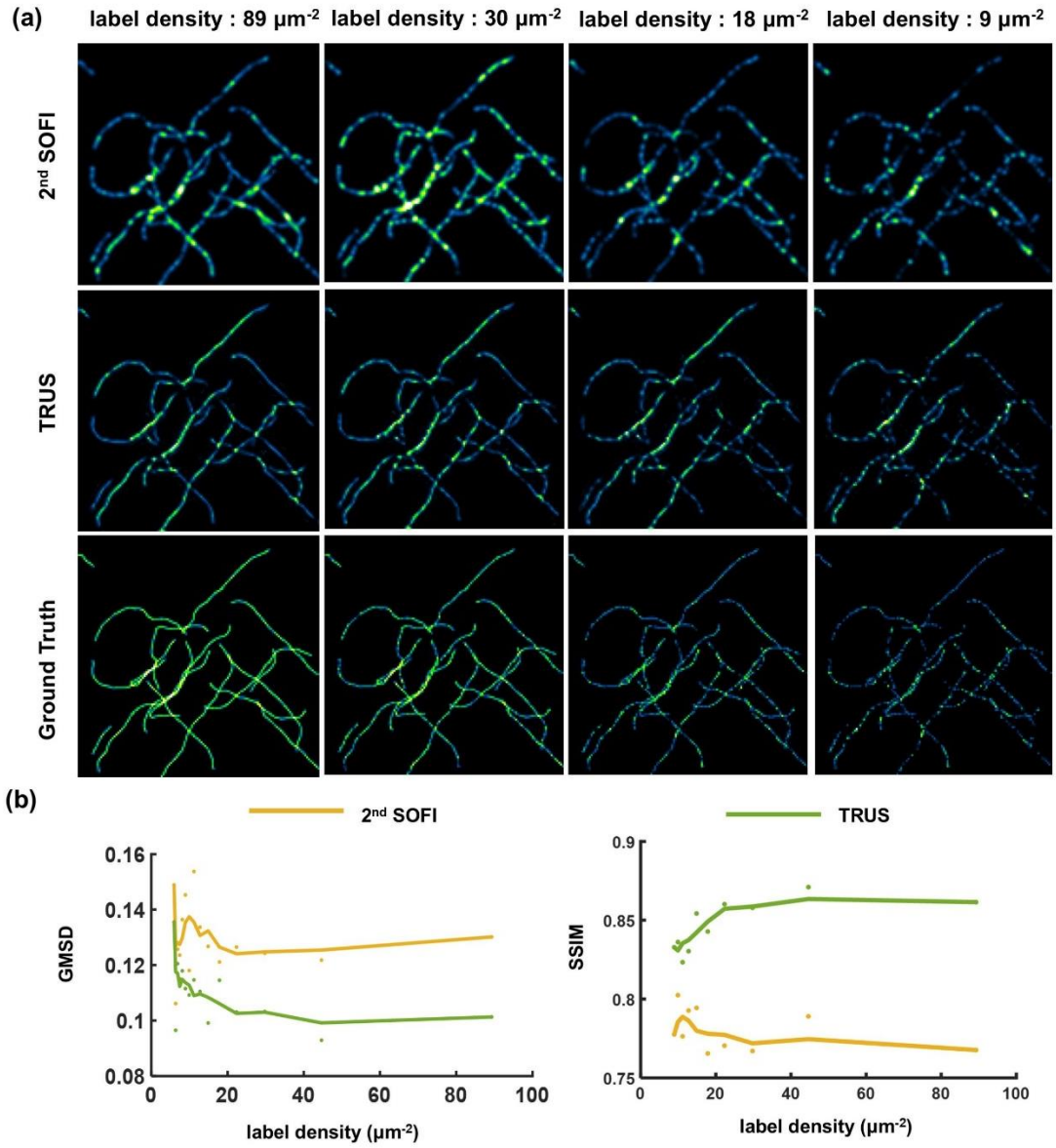


Fig. S3. (a) Comparison of second-order SOFI (top row), TRUS (middle row), and ground truth (bottom row) at sequentially decreasing label densities (left to right). (b) GMSD and SSIM metrics for second-order SOFI and TRUS versus increasing label density.

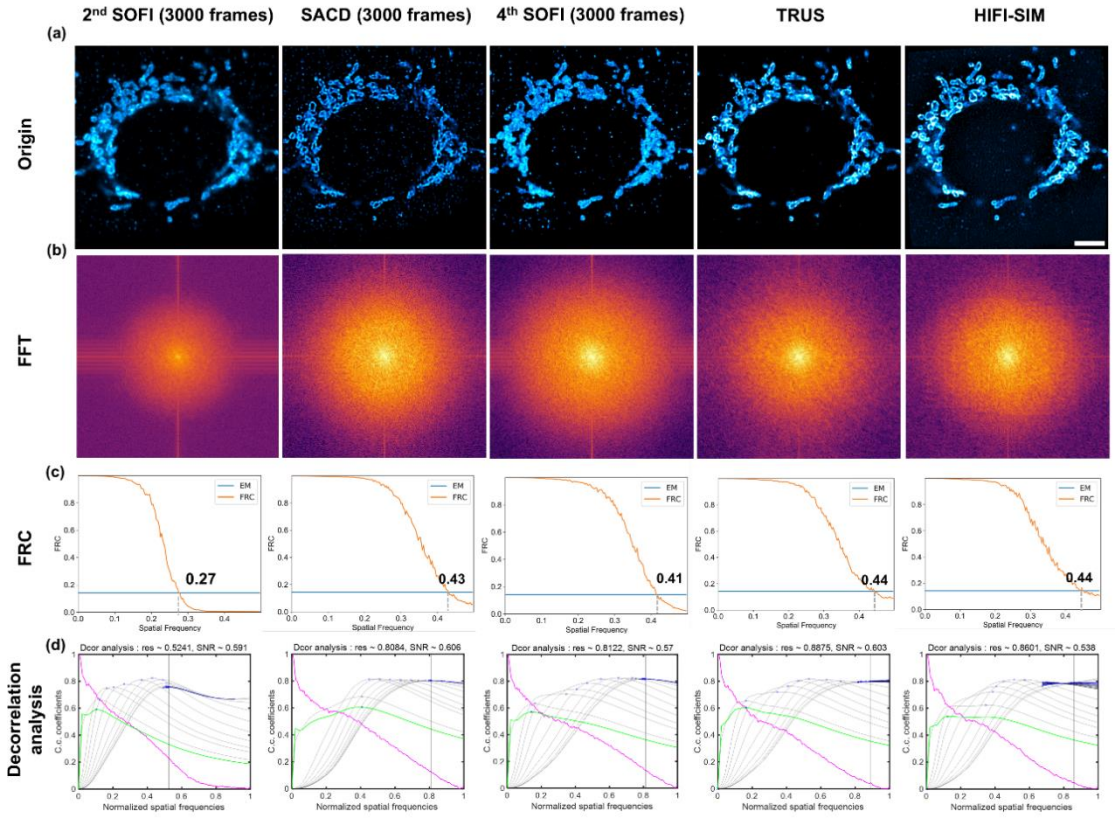


Fig. S4. (a) Mitochondrial outer membrane in a COS-7 cell labeled with Alexa Fluor 647, imaged by second-order SOFI, SACD (reconstructed from 3000 frames), fourth-order SOFI, TRUS, and corresponding HIFI-SIM. (b) Spatial frequency spectra corresponding to (a). (c) Fourier ring correlation (FRC) analysis results with threshold = 0.143 (cut-off frequency defined at FRC-threshold intersection). (d) Decorrelation coefficient profiles. Scale bar: 5 μm.

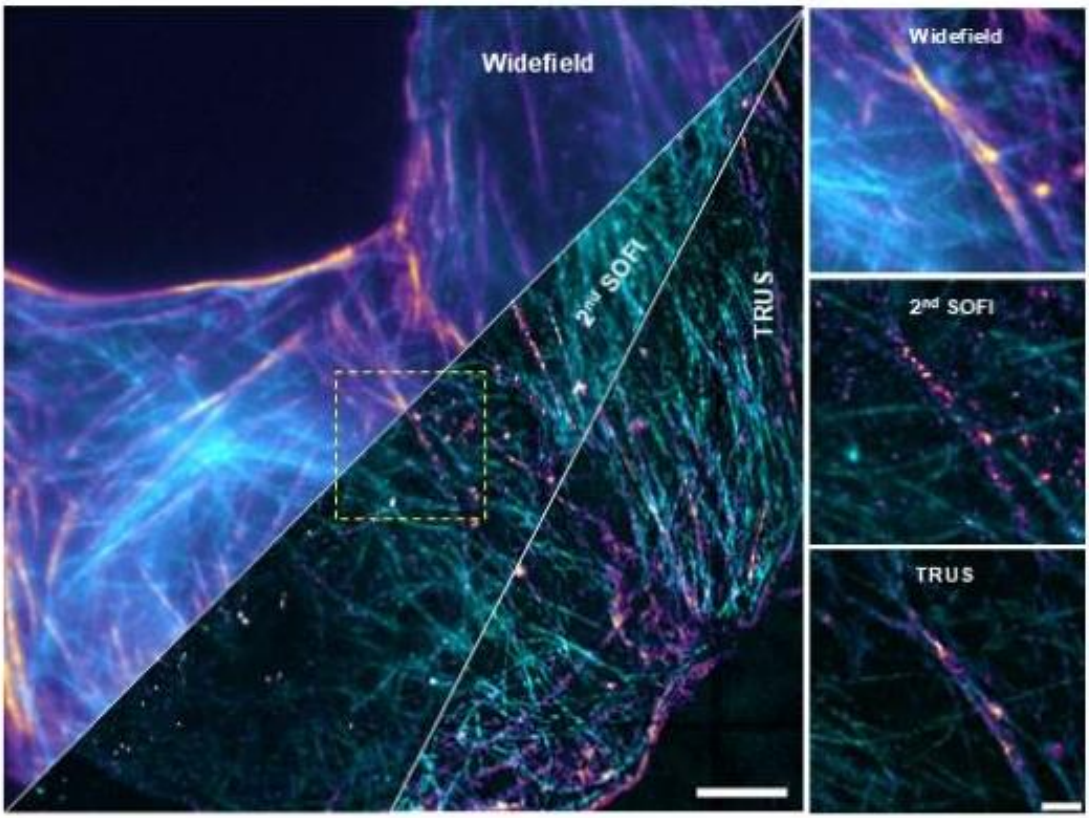


Fig. S5. Dual-color imaging with TRUS. Widefield images and results reconstructed by second-order SOFI (reconstructed from 1000 frames), and TRUS of microtubule (cyan) labeled with Alexa Fluor 647 and actin (magenra) labeled with Alexa Fluor 488. Scale bars: 5 μm in (**a**), 2 μm in (**b**).

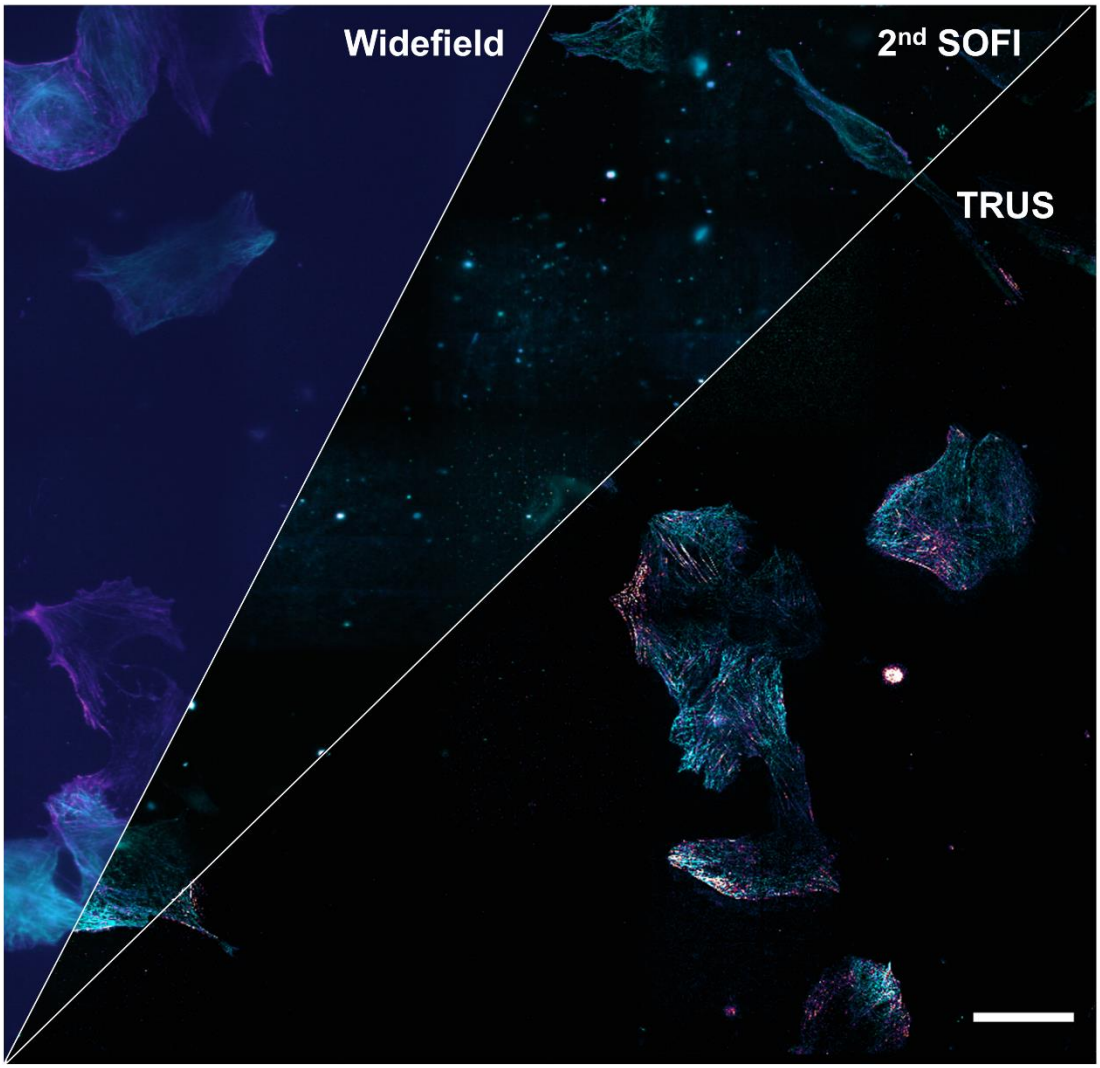


Fig. S6. Dual-color high though-put imaging. Application of TRUS to high-throughput SR imaging of an ~0.25 mm^2 area. Microtubule (cyan) labeled with Alexa Fluor 647 and actin (magenra) labeled with Alexa Fluor 488 from COS-7 cells. Scale bar: 50 um.

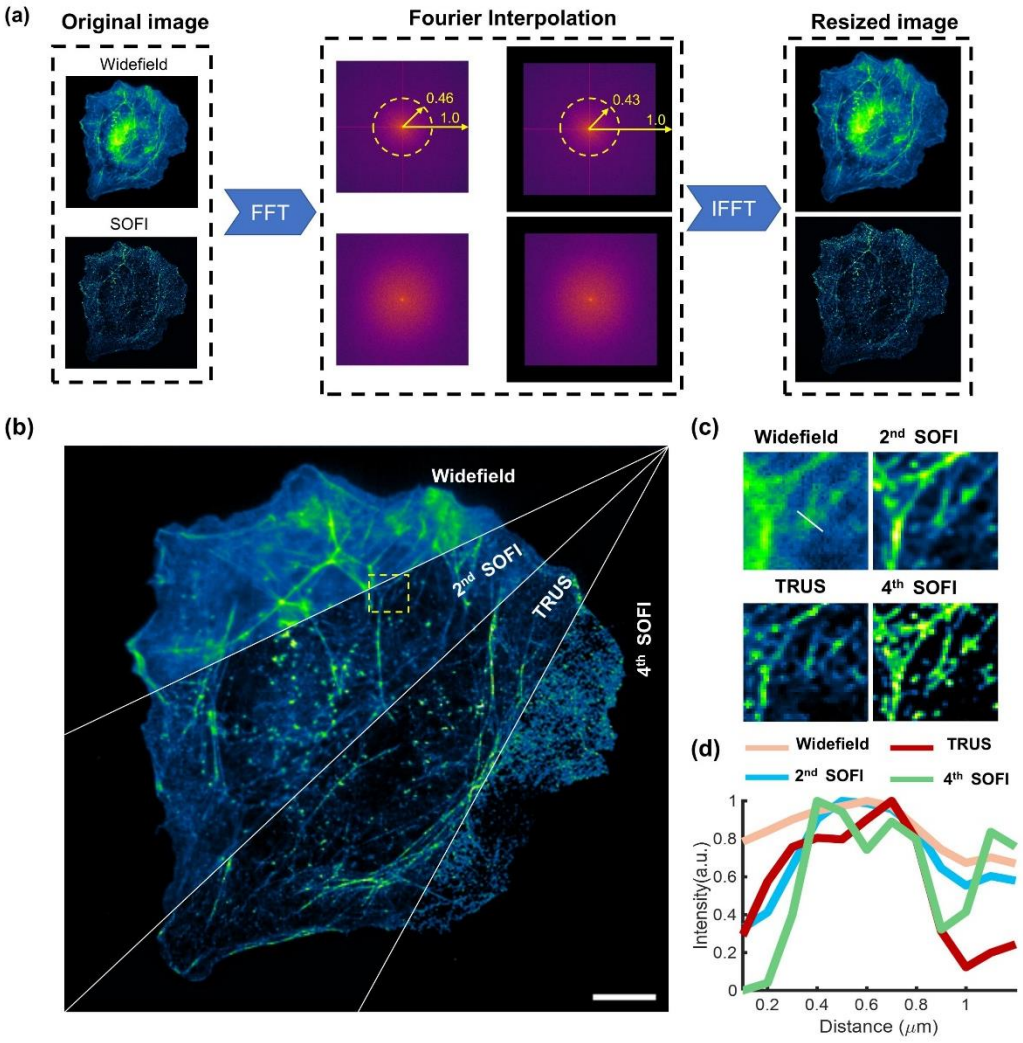


Fig. S7. Validation on data with different ideal numerical aperture (a) The process of Fourier interpolation for input data with different PSF. (b) Comparision of widefield, second-order SOFI, TRUS with corresponding ground truth image on F-actin labeled with. (c) Maginified view of yellow dash box region in (a). (d) Intensity profiles of white line in (c). Scale bar: 8 μm

Supplementary Note 1: Network architechture

Experimentally, we employ a network structure comprising four squeeze blocks, four excitation blocks, and two swin transformer blocks. The detailed configuration for TransNet is provided in Table S1. Regarding model complexity, TransNet contains approximately 12M parameters. For an input tensor of size 256×256×2, the estimated model size is 8719 MB. In terms of processing speed, inference on a 256×256×2 input takes approximately 2 seconds on CPU and ~0.2 seconds on GPU.

Table. S1 TransNet architectures

layer	Output size	Kernel size
Double Convolutional layer (1)	128×128×64	[3 × 3 conv] × 2
Convolution layer (1)	64×64×64	[3 × 3 conv] × 1, stride=2
Double Convolutional layer (2)	64×64×128	[3 × 3 conv] × 2
Convolution layer (2)	32×32×128	[3 × 3 conv] × 1, stride=2
Double Convolutional layer (3)	32×32×256	[3 × 3 conv] × 2
Convolution layer (3)	16×16×256	[3 × 3 conv] × 1, stride=2
Double Convolutional layer (4)	16×16×512	[3 × 3 conv] × 2
Convolution layer (4)	8×8×512	[3 × 3 conv] × 1, stride=2
Single Convolutional layer (1)	8×8×128	[3 × 3 conv] × 1
Swin-transformer blocks (1)	8×8×128	
Swin-transformer blocks (2)	8×8×128	
Single Convolutional layer (2)	8×8×512	[3 × 3 conv] × 1
ConvTranspose layer (1)	16×16×512	[4 × 4 conv] × 1, stride=2
Double Convolutional layer (6)	16×16×256	[3 × 3 conv] × 2
ConvTranspose layer (2)	32×32×256	[4 × 4 conv] × 1, stride=2
Double Convolutional layer (7)	32×32×128	[3 × 3 conv] × 2
ConvTranspose layer (3)	64×64×128	[4 × 4 conv] × 1, stride=2
Double Convolutional layer (8)	64×64×64	[3 × 3 conv] × 2
ConvTranspose layer (4)	128×128×64	[4 × 4 conv] × 1, stride=2
Double Convolutional layer (9)	128×128×2	[3 × 3 conv] × 2
ConvTranspose layer (5)	256×256×2	[4 × 4 conv] × 1, stride=2
Double Convolutional layer (9)	256×256×2	[3 × 3 conv] × 2
Single Convolutional layer (3)	256×256×1	[1 × 1 conv] × 1
Sigmoid	256×256×1	

Supplementary Note 2: Comparition of TRUS with different input combinations

Within the TRUS computational framework, we experimentally validated that the performance of TRUS with several combinations of inputs. Fig S7 highlights the limitations of reconstructing TRUS images using sparse SOFI data alone, which fails to resolve accurate structural features. While widefield microscopy reconstruction produces clear, continuous filaments with partial correspondence to the reference structured illumination microscopy (SIM) image (Fig. S8(a)) in isolated regions, significant artifacts emerge in densely structured areas. These include erroneous displacements, spurious splitting, and artificial merging of filamentous structures (Fig. S8(b), white arrow, bottom panel). In stark contrast, TRUS reconstructions integrating both sparse SOFI and widefield inputs achieve markedly improved fidelity, resolving sharp, continuous filaments that align closely with the ground-truth SIM reconstruction (Fig. S8(a)).

To further quantitatively evaluate the performance of the three input strategies, we performed an ablation study on the test dataset. Specifically, we systematically compared the results obtained under three distinct training configurations: (i) The network was trained excluding the SOFI result (reconstructed from 20 frames), (ii) The network was trained excluding the widefield image, and (iii) The network was trained utilizing both the widefield and SOFI images. Performance was evaluated across a set of 6 image pairs, with the network output compared against the HIFI-SIM result as the ground truth reference. As shown in Table S2, the TRUS model trained with both widefield and SOFI inputs demonstrates superior performance compared to the other two strategies, achieving the highest PSNR value.

Table. S2 Evaluation of the average PSNR of the TRUS results with different input combination on biological samples. (n=6)

	Only widefield	Only SOFI	Widefield and SOFI
PSNR	23.6	21.7	**24.0**

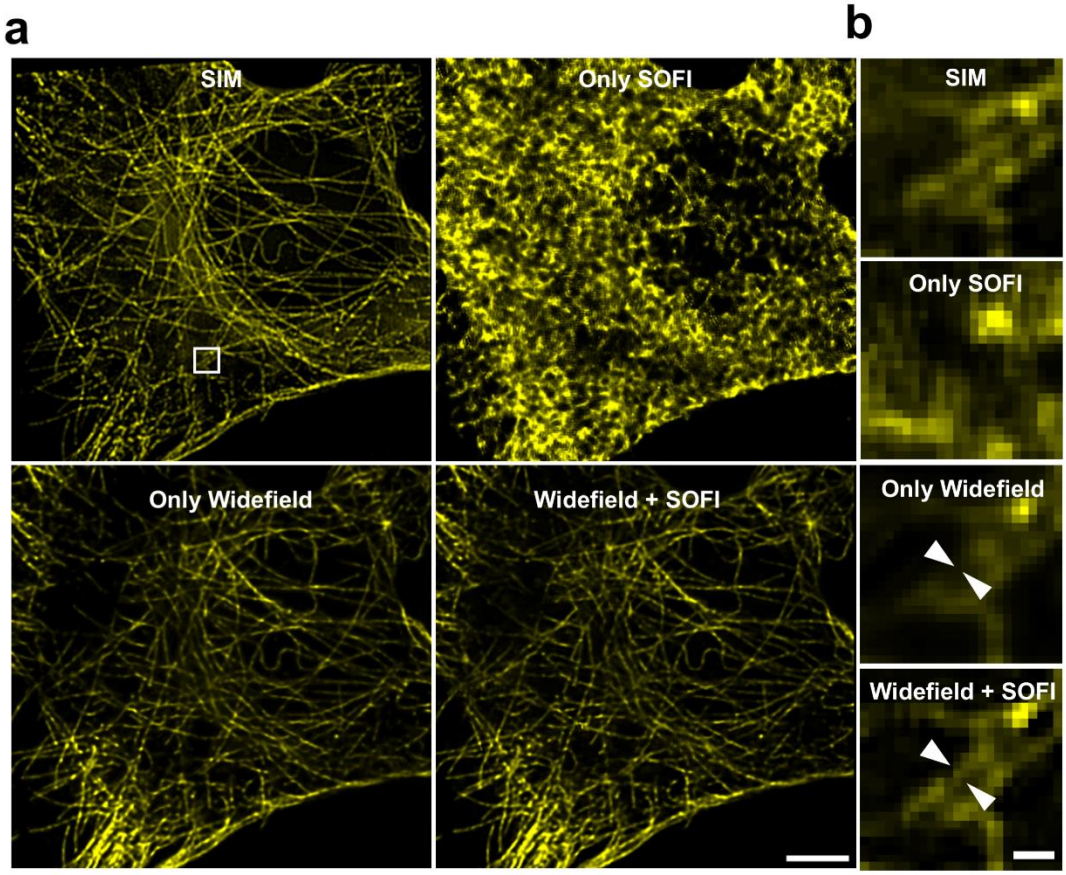
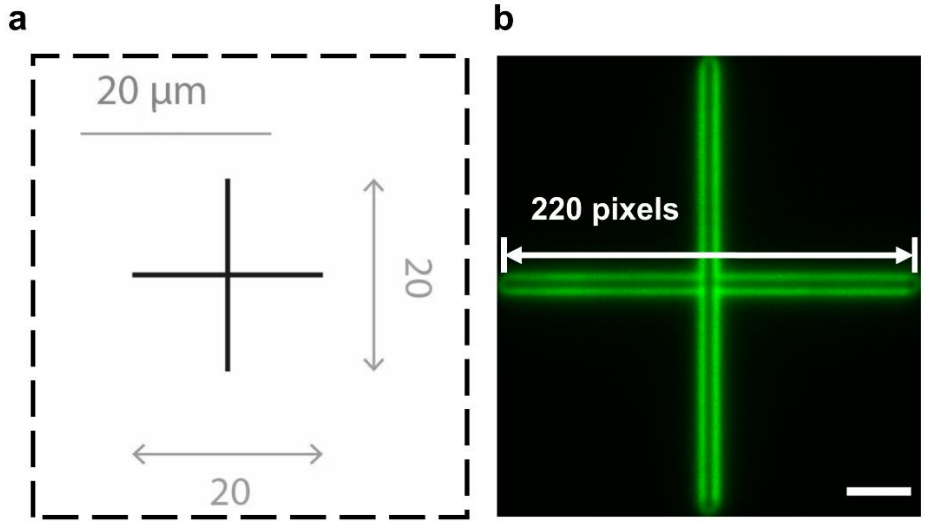


Fig. S8. Comparition of TRUS with different input combinations. (a) SIM image and corresponding TRUS reconstruction from the sparse SOFI only, widefield image only and widefield image and sparse SOFI combined. (b) Magnified view of white box region in (a). Scale bars: 5 μm in (a), 500 nm in (b).

Supplementary Note 3: Optical System configuration

Fluorescence emission was collected through a high-numerical-aperture oil immersion objective (Nikon CFI Apo TIRF 100×, NA 1.49) for simultaneous dual-channel detection. The separated signals were recorded using two scientific complementary metal-oxide-semiconductor (sCMOS) cameras (Hamamatsu ORCA-Fusion BT, Model C15440-20UP) operating in synchronized acquisition mode. Spatial calibration was performed using a fluorescent calibration standard (Argo-SIM V2.0), establishing an effective pixel size of 90 nm in object space (Fig.S9).



.

Fig. S9. Pixel size calibration. (a) Schematics of one of the repositioning crosses. All dimensions are in μm. (b) Corresponding widefield image of patten in (a). Scale bar: 3 um in (b).

Supplementary Note 4: Simulations

In this section, we use simulation to verify the ability of the TRUS to correctly reconstruct SOFI image on broad emitter dynamic range. For this purpose, we first generated fiber-like structure as shown in Fig. 3. The blinking dynamic is simulated using SOFI Simulation tool[1].T And the widefield image was simulated from ground-truth image convolved with simulated point spread function (PSF). The PSF was calculated from equation S1:

$$PSF(r) = abs(\frac{2 \bullet J_0(2\pi \bullet NA \bullet r / \lambda)}{r})^2 \tag{S1}$$

The r represents the distance between the point with center. And NA is numerical aperture and λ is the wavelength. Given the observed discrepancies between the theoretical point spread function (PSF) and experimental measurements (Fig. S10), we implemented decorrelation analysis[2] to derive the empirical PSF.

The $K_{c\max}$ representing the cut-off frequency, calculated by following steps: Firstly, compute fast fourier transform of input image $I(r)$ and define as $I(k)$. Then, generating a series of binary mask $M_i(k)$, $i = [1,2,.....,m]$ with gradually increasing radius r_i by steps of $1/m$ and Gaussian high-pass filters defined as $H_j(k)$, $j = [1,2,.....,n]$. Calculate the Pearson correlation factor $d_{ij}(r_i)$ as:

$$d_{ij}(r_i) = \frac{\iint real(I(k)M_i(k)norm(H_j(k)I(k)))dk_x dk_y}{\sqrt{\iint |I(k)|^2 dk_x dk_y \iint |M_i(k)norm(H_j(k)I(k))|^2 dk_x dk_y}} \tag{S2}$$

The radius of mask $M_i(k)$ produce the maxium $d_{ij}(r_i)$ is the cut-off frequency $K_{c\max}$.

The NA, λ and pixelsize are subject to:

$$\frac{4 \bullet pixelsize \bullet NA}{\lambda} = K_{c\max} \tag{S3}$$

We used standard fluorescent sample to calibrate the cut-off frequency of our optical system as 0.452. The pixelsize was set to 60 nm, and wavelength was 640 nm, so the numerical aperture was 1.2 in simulation experiment

The blinking dynamic is described by on-time ratio defined as:

$$Ratio_{on-time} = \frac{time_{on-state}}{time_{on-state} + time_{off-state}} \tag{S4}$$

The durations of both on-state and off-state periods range from 1 to 15 frames each, resulting in an on-time ratio spanning from 0.069 to 1. The probability distribution governing the occurrence of on-states

within cycles follows a normal distribution truncated between 0 and 1. Furthermore, the maximum photon emission per emitter per single frame is capped at 1000 photons.

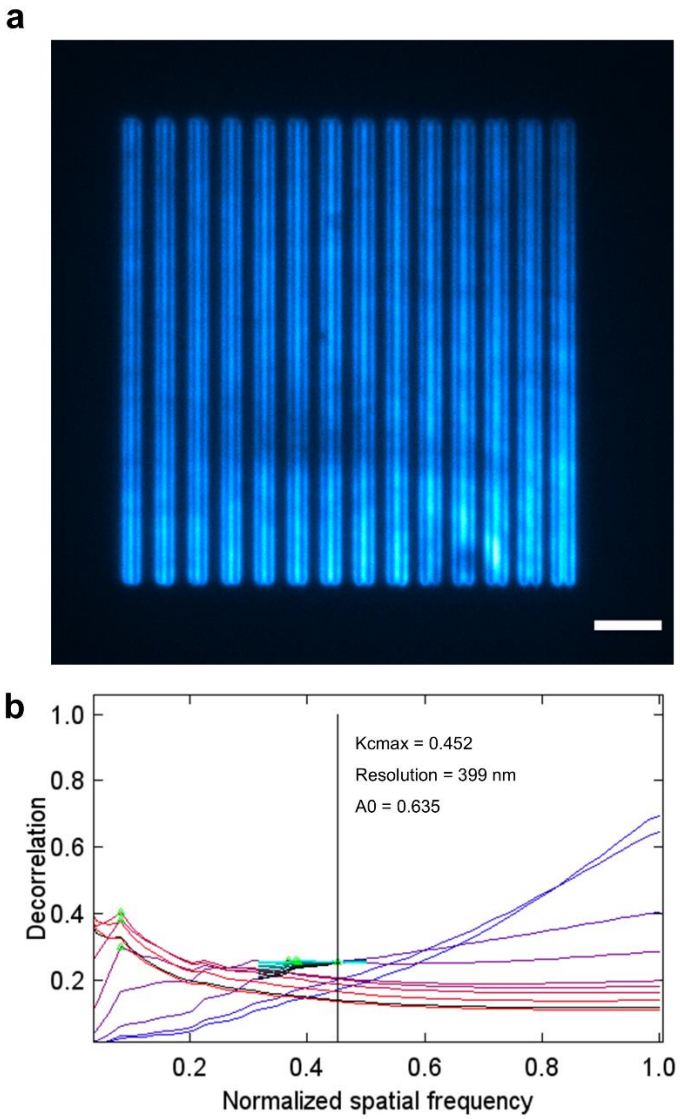


Fig. S10. Spaced fluorescent line pattern sample separated by steps of 30 nm (from 0 nm to 390 nm). Scale bar: 5 μm in (a).

Supplementary Note 5: Sample preparation

U2OS (human osteosarcoma cell line) cells were purchased from the American Type Culture Collection and cultured in McCoy's 5A medium (Cat.No:16600-082; Thermo Fisher Scientific, Inc.) supplemented with 10% (v/v) fetal bovine serum (FBS; Cat.No:10091-148; Thermo Fisher Scientific, Inc.). The cultures were maintained at 37°C in a humidified 5% CO2 environment and seeded into glass bottom dishes (Cat.No:81158; ibidi, GmbH.) at a density of 1.5~2.0 × 104 per well before labeling.

For immunofluorescence staining, the cells were fixed with 3% (m/v) paraformaldehyde (Electron Microscopy Sciences) and 0.1% (v/v) glutaraldehyde (Sigma-Aldrich Co., LLC) for 13 min at 37 °C, and quenched with $NaBH_4$ for 7 min at room temperature. Then, the cells were incubated with 0.2% (v/v) Triton X-100 (Sigma-Aldrich Co., LLC) and 5% goat serum (Thermo Fisher Scientific, Inc.) for 1 h at room temperature. The tubulin were stained with mouse anti-α tubulin (1:200 dilution; ab7291; abcam, Inc.) in PBS overnight at 4 °C and goat anti-mouse Alexa Fluor 647 (1:1000 dilution; A21236; Thermo Fisher Scientific, Inc.) in PBS for 1 h. The mitochondria were stained with rabbit anti-TOMM20 (1:100 dilution; MA5-24859; Thermo Fisher Scientific, Inc.) in PBS overnight at 4°C and goat anti-rabbit Alexa Fluor Plus 647 (1:500 dilution; A32731; Thermo Fisher Scientific, Inc.) in PBS for 1 h. Before imaging, buffer[3] was added to the samples.

For two-color sample staining, sample fixation follow the same procedure with single-color sample preparation. The tubulin and mitochondria were stained with mouse anti-α tubulin (1:200 dilution; ab7291; abcam, Inc.) mixed with rabbit anti-TOMM20 (1:100 dilution; MA5-24859; Thermo Fisher Scientific, Inc.) in PBS overnight at 4 °C Afterwards, the cells were incubated in goat anti-mouse Alexa Fluor 647 (1:1000 dilution; A21236; Thermo Fisher Scientific, Inc.) and goat anti-rabbit Alexa Fluor Plus 488 (1:500 dilution; A32731; Thermo Fisher Scientific, Inc.) mixed up in PBS for 1 h.

For actin-microtubule co-visualization, tubulin was labeled as described previously. After tubulin staining, actin was stained with 1:400 Alexa Fluor 488-phalloidin (ThermoFisher A12379) for 30 min at room temperature. Samples were washed thrice for 3, 3, and 15 min respectively with PBS on orbital shaker and imaged immediately in imaging buffer. Under high intensity laser beam, fluorophores enter a long-lived triplet state. In buffer contained MEA, thiolates (RSH/RS−) donate electrons to convert fluorophores into non-fluorescent radical anions. Stochastic reactivation occurs when residual oxygen reoxidize the radical anions, regenerating the fluorescent ground state.

The immunofluorescence protocols and imaging conditions for samples in Figs. 4-6 are detailed in Table S3.

Table. S3 Immunofluorescence Protocols for Figs

Fig caption	Primary antibodies	Secondary antibodies	Excitation Wavelength (nm)
Fig. 4	rabbit anti-TOMM20	goat anti-rabbit Alexa Fluor Plus 647	642
Fig. 5(a)	rabbit anti-TOMM20	goat anti-rabbit Alexa Fluor Plus 647	642
Fig. 5(b)	Alexa Fluor 488-phalloidin		488
Fig. 5(c)	mouse anti-α tubulin	goat anti-mouse Alexa Fluor 647	642
Fig. 6(a)	rabbit anti-TOMM20	goat anti-rabbit Alexa Fluor 488	488
Fig. 6(a)	mouse anti-α tubulin	goat anti-mouse Alexa Fluor Plus 647	642
Fig. 6(b)	mouse anti-α tubulin	goat anti-mouse Alexa Fluor Plus 647	642
Fig. 7	mouse anti-α tubulin	goat anti-mouse Alexa Fluor Plus 647	642

Supplementary Note 6: Gradient Magnitude Similarity Deviation (GMSD)

GMSD assesses structural continuity by calculating the standard deviation of the gradient magnitude similarity between a reference image and a distorted image. The mathematical formulation shows below:

$$G_{ref}(i,j) = \sqrt{(I_{ref} * h_x)(i,j)^2 + (I_{ref} * h_y)(i,j)^2} \tag{S5}$$

$$G_{dis}(i,j) = \sqrt{(I_{dis} * h_x)(i,j)^2 + (I_{dis} * h_y)(i,j)^2} \tag{S6}$$

The h_x , h_y are Prewitt operators. I_{ref}, I_{dis} are reference image and distorted image. The local similarity map is defined as:

$$S(i,j) = \frac{2 \bullet G_{ref}(i,j) \bullet G_{dis}(i,j) + C}{G_{ref}(i,j)^2 + G_{ref}(i,j)^2 + C} \tag{S7}$$

The GMSD score is defined as the standard deviation of $S(i,j)$ the over all pixels:

$$GMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{W}\sum_{j=1}^{H}(S(i,j) - \bar{S})} \tag{S8}$$

The local similarity map is relatively uniform and high, leading to a small GMSD value. In contrast, high GMSD indicates significant distortion.

References:

1. Girsault Arik, Tomas Lukes, Azat Sharipov , *et al.* SOFI Simulation Tool: A Software Package for Simulating and Testing Super-Resolution Optical Fluctuation Imaging. *PLOS ONE* **11**, e0161602 (2016).
2. Descloux A., K. S. Grußmayer, and A. Radenovic. Parameter-free image resolution estimation based on decorrelation analysis. *Nat Methods* **16**, 918-924 (2019).
3. Power Rory M., Aline Tschanz, Timo Zimmermann , *et al.* Build and operation of a custom 3D, multicolor, single-molecule localization microscope. *Nature Protocols* **19**, 2467-2525 (2024).