Spatial redundancy transformer for self-supervised fluorescence image denoising

Xinyang Li[1,2,3†], Xiaowan Hu[2†], Xingye Chen[1,3†], Jiaqi Fan[2,4], Zhifeng Zhao[1,3], Jiamin Wu[1,3,5,6*], Haoqian Wang[2,7*] & Qionghai Dai[1,3,5,6*]

[1]*Department of Automation, Tsinghua University, Beijing 100084, China.*

[2]*Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, China.*

[3]*Institute for Brain and Cognitive Sciences, Tsinghua University, Beijing 100084, China.*

[4]*Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.*

[5]*Beijing Key Laboratory of Multi-dimension & Multi-scale Computational Photography (MMCP), Tsinghua University, Beijing 100084, China.*

[6]*IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing, China.*

[7]*The Shenzhen Institute of Future Media Technology, Shenzhen 518071, China*

[†]*These authors contributed equally to this work.*

**Correspondence: wujiamin@tsinghua.edu.cn; wanghaoqian@tsinghua.edu.cn; qhdai@tsinghua.edu.cn*

Abstract

Fluorescence imaging with high signal-to-noise ratios has become the foundation of accurate visualization and analysis of biological phenomena. However, the inevitable photon shot noise poses a formidable challenge on imaging sensitivity. In this paper, we provide a spatial redundancy denoising transformer (SRDTrans) to remove noise from fluorescence time-lapse images in a self-supervised manner. First, a sampling strategy based on spatial redundancy is proposed to extract adjacent orthogonal training pairs, which eliminates the dependence on high imaging speed. Secondly, to break the performance bottleneck of convolutional neural networks (CNNs), we designed a lightweight spatiotemporal transformer architecture to capture long-range dependencies and high-resolution features at a low computational cost. SRDTrans can overcome the inherent spectral bias of CNNs and restore high-frequency information without producing over-smoothed structures and distorted fluorescence traces. Finally, we demonstrate the state-of-the-art denoising performance of SRDTrans on single-molecule localization microscopy and two-photon volumetric calcium imaging. SRDTrans does not contain any assumptions about the imaging process and the sample, thus can be easily extended to a wide range of imaging modalities and biological applications.

Introduction

The rapid development of intravital imaging techniques enables researchers to observe biological structures and activities at micrometer and even nanometer scales[1-4]. As an imaging method with great prevalence, fluorescence microscopy has contributed to the discovery of a

series of new physiological and pathological mechanisms due to its high spatiotemporal resolution and molecular specificity[5-8]. The fundamental goal of fluorescence microscopy is to obtain clean and sharp images containing sufficient information about the sample, which can guarantee the accuracy of downstream analysis and support convincing conclusions. However, limited by multiple biophysical and biochemical factors (*e.g.*, labeling concentration, fluorophore brightness, phototoxicity, photobleaching, *etc.*), fluorescence imaging is conducted in photon-limited conditions and the inherent photon shot noise degrades the image signal-to-noise ratio (SNR) severely, especially in low-illumination and high-speed observations[9-11].

Various methods have been proposed to remove noise from fluorescence images. Conventional denoising algorithms based on numerical filtering and mathematical optimization suffer from unsatisfactory performance and limited applicability[12-14]. In the past few years, deep learning has shown remarkable performance in image denoising[15-17]. After iterative training on a dataset with ground truth (GT), deep neural networks can learn the mapping between noisy images and their clean counterparts. Such a supervised manner depends heavily on paired GT images[18-22]. When observing the activity of living organisms, obtaining pixel-wise registered clean images is a great challenge because the sample often undergoes fast dynamics. To alleviate this contradiction, some self-supervised methods have been proposed for more applicable and practical denoising in fluorescence imaging[9, 23-29]. Among them, the first kind of methods rely on the similarity between adjacent frames[9,23-25]. But when the sample changes very quickly or the imaging speed is too slow, the time-lapse data cannot provide enough temporal redundancy. This is a common problem in volumetric imaging

since the volume rate decreases proportionally to the number of imaging planes. The dissimilarity between adjacent frames will result in inferior performance and distorted structures and fluorescence kinetics. The other kind of methods learn to denoise only from spatially adjacent pixels in 2D frames[26-29]. However, without utilizing endogenous temporal correlations, these methods perform poorly on time-lapse imaging. Therefore, to achieve better denoising performance, the ability to simultaneously extract global spatial information and long-range temporal correlations is essential, which is lacking in CNNs because of the locality of convolutional kernels[30,31]. Moreover, the inherent spectral bias makes CNNs tend to fit low-frequency features preferentially while ignoring high-frequency features, inevitably producing over-smoothed denoising results[32, 33].

Here, we present the spatial redundancy denoising transformer (SRDTrans) to address these dilemmas. On the one hand, a spatial redundancy sampling strategy is proposed to extract three-dimensional (3D) training pairs from the original time-lapse data in two orthogonal directions. This scheme has no dependence on the similarity between two adjacent frames, so SRDTrans is applicable to very fast activities and extremely low imaging speed, which is complementary to our previously proposed DeepCAD that leverages temporal redundancy[9,26]. On the other hand, we designed a lightweight spatiotemporal transformer network to fully exploit long-range correlations. The optimized feature interaction mechanism allows our model to obtain high-resolution features with a small number of parameters. Compared with classical CNNs, the proposed SRDTrans has stronger abilities for global perception and high-frequency maintenance, enabling the revelation of fine-grained spatiotemporal patterns that were

previously indiscernible. We demonstrate the superior denoising performance of SRDTrans on two representative applications. The first one is single-molecule localization microscopy (SMLM) with adjacent frames being random subsets of fluorophores[34,35]. The other one is two-photon calcium imaging of large 3D neuronal populations with a volumetric speed as low as 0.3 Hz. Extensive qualitative and quantitative results indicate that SRDTrans can serve as a fundamental denoising tool for fluorescence imaging to observe various cellular and subcellular phenomena.

Results

Principle of SRDTrans. The self-supervised framework of SRDTrans is schematized in Fig. 1a. For spatial redundancy sampling, spatially adjacent sub-stacks are sampled by orthogonal masks from the original low-SNR image stack. Each target is adjacent to the input stack horizontally or vertically to exploit spatial correlations fully and isotropically. A simplified implementation of sampling is depicted in Fig. 1b. Since the noise of adjacent pixels is independent while the signals are closely correlated, the sub-stack filled by the central pixels can be used as the training input and the other two spatially adjacent sub-stacks are used as corresponding targets to optimize the network parameters (Supplementary Note 1). Compared with other sampling methods[26-28], our sampling strategy is more effective and comprehensive in preserving both spatial and temporal information (Supplementary Fig. 1). In the inference stage, low-SNR stacks will be fed into pre-trained SRDTrans models without spatial down-sampling. To overcome the locality of convolutional kernels, we constructed a transformer network to capture endogenous non-local spatial features and long-range temporal

dependencies using the self-attention mechanism (Fig. 1c). The restoration of each pixel can simultaneously integrate the temporal information of all frames and the spatial information of all pixels, even if they are far from each other. Besides, the network does not contain any spatial down-sampling module, allowing more high-frequency components to flow through the network and avoiding the loss of spatial resolution. Furthermore, as the amount of data in fluorescence time-lapse imaging is often very large, sometimes at the scale of petabyte, the transformer network was designed to be as lightweight as possible to relieve the computational burden of large-scale data processing. Compared with other transformer networks[36-39], our architecture can achieve the best performance with more than one order of magnitude fewer parameters (Supplementary Tab. 1 and Supplementary Tab. 2). The lightweight architecture of SRDTrans also makes it easy to train a good model even with a small amount of training data (*e.g.*, 500 frames, 490×490 pixels each frame), relieving the pressure of capturing large-scale datasets. (Supplementary Fig. 2).

To demonstrate the predominance of our transformer network over CNNs, we generated simulated calcium imaging data (Method and Supplementary Fig. 3) and used them to train a 3D-UNet[16], as well as the transformer of SRDTrans, using the same spatial redundancy sampling strategy. We term the former as spatial redundancy denoising CNN (SRDCNN). The visualized feature maps of a typical deep layer intuitively show the superiority of SRDTrans in revealing fine-grained patterns (Fig. 1d). As features flow through the network, the limited receptive field of convolutional kernels makes CNN-based methods only focus on rough features while our transformer architecture still has a strong perception of sophisticated

structures. We also compare the denoised images of the two architectures (Fig. 1e). The result of SRDCNN is obviously over-smoothed, especially in regions with sharp edges, which is a manifestation of spectral bias that the low-frequency information is overfitted while the high-frequency information is hardly preserved (Supplementary Fig. 4). This deficiency is largely alleviated by SRDTrans, and more subcellular structures such as dendritic fibers can be restored accurately. The intensity profile deconstructed from the SRDTrans denoised image is more consistent with GT (Fig. 1f). Moreover, lacking the ability to capture long-range temporal correlations also drags down the denoising performance of SRDCNN on time-lapse imaging data. However, for SRDTrans, the output SNR continuously grows as the input temporal scale (T) increases (Fig. 1g). A more comprehensive evaluation of the influence of input temporal scale suggests that SRDTrans can make full use of the information offered by temporally distant pixels (Supplementary Fig. 5). To show the practicality of SRDTrans at extremely low imaging speed, we compared the performance of methods combining different networks and sampling schemes on simulated calcium imaging data sampled at 0.3 Hz (Supplementary Video 1). When the similarity between adjacent frames is low, using spatial redundancy sampling is more reasonable. Succinctly, the synergy between spatial redundancy sampling and dedicated transformer architecture endows SRDTrans with the ability to restore high-resolution structures and fast dynamics.

SRDTrans improves the performance of SMLM. Given N detected fluorescence photons, the lower bound of the precision of SMLM scales to $1/\sqrt{N}$[34], which is the mathematical formula of the shot-noise limit or the well-known standard quantum limit[40,41]. This indicates

that the fundamental physical limit of localization precision is shot noise. To investigate the benefits that our denoising method can bring to SMLM, we first applied SRDTrans to simulated stochastic optical reconstruction microscopy (STORM) data with GT for quantitative evaluation[42,43]. The noise-free single-molecule emitting images were synthesized by TestSTORM[44] and corresponding noisy images with different SNRs were then generated by applying different levels of Mixed Poisson-Gaussian noise (Method and Supplementary Fig. 6). For the image stack of each SNR, we trained a specific model for it and then processed it using the model to obtain the denoised image stack. Quantitative comparisons of the visualized images (Fig. 2a and Supplementary Video 2) and the extracted intensity profiles (Fig. 2b) show that the results of SRDTrans are highly consistent with GT. Over a wide range of imaging SNR, including some extremely low-SNR conditions, SRDTrans can substantially improve the image quality evaluated by SNR at a pixel-intensity level and structural similarity (SSIM) at a perception level (Fig. 2c). Compared with other self-supervised methods[23-28], SRDTrans can better preserve the distribution and intensity of emitters owing to its strong ability in exploiting high-resolution features and long-range dependencies (Supplementary Fig. 7).

Next, we evaluate the improvement of single-molecule localization performance with the enhancement of image SNR. To reconstruct super-resolved images, we applied ThunderSTORM[45], one of the most frequently used localization software with an excellent balance between accuracy and execution runtime[46,47]. Contaminated by noise, the image reconstructed from the original noisy data contains many misidentified molecules (Fig. 2d). By contrast, the reconstructed image from SRDTrans denoised images can reveal clear and

continuous cytoskeletal filaments that are not previously recognizable because of suppressed localization error and improved resolution (Fig. 2e and Supplementary Fig. 8). For better quantitative analysis, we matched detected fluorescent molecules with GT using the Hungarian algorithm[46]. From raw images, few fluorescent molecules can be detected and most of them are wrongly localized. After SRDTrans denoising, almost all molecules can be detected in high agreement with GT (Fig. 2f and Supplementary Fig. 9). Using Jaccard index and root mean squared error (RMSE) as the metrics to respectively quantify the proportion of correctly detected molecules and the localization accuracy of those detected molecules, we found that Jaccard index was improved by ~6-fold (85.7 % versus 14.1 %) and RMSE was reduced by ~3.4-fold (24.86 nm versus 85.94 nm) after denoising (Fig. 2g). From a more comprehensive perspective, we further adopted the metric termed efficiency that combines Jaccard index and RMSE[47]. The results show that SRDTrans improved the efficiency of single-molecule localization from -21.54 to 71.33 (Fig. 2h), fully demonstrating the benefits of SRDTrans on SMLM.

We further applied SRDTrans to experimental SMLM data of microtubules to validate its ability in revealing subcellular structures. Raw frames were captured with low excitation power and short exposure time to reduce phototoxicity and emitter density. The experimentally obtained single-molecule emission images and SRDTrans denoised images are shown in Fig. 3a. Disturbed by the noise, the reconstruction algorithm can hardly localize the fluorescent molecules in the raw frames. The reconstructed super-resolution image contains many erroneous spots and cannot reveal any meaningful structures (Fig. 3b). Whereas

SRDTrans can effectively suppress the noise and remove localization artifacts in the reconstructed image, making the distribution and extension directions of microtubules visible. We computed the Fourier ring correlation (FRC)[48,49] curve to quantify the resolution from the SMLM images (Fig. 3c). The image resolution is defined as the inverse of the spatial frequency at the intersection of the FRC curve and the threshold line (~0.143). Benefitting from the removal of noise, the resolution of SRDTrans denoised data was improved from 52.4 nm to 36.0 nm and the localization uncertainty was reduced from 8.0 nm to 5.0 nm (Fig. 3d). In addition to the data acquired by our instrument, we also used SRDTrans to denoise publicly available SMLM data contributed by other laboratories[50] (Fig. 3e). The reconstructed super-resolution images indicate that SRDTrans can eliminate the artifacts and bring more complete organelle structures (Fig. 3f,g). We applied Gaussian fitting to the intensity profile perpendicular to the microtube filaments and measured the full width at half maximum (FWHM) to quantify the image resolution (Fig. 3h). The SRDTrans denoised data shows improved resolution as the averaged FWHM dropped from 187.89 nm to 60.96 nm (Fig. 3i). As SMLM is heading towards live-cell imaging and long-term observation[51,52] our denoising method promises to be a beneficial tool to reduce the laser power by resolving fluorescent molecules from very low-SNR frames. For 3D SMLM, since the axial positions of molecules are estimated through PSF engineering[34], SRDTrans is expected to offer great help by resolving single-molecule emission patterns from low-SNR images.

SRDTrans enhances two-photon volumetric calcium imaging in the mouse cortex. In multiphoton imaging, the volumetric imaging speed decreases linearly as the number of

206 scanning planes increases. Thus, the achievable sampling rate for observing neuronal
207 populations with large axial ranges is often quite low, making those denoising methods that rely
208 on the similarity between temporally adjacent frames infeasible[9,23-25]. However, SRDTrans
209 provides an opportunity to restore those highly degraded fluorescence signals in large-scale
210 volumetric calcium imaging by utilizing the similarity between spatially adjacent pixels. To
211 evaluate the denoising performance of SRDTrans on calcium imaging with different sampling
212 rates, we generated realistic calcium imaging data with synchronized GT using NAOMi[53]. We
213 started from denoising high-sampling-rate (30 Hz) data and found that SRDTrans can
214 effectively remove noise and recover previously indiscernible structures such as soma, neurites,
215 and vascular shadows (Fig. 4a, Supplementary Fig. 10, and Supplementary Video 3). The
216 enhancement is not only manifested in the visual effect, but more importantly, in the accurate
217 restoration of pixel intensities (Fig. 4b). Visualization in the frequency domain (calculated by
218 discrete Fourier transform) shows that SRDTrans can restore most of the frequency
219 components (Fig. 4c), especially those high-frequency components lost by DeepCAD[9,25] and
220 DeepInterpolation[24], thereby leading to the best performance that has ever been achieved in
221 denoising calcium imaging data (Supplementary Fig. 11). Such a remarkable denoising
222 capability can be maintained over a wide range of input SNRs (from -2.08 dB to 17.68 dB),
223 and the average SNR improvement is about 22 dB (Fig. 4d).

224 Then, we investigate the denoising performance of SRDTrans on calcium imaging data
225 sampled at 0.3 Hz, which is 100 times lower than the imaging speed demonstrated above.
226 Bilateral assessments both in the space domain and the frequency domain reveal that SRDTrans

227 can accurately retrieve the fluorescence signals from the original highly degraded images
228 without structural blurring and frequency deficiency (Fig. 4e and Supplementary Fig. 11).
229 When the sampling rate is much lower than the fluorescence dynamics, the large discrepancy
230 between the signals in two adjacent frames cannot provide the temporal correlation required by
231 DeepCAD and DeepInterpolation, so they are not accurate enough to be used in conditions of
232 low imaging speed or fast activity (Supplementary Fig. 12). To figure out how SRDTrans works
233 at different imaging speeds, we performed an ablation study on different sampling strategies
234 and network architectures (Fig. 4f and Supplementary Tab. 3). The results indicate that spatial
235 redundancy sampling performs better at low imaging speeds, whereas temporal redundancy
236 sampling performs better at high imaging speeds. Almost equally for all imaging speeds, our
237 transformer architecture offers an additional improvement (~2.05 dB on average) over
238 conventional CNNs. In general, the synergistic combination of the spatial redundancy sampling
239 and the transformer architecture in SRDTrans provides better performance than DeepCAD at
240 all imaging speeds (Supplementary Fig. 13). In the time domain, the superior ability of
241 SRDTrans can reveal high-fidelity calcium transitions without distorting fluorescence kinetics
242 (Fig. 4g and Supplementary Fig. 14). Moreover, we also simulated fast-moving objects to
243 imitate migrating cells that are widely existed in living organisms. The quantitative evaluation
244 shows that SRDTrans can preserve the structure of densely distributed objects even if they are
245 moving at a speed of up to 9 pixel/frame (Supplementary Fig. 15 and Supplementary Video 4),
246 alleviating the shortage of denoising methods for fast-moving cells and organelles.

Finally, we went a step further in denoising calcium imaging data by applying SRDTrans to volumetric recordings, which is not achievable for other self-supervised denoising methods[9,24-25] because of their heavy reliance on high sampling rates. We used transgenic mice expressing the GCaMP6f genetically encoded calcium indicator[54] and imaged a large brain volume of 500×500×200 μm in the mouse cortex using a two-photon microscope. We scanned 100 planes with a frame rate of 30 Hz, and thus the overall volume rate was 0.3 Hz. 3D visualization of the neural volume shows that the spatial profiles and firing states of the neurons can be revealed after denoising, which otherwise would be swamped by severe shot noise (Fig. 5a and Supplementary Video 5). For better comparison, we present the snapshots of a certain imaging plane at two different moments. With the enhancement of SRDTrans, the structure and distribution of the neurons become clearly observable (Fig. 5b). We also extracted the fluorescence traces along the time axis and found that a large number of calcium transients can be restored after denoising (Fig. 5c). The dramatically improved SNR would propel the decoding of underlying neural activity from fluorescence signals. Since neural circuits in the mammalian brain are spatially coordinated and temporally orchestrated, deciphering the function of large neuronal ensembles requires large-scale volumetric imaging with high SNR. The superior denoising performance of SRDTrans provides an opportunity to implement high-sensitivity volumetric calcium imaging for investigating functionally concerted neurons and recognizing circuit motifs, especially those distributed across multiple cortical layers.

Discussion

In this work, we proposed SRDTrans, a self-supervised denoising method for fluorescence microscopy that achieves excellent performance on various time-lapse imaging data, even for samples with very fast activities or microscopes with extremely low imaging speed. The spatial redundancy sampling strategy eliminates the reliance on temporal redundancy, making SRDTrans robust to image sequences with very low correlations between adjacent frames. Benefiting from the spatiotemporal transformer architecture with global receptive fields, SRDTrans can alleviate the inherent spectral bias in CNNs and restore more high-frequency components. The strong capability of long-range perception also endows SRDTrans with better performance than all previous methods. For demonstration, we focused on applications with low temporal redundancy, such as SMLM of cytoskeletal filaments and two-photon calcium imaging of large neural volumes, which verifies the effectiveness and superiority of SRDTrans in visualizing and analyzing complex biological structures and functions. Additionally, our method does not rely on any assumptions about the contrast mechanism, noise model, sample dynamics, and imaging speed. Thus, it can be readily extended to other biological samples and imaging modalities, such as membrane voltage imaging, single-protein detection, light sheet microscopy, light field microscopy, and super-resolution microscopy[55-59].

As the development of fluorescent indicators heading towards faster kinetics to monitor biological dynamics at the millisecond scale[60-62], the imaging speed required to record these fast activities is continuously growing. Obtaining adequate sampling rates is becoming more and more challenging for denoising methods relying on temporal redundancy. Our rationale is to fill the niche by seeking to utilize spatial redundancy as an alternative to enable self-

supervised denoising in more imaging applications. Although the perfect case for spatial redundancy sampling is that the spatial sampling rate is two times higher than the Nyquist sampling of the diffraction limit to ensure that two adjacent pixels have nearly identical optical signals, the endogenous similarity between two spatially downsampled sub-stacks is sufficient to guide the training of the network. However, this does not mean that the proposed spatial redundancy sampling strategy can fully replace temporal redundancy sampling, since the ablation study (Fig. 3f) shows that, if equipped with the same network architecture, the temporal redundancy sampling can achieve better performance in high-speed imaging. The superiority of SRDTrans over DeepCAD at high imaging speeds actually derives from the transformer architecture. In general, spatial redundancy and temporal redundancy are two complementary sampling strategies to enable self-supervised training of denoising networks for fluorescence time-lapse imaging. Using which sampling strategy depends on which kind of redundancy is more sufficient in the data. It is noteworthy that there are still many cases where neither redundancy is sufficient to support current sampling strategies. Developing specific or more general self-supervised denoising methods is of persistent value for fluorescence imaging.

Methods

The spatial redundancy sampling scheme. In SRDTrans, we employed a spatial redundancy sampling strategy to produce training pairs. The detailed implementation for generating training pairs in SRDTrans is shown in Fig. 1b and Supplementary Fig. 1d. For each image of an input training stack with H×W×T pixels, we spatially divided it into many adjacent small patches with 2×2 pixels. Next, we randomly selected three adjacent pixels from each patch. The central pixel was used to compose the input sub-stack and the two spatially adjacent pixels were used to compose the two target sub-stacks. After traversing all patches, we can finally obtain three downsampled sub-stacks with the size of H/2×W/2×T pixels. Since the fluorescence signals of spatially adjacent pixels are closely correlated, the input sub-stack and each target sub-stack can be considered as two independent samples of the same underlying pattern. Thus, the input sub-stack and the two target sub-stack can form two training pairs, which can be used for the self-supervised training of denoising networks (Supplementary Note 1).

Network architecture and loss function. The transformer architecture of SRDTrans is composed of three parts: a temporal encoder module, a spatiotemporal transformer block (STB), and a temporal decoder module (Fig. 1c). The temporal encoder module is equipped with two temporal encoders. Each temporal encoder can compress the temporal scale of the input sub-stack by a factor of r (r is 4 in this work) using a convolutional layer with 3×3 kernels. Different from those UNet-type architectures with many up-sampling and down-sampling operations, SRDTrans do not reduce the size of feature maps in these encoders (Supplementary

Fig. 16). Thus, for an input sub-stack with a size of H/2×W/2×T pixels, the output size of the temporal encoder module is H/2×W/2×T/r^2. The output from the temporal encoder module will be fed into the STB to extract global information both in space and time. In the STB, the input is divided into patches with the size of p×p×T/r^2 (p is 7 in this work). These patches are first flattened into one-dimensional vectors and input into the position embedding layer, where spatial concatenation and linear transformation are performed. Two multi-head self-attention (MSA) layers are then stacked to calculate the global attention between these vectors. Local features flow fully in MSA layers and densely interact with long-range global features. Finally, the output of the STB is remapped by the temporal decoder module, and its temporal scale can be rescaled to T. This decompression operation is implemented by two cascaded temporal decoders using convolutional layers with 3×3 kernels.

We used a linear combination of L1-norm loss and L2-norm loss as the loss function to optimize the parameters of SRDTrans. We defined the input sub-stack filled with central pixels as S_c, the target sub-stack filled with its vertically adjacent pixels as S_v, and the target sub-stack filled with its horizontally adjacent pixels as S_h. The total loss consists of two pairs of training losses, which is defined as:

$$
\begin{aligned}
L_{ver} &= \left\| \mathrm{F}_{SRDTrans}(S_i) - S_v \right\|_2^2 + \left| \mathrm{F}_{SRDTrans}(S_i) - S_v \right|_1, \\
L_{hor} &= \left\| \mathrm{F}_{SRDTrans}(S_i) - S_h \right\|_2^2 + \left| \mathrm{F}_{SRDTrans}(S_i) - S_h \right|_1, \\
L_{total} &= L_{ver} + L_{hor}.
\end{aligned}
$$

where L_{ver} and L_{hor} denote the loss of the vertically and horizontally adjacent sub-stacks, respectively.

Training and inference. To achieve optimal performance, specific models were trained for stacks with different SNRs. One or more training stacks were divided into a specified number of 3D (xy-t) training pairs (6000 by default). The batch size for all experiments was set to the number of GPUs (NVIDIA GeForce RTX 3090) being used. The patch size was set to 128×128×128 pixels. All extracted training pairs were geometrically transformed by random flipping or rotation for 8-fold data augmentation. The compression factor of each temporal encoder was set to 4. In the STB, we set the internal patch size to 7, the number of heads in the multi-headed self-attention block to 8, and the embedded feature channels to 128. For model optimization, we used the Adam optimizer and the exponential decay rate for the first moment was 0.9, the exponential decay rate for the second moment was 0.999, and the learning rate was 0.00001. PyTorch was used to construct the network and implement all operations. In the interference stage, the model of the last training epoch was selected for final processing. The denoised result of each image stack was saved as a separate TIF file.

Data simulation. Quantitative evaluations were performed on simulated data because noise-free images (ground truth) are available. To synthesize noise-free two-photon calcium imaging data, we used the Neural Anatomy and Optical Microscopy (NAOMi)[53], which can generate realistic calcium imaging data with high-fidelity tissue characteristics and fluorescence kinetics. Then we applied different levels of Mixed Poisson Gaussian (MPG) noise to generate calcium imaging data of different imaging SNRs[9,25]. We first synthesized calcium imaging data sampled at 30 Hz and 1 Hz. The data of other sampling rates were obtained by extracting frames at

365 different intervals. The image size for all simulated calcium imaging data was 490×490 pixels
366 and the pixel size was 1.02 μm.

367 To generate realistic SMLM data, we first acquired reconstructed super-resolution images
368 from the ShareLoc.XYZ platform (https://shareloc.xyz/)[50]. These images were experimentally
369 obtained on a Nikon N-STORM microscope and contained densely distributed microtubules
370 immuno-labeled with Alexa-647[63]. The tracks of all microtubules in a selected ROI (region of
371 interest) were extracted semi-automatically using the JFilament plugin of ImageJ[64]. We then
372 generated synthetic single-molecule emitting image stacks (ground-truth images without noise)
373 using the TestSTORM[44] simulator by loading the microtubule patterns from JFilament. All
374 fluorescent molecules were linked on the microtubule pattern with a radius of 12.5 nm. For
375 imaging parameters, the numerical aperture was 1.4 and the frame rate was 200 Hz (5-ms
376 exposure time). The image size was 328×328 pixels and the pixel size was 30 nm. Noisy stacks
377 were generated by applying MPG noise post hoc with a customized Matlab script[9,25].

378 We synthesized moving objects with different moving speeds using the MNIST
379 dataset[65,66]. Each frame was defined as an image of 512×512 pixels with a black
380 background and many bright moving digits. Each digit was an image patch (28×28 pixels)
381 randomly extracted from the MNIST dataset, moved in a random direction, and
382 appeared or disappeared only once. The total number of digits in the field of view was
383 500. We first generated the data with a moving speed of 0.5 pixel/frame. Those data
384 with higher moving speeds were then generated by extracting frames at different
385 intervals. The total frame number for all moving speeds was 5000. The final experiment

was implemented on twenty datasets with moving speeds from 0.5 to 10 pixel/frame. The sampling interval of the moving speed was 0.5 pixel/frame.

SMLM imaging. The imaging samples (including the buffer solution and the sample holder) for the SMLM experiments were purchased from Standard Imaging (Beijing) Biotechnology Co., Ltd (https://www.standardimaging.cn/standardsample?lang=en). The SMLM experiments were performed on a commercial microscope (Nikon N-STORM) equipped with laser sources of 405 nm and 640 nm, which were used for activation and excitation, respectively. An sCMOS camera (Hamamatsu Flash 4.0) was placed on the image plane to capture the emission signals. To mimic living-cell imaging, we used low excitation power to reduce phototoxicity and short exposure time to obtain images with low emitter density. The detailed settings are summarized in Supplementary Tab. 5.

SMLM sample preparation. The Biologics Standards-Cercopithecus-1 (BSC-1) cell line purchased from Pricella Life Technology Co. LTD was used for SMLM imaging. BSC-1 cells were cultured in DMEM (Invitrogen, #11965-118) supplemented with 10% fetal bovine serum (FBS) (Gibco, #16010-159). To prevent bacterial contamination, 100 μg/ml penicillin and streptomycin (Invitrogen, #15140122) were added into the DMEM medium. Cells were grown under standard cell culture conditions (5% CO2, humidified atmosphere at 37°C). BSC-1 cells were plated on #1.5 glass-bottom dishes over 48 h before sample preparation. For cell passage, cells were washed with pre-warmed PBS (Life Technologies, #14190500BT) 3 times and digested with 25% trypsin (Gibco, #25200-056) for 30 s. BSC-1 1 cell lines were tested for potential mycoplasma contamination (MycoAlert, Lonza), and all tests showed negative results.

For immunofluorescence staining, cells were grown on 35 mm, #1.5 glass coverslips. We pre-treated glass-bottom dishes with fibronectin (Invitrogen, #33016015) for 1 h at 37°C to increase cell adhesion. On the day of sample preparation, the cell density should be about 50%-70%. Cells were fixed with 37°C pre-warmed fixation buffer for 10 min, containing 4% paraformaldehyde (EMS) and 0.1% glutaraldehyde in PBS. Then the sample was washed three times with PBS. For quenching the background fluorescence, we incubated the cells with 2 ml 0.1% NaBH4 solution in PBS for 7 min, optionally shaking on the shaker (<1 Hz). The sample was washed three times with 2 ml PBS and then incubated for 30 min in PBS containing 5% BSA (Jackson, #001-000-162) and 0.5% Triton X-100 (Fisher Scientific) at 37°C. All antibodies were diluted in the 5% BSA+ 0.5% triton solution described above. Next, we incubated the sample for 40 min with the appropriate dilution of primary antibodies: beta-tubulin (E7, DSHB) at 25°C. After primary antibodies incubation, the cells were washed three times with 2 ml PBS for 5 min. Secondary antibodies were incubated for 60 min with the appropriate dilutions of secondary antibodies (Jackson Immuno Research, conjugated with Cy5 in our laboratory) at 25°C. After being washed 3 times with PBS, cells were fixed with post-fixation buffer for 10 min. The sample was stored at 4°C in PBS and protected from light. Before imaging, we used an imaging buffer (Standard Imaging (Beijing) Biotechnology Co., Ltd, #ST031) to replace PBS.

SMLM reconstruction. The super-resolution SMLM images were reconstructed by the Fiji plugin ThunderSTORM[45]. The detailed configuration is set as: the image filter was the wavelet filter (B-spline) with the order of 3 and the scale of 2; the algorithm for determining

the approximate position of molecules was the local maximum algorithm; the sub-pixel localization is performed by the integrated Gaussian PSF model with a fitting radius of 3 pixels; a fitting method of "weighted least squares", and an initial sigma of 1.6 pixels. Both visualization images are generated by averaged shifted histogram with a magnification of 5.

Mouse preparation and calcium imaging. All experiments involving mice were performed in accordance with the institutional guidelines for animal welfare and have been approved by the Animal Care and Use Committee of Tsinghua University. Transgenic mice hybridized between Rasgrf2-2A-dCre mice and Ai148 (TIT2L-GC6f-ICL-tTA2)-D mice expressing Cre-dependent GCaMP6f genetically encoded calcium indicator (GECI) were used for calcium imaging of neural circuits. Craniotomy surgeries were conducted to remove the skull and a coverslip was implanted on the craniotomy region for chronic imaging. Two-photon volumetric imaging of the mouse cortex was performed on head-fixed mice without anesthesia using a standard two-photon microscope. The neural volume being recorded was located at the primary visual cortex with a depth of about 150~350 μm below the dura, and was scanned for 100 planes with an axial step of 2 μm. The whole imaging session lasted 30 minutes with a volume rate of 0.3 Hz.

Three-dimensional visualization of neural activity. For volumetric calcium imaging, we used Imaris 9.0 (Oxford Instruments) to visualize the calcium activity of the neuronal population before and after denoising. Specifically, we imported the 4D (xyz-t) data into Imaris, applied pseudo-color to the images, and then performed 3D rendering using the maximum

intensity projection mode. The contrast and brightness were adjusted to make structures in the volume as clear as possible. All values for gamma correction were set to one.

Method comparison. We compared the performance of SRDTrans with six baseline self-supervised methods, including Noise2Noise[23], Noise2Void[26], Noise2Self[27], Neighbor2Neighbor[28], DeepInterpolation[24], and DeepCAD[9,25]. These methods were all implemented by open-source codes released by relevant papers. The denoising model of each method was trained and tested on the same datasets. For the methods designed for 2D images, we split the time-lapse (xy-t) image stack into a series of 2D frames to match the input dimension. Training and inference were performed frame-by-frame. We followed the default training settings about network architectures and hyperparameters for all methods. Specifically, the model of DeepInterpolation was fine-tuned based on a public pre-trained model (pre-trained with 225,000 two-photon images of the Ai93 reporter line). Other methods were trained from scratch. The detailed settings of each method are listed in Supplementary Tab. 4.

Evaluation metrics. We used several metrics to evaluate the performance of different denoising methods. For an image (or an image stack) S_x and its ground truth S_y, the metrics are defined as follows:

Signal-to-noise ratio (SNR) measures the pixel-level deviation between two images using the logarithmic decibel scale, which is formulated as

$$SNR = 10 \log_{10} \frac{\left\| S_y \right\|_2^2}{\left\| S_x - S_y \right\|_2^2} .$$

Structural similarity (SSIM) measures the similarity between two images on a perceptual level, including luminance, contrast, and structure. The definition is

$$SSIM = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)},$$

where $\{\mu_x, \mu_y\}$ and $\{\sigma_x, \sigma_y\}$ are the means and variances of S_x and S_y, respectively. σ_{xy} is the covariance of S_x and S_y. The two constants c_1 and c_2 are defined as $c_1=(k_1L)^2$ and $c_2=(k_2L)^2$ with k_1=0.01, k_2=0.03, and L=65535.

Jaccard index measures the proportion of correctly detected molecules in SMLM. Those correctly localized fluorescent molecules are true positives (TP). Those incorrectly localized molecules are false positives (FP) and those undetected molecules are false negatives (FN). The Jaccard index is formulated as:

$$Jaccard = 100\frac{TP}{TP+FP+FN}\% .$$

Root mean squared error (RMSE) quantifies the mean difference between the localized positions (P_x) and ground-truth positions (P_y) of all detected fluorescent molecules:

$$RMSE = \sqrt{\frac{1}{TP}\sum_{TP}\left\|P_y - P_x\right\|_2^2} ,$$

Efficiency (E) is a comprehensive metric combining the Jaccard index and RMSE to measure the performance of single-molecule localization[43]. It can simultaneously reflect the ability to detect molecules from images (measured by Jaccard) and the ability to precisely locate molecules (measured by RMSE), which is defined as:

$$Efficiency = 100 - \sqrt{(100 - JAC)^2 + \alpha^2 RMSE^2} ,$$

where α=1 nm^{-1} controls the trade-off between Jaccard and RMSE.

Pearson correlation coefficient measures the similarity between a variable (images and fluorescence traces) and its ground truth, which is formulated as

$$R = \frac{E\left[(S_x - \mu_x)(S_y - \mu_y)\right]}{\sigma_x \sigma_y},$$

where E represents the arithmetic mean. $\{\mu_x, \mu_y\}$ and $\{\sigma_x, \sigma_y\}$ are the means and variances of S_x and S_y, respectively.

Logarithmic frequency distance (LFD) quantifies the spectral difference between two images in the frequency domain. For images with a size of H×W pixels, LFD is formulated as:

$$d(u,v) = \left\| F_{S_x}(u,v) - F_{S_y}(u,v) \right\|^2,$$
$$LFD = \log_{10}\left[\frac{1}{HW}\left(\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\left|d(u,v)\right|\right) + 1\right].$$

F_{Sx} and F_{Sx} are the discrete Fourier transform of S_x and S_y, respectively.

Statistics and reproducibility. To ensure the reproducibility of the findings, we report the sample size and statistics in the legend and text of each experiment. All boxplots are drawn in the standard Tukey boxplot format. The upper and lower quartiles are represented by boxes, and the lines in the boxes indicate the median. The lower whisker represents the minimum observed value, equal to the lower quartile minus 1.5× the interquartile range. The upper whisker the maximum observed value, equal to the upper quartile plus 1.5× the interquartile range. Results obtained through experimental or observational studies or statistical analysis of data sets can be reproduced with high

reliability when the study is repeated. In the figure shown are representative frames, and all experiments obtained similar results on more than 1,000 frames.

Data availability

We have no restrictions on data availability. Both the simulated and experimental data of single-molecule localization microscopy and two-photon calcium imaging in this work are available at https://github.com/cabooster/SRDTrans/tree/main/datasets.

Code availability

The source code and pre-trained models are publicly available at https://github.com/cabooster/SRDTrans.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (62088102, 62222508, 62071272) and National Key Research and Development Program of China (Project No. 2022YFB36066), in part by the Shenzhen Science and Technology Project under Grant (CJGJZD20200617102601004, JCYJ20220818101001004). This work was also supported by the China Nation Postdoctoral Program for Innovative Talents (BX2021159) and Shuimu Tsinghua Scholar Program. We thanked Dr. Huiwen Hao from Standard Imaging (Beijing) Biotechnology Co., Ltd for providing information about SMLM sample preparation.

Author Contributions

Q.D., H.W. and J.W. supervised this research. Q.D., H.W., J.W. and X.L. conceived and initiated this project. X.L., X.H. and X.C. designed detailed implementations and performed imaging experiments. X.H. and X.L. developed the Python code, performed simulations, and processed relevant imaging data. J.F. prepared samples and provided models animals. Z.Z. gave critical support on the two-photon imaging system and imaging procedures. X.H., X.L., and X.C. analyzed the data, prepared figures and videos. X.L., X.H., X.C., J.F., Z.Z. and J.W. participated in discussions about the results and gave valuable advice. All authors participated in the drafting of the manuscript.

Competing interests

The authors declare no competing interests.

Materials & Correspondence

Correspondence and requests for materials should be addressed to Jiamin Wu, Haoqian Wang, or Qionghai Dai.

References

1. Royer, L. A. et al. Adaptive light-sheet microscopy for long-term, high-resolution imaging in living organisms. *Nat. Biotechnol.* **34**, 1267-1278 (2016).
2. Fan, J. et al. Video-rate imaging of biological dynamics at centimetre scale and micrometre resolution. *Nat. Photon.* **13**, 809-816 (2019).
3. Keller, P.J. & Ahrens, M.B. Visualizing whole-brain activity and development at the single-cell level using light-sheet microscopy. *Neuron* **85**, 462-483 (2015).
4. Balzarotti, F. et al. Nanometer resolution imaging and tracking of fluorescent molecules with minimal photon fluxes. *Science* **355**, 606-612 (2017).
5. Baruch, K. et al. PD-1 immune checkpoint blockade reduces pathology and improves memory in mouse models of Alzheimer's disease. *Nature medicine* **22**, 135-137 (2016).
6. Wu, J. et al. Iterative tomography with digital adaptive optics permits hour-long intravital observation of 3D subcellular dynamics at millisecond scale. *Cell* **184**, 3318–3332 (2021).
7. Verweij, F. J. et al. The power of imaging to understand extracellular vesicle biology in vivo. *Nat. Methods* **18**, 1013–1026 (2021).
8. Mulay, S.V. et al. Enhanced fluorescence turn‐on imaging of hypochlorous acid in living immune and cancer cells. *Chemistry‐A European Journal* **22**, 9642-9648 (2016).
9. Li, X. et al. Real-time denoising enables high-sensitivity fluorescence time-lapse imaging beyond the shot-noise limit. *Nat. Biotechnol.* **1-11** (2022).
10. Casacio, C.A. et al. Quantum-enhanced nonlinear microscopy. *Nature* **594**, 201-206 (2021).
11. Brida, G., Genovese, M. & Ruo Berchera, I. Experimental realization of sub-shot-noise quantum imaging. *Nat.Photon* **4**, 227–230 (2010).
12. Meiniel, W., Olivo-Marin, J. C. & Angelini, E. D. Denoising of microscopy images: a review of the state-of-the-art, and a new sparsity-based method. *IEEE Trans. Image Process*. **27**, 3842–3856 (2018).
13. Dabov, K., Foi, A., Katkovnik, V. & Egiazarian, K. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Trans. Image Process*. **16**, 2080-2095 (2007).
14. Gu, S., Zhang, L., Zuo, W. & Feng, X. Weighted nuclear norm minimization with application to image denoising. In *Proc. of the IEEE/CVF conference on computer vision and pattern recognition* 2862-2869 (2014).
15. Zhang K. et al.. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Trans. Image Process*. **26**, 3142-3155 (2017).
16. Mao, X., Shen, C. & Yang, Y.-B. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in Neural Information Processing Systems* **29** (2016).
17. Tai, Y., Yang, J., Liu, X. & Xu, C. Memnet: A persistent memory network for

image restoration. in *Proc.* In *Proc. of the IEEE/CVF conference on computer vision and pattern recognition* 4539-4547 (2017).

18. Weigert, M. et al. Content-aware image restoration: pushing the limits of fluorescence microscopy. *Nat. Methods* **15**, 1090–1097 (2018).
19. Belthangady, C. & Royer, L. A. Applications, promises, and pitfalls of deep learning for fluorescence image reconstruction. *Nat. Methods* **16**, 1215–1225 (2019).
20. Chen, J. et al. Three-dimensional residual channel attention networks denoise and sharpen fluorescence microscopy image volumes. *Nat. Methods* **18**, 678–687 (2021)
21. Chaudhary, S., Moon, S. & Lu, H. Fast, efficient, and accurate neuro-imaging denoising via supervised deep learning. *Nat Commun.* **13**, 5165 (2022)
22. Wang, Z., Xie, Y. & Ji, S. Global voxel transformer networks for augmented microscopy. *Nat. Mach. Intell.* **3**, 161–171 (2021).
23. Lehtinen, J. et al. Noise2Noise: learning image restoration without clean data. In *Proc. of the 35th International Conference on Machine Learning* (eds Dy, J. & Krause, A.) 2965–2974 (PMLR, 2018).
24. Lecoq, Jérôme, et al. Removing independent noise in systems neuroscience data using DeepInterpolation. *Nat. Methods* **18**, 1401-1408 (2021).
25. Li, X. et al. Reinforcing neuron extraction and spike inference in calcium imaging using deep self-supervised denoising. *Nat. Methods* **18**, 1395-1400 (2021).
26. Krull, A., Buchholz, T.-O. & Jug, F. Noise2Void—learning denoising from single noisy images. In *Proc. of the IEEE/CVF conference on computer vision and pattern recognition* 2129-2137 (2019).
27. Batson, J. & Royer, L. Noise2Self: blind denoising by self-supervision. In *Proc. of the 36th International Conference on Machine Learning* 524–533 (PMLR, 2019).
28. Huang, T. et al. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *Proc. of the IEEE/CVF conference on computer vision and pattern recognition* 14781-14790 (2021).
29. Lequyer, J. et al. A fast blind zero-shot denoiser. *Nat. Mach. Intell.* **4**, 953–963 (2022).
30. Luo, W. et al. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems* **29** (2016).
31. Araujo, André, Wade Norris, and Jack Sim. Computing receptive fields of convolutional neural networks." *Distill* **4**.11 (2019).
32. Rahaman N. et al. On the spectral bias of neural networks. In *International Conference on Machine Learning* 5301-5310 (PMLR, 2019).
33. Khayatkhoei M, Elgammal A. Spatial frequency bias in convolutional generative adversarial networks. In *Proc. of the AAAI Conference on Artificial*

Intelligence, Vol. **36** 7152-7159 (2022).

34. Lelek, M. et al. Single-molecule localization microscopy. *Nat. Rev. Methods Primers* **1**, 39 (2021).
35. Speiser, A. et al. Deep learning enables fast and dense single-molecule localization with high accuracy. *Nat. Methods* **18**, 1082–1090 (2021).
36. Liu Z. et al. Video swin transformer. In *Proc. of the IEEE/CVF conference on computer vision and pattern recognition* 3202-3211 (2022).
37. Zhou H. et al. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201* (2021).
38. Hatamizadeh A. et al. Unetr: Transformers for 3d medical image segmentation. In *Proc. of the IEEE/CVF winter conference on applications of computer vision* (2022).
39. Hatamizadeh A. et al. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *7th International Workshop of Medical Image Computing and Computer-Assisted Intervention* (2021).
40. Taylor, M. A. & Bowen, W. P. Quantum metrology and its application in biology. *Phys. Rep.* **615**, 1–59 (2016).
41. T. Nagata. et al. Beating the standard quantum limit with four-entangled photons. *Science* **316**, 726–729 (2007).
42. Rust, M., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3**, 793–796 (2006).
43. Elias Nehme, Lucien E. Weiss, Tomer Michaeli, and Yoav Shechtman. Deep-STORM: super-resolution single-molecule microscopy by deep learning. *Optica* **5**, 458-464 (2018).
44. Sinkó J. et al. TestSTORM: Simulator for optimizing sample labeling and image acquisition in localization based super-resolution microscopy. *Biomedical optics express* **5**, 778-787 (2014).
45. Ovesný, M. et al. ThunderSTORM: a comprehensive ImageJ plug-in for PALM and STORM data analysis and super-resolution imaging. *Bioinformatics* **30**, 2389-2390 (2014).
46. Sage, D. et al. Quantitative evaluation of software packages for singlemolecule localization microscopy. *Nat. Methods* **12**, 717–724 (2015).
47. Sage, D. et al. Super-resolution fight club: assessment of 2D and 3D single-molecule localization microscopy software. *Nat. Methods* **16**, 387–395 (2019).
48. Nieuwenhuizen, R., Lidke, K., Bates, M. *et al.* Measuring image resolution in optical nanoscopy. *Nat. Methods* **10**, 557–562 (2013).
49. Descloux, A., Grußmayer, K.S. & Radenovic, A. Parameter-free image resolution estimation based on decorrelation analysis. *Nat. Methods* **16**, 918–924 (2019).
50. Ouyang, W. et al. ShareLoc — an open platform for sharing localization microscopy data. *Nat. Methods* **19**, 1331–1333 (2022).
51. Jones, S. et al. Fast, three-dimensional super-resolution imaging of live

cells. *Nat. Methods* **8**, 499–505 (2011).

52. Godin A G, Lounis B, Cognet L. Super-resolution microscopy approaches for live cell imaging. *Biophys. J.* **107**.8, 1777-1784 (2014).

53. Song, A., Gauthier, J.L., Pillow, J.W., Tank, D.W. & Charles, A.S. Neural anatomy and optical microscopy (NAOMi) simulation for evaluating calcium imaging methods. *J. Neurosci. Methods* **358**, 109173 (2021).

54. Chen, TW. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).

55. Huang, F., Schwartz, S. L., Byars, J. M. & Lidke, K. A. Simultaneous multiple-emitter fitting for single molecule super-resolution imaging. *Biomed. Opt. Express* **2**, 1377–1393 (2011).

56. Platisa, J. et al. High-speed low-light in vivo two-photon voltage imaging of large neuronal populations. *Nat. Methods* (2023).

57. Zhao, W. et al. Sparse deconvolution improves the resolution of live-cell super-resolution fluorescence microscopy. *Nat. Biotechnol.* **40**, 606–617 (2022).

58. Dahmardeh, M. et al. Self-supervised machine learning pushes the sensitivity limit in label-free detection of single proteins below 10 kDa. *Nat. Methods* **20**, 442–447 (2023).

59. Li, X. et al. Unsupervised content-preserving transformation for optical microscopy. *Light. Sci. Appl.* **10**, 44 (2021).

60. Zhang, Y. et al. Fast and sensitive GCaMP calcium indicators for imaging neural populations. *Nature* **615**, 884–891 (2023).

61. Knöpfel, T., Song, C. Optical voltage imaging in neurons: moving from technology development to practical tool. *Nat. Rev. Neurosci.* **20**, 719–727 (2019).

62. Liu Z. et al. Sustained deep-tissue voltage recording using a fast indicator evolved for two-photon microscopy. *Cell* **185**, 3408-3425(2022).

63. Jimenez A, Friedl K, Leterrier C. About samples, giving examples: optimized single molecule localization microscopy. *Methods* **174** , 100-114 (2020).

64. Smith M B. et al. Segmentation and tracking of cytoskeletal filaments using open active contours. *Cytoskeleton* **67**, 693-705 (2010).

65. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. In *Proc. IEEE* **86**, 2278-2324 (1998).

66. He, Z. et al. Tracking by animation: Unsupervised learning of multi-object attentive trackers. In *Proc. of the IEEE/CVF conference on computer vision and pattern recognition* 1318-1327 (2019).

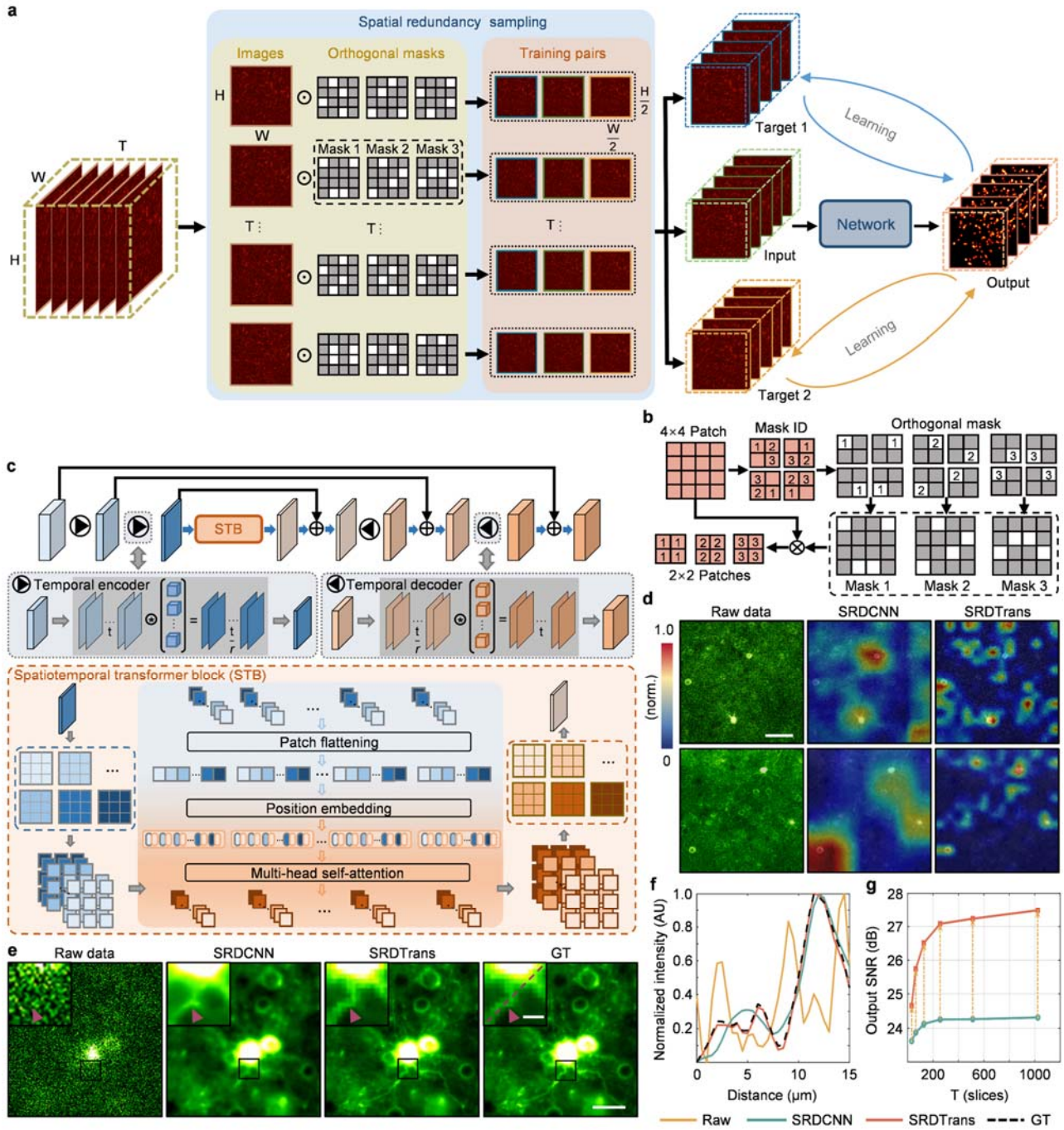


Fig. 1 | Principle of SRDTrans and performance evaluation. a, Self-supervised training strategy of SRDTrans. In a single instance of spatial redundancy sampling, the original low-SNR stack of H×W×T pixels is sampled by orthogonal masks, producing three downsampled sub-stacks (input, target 1, and target 2) of H/2×W/2×T pixels. The "input" sub-stack is fed into the transformer network, and the corresponding output is compared with the "target" sub-stacks to calculate the loss function for parameter optimization. **b**, Simplified schematic of spatial redundancy sampling (H=4, W=4, T=1). A 4×4 patch is split into four 2×2 blocks and three adjacent pixels are randomly selected in each block. The central pixel (labeled as "2") is horizontally or vertically adjacent to the other two pixels (labeled as "1" and "3"). Pixels with the same label are recombined into new 2×2 patches. **c**, The architecture of the lightweight spatiotemporal transformer. It consists of a temporal encoder module, a spatiotemporal transformer block (STB), and a temporal decoder module. Each temporal encoder compresses the temporal scale (t) of the input by a factor

of r (adjustable and set as 4 in this work) using convolution. In the STB module, the input is divided into small patches, and different feature maps of the same spatial position are stitched together in the patch flattening layer. The position embedding layer records the spatial position of each patch so that it can be mapped back after the global interaction in the multi-head self-attention layer. The self-attention mechanism can calculate the spatiotemporal correlation between all local patches. Then the output patches are remapped to their original spatial positions. The output of the STB module will be uncompressed to the original temporal scale by the following temporal decoder module. Multiple skip connections are configured to stabilize the training of the network. **d**, Visualizing the feature responses in deep layers of SRDCNN and SRDTrans. SRDCNN represents the method that replaces the transformer network in SRDTrans with a 3D-UNet. Scale bar, 60 μm. **e**, Comparing the denoising performance of SRDCNN and SRDTrans on simulated calcium imaging data (30 Hz). Magnified views of boxed regions are shown at the top left of the images. Purple arrowheads point to a dendritic fiber. Scale bar, 40 μm for the whole field of view (FOV) and 10 μm for magnified views. **f**, Pixel intensity along the red dashed line in **e**. **g**, Evaluating the ability of SRDCNN and SRDTrans to capture long-range dependencies. Models were trained and validated on simulated calcium imaging data (30 Hz) of different input temporal scales (T).

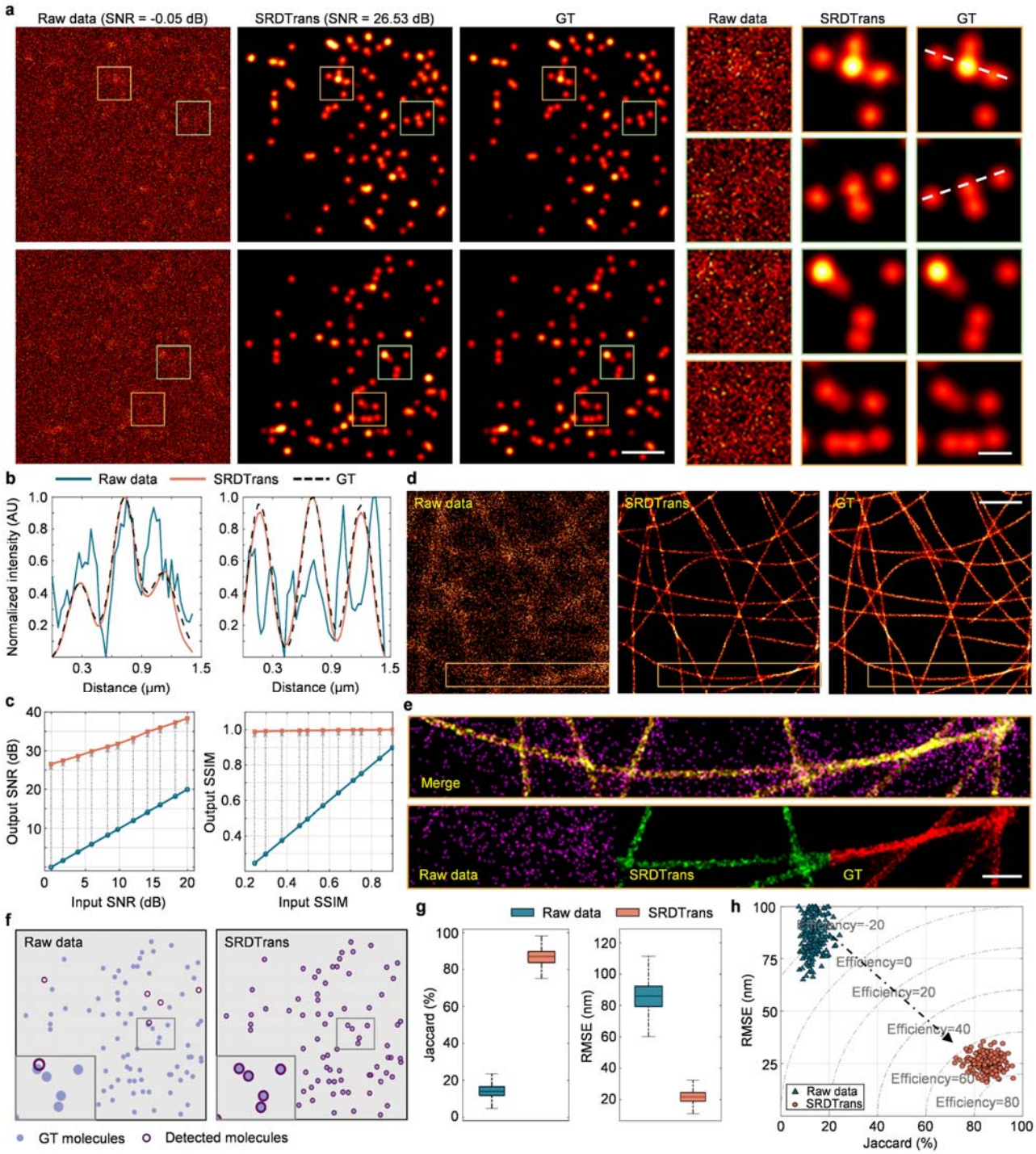


Fig. 2 | Validation of SRDTrans on simulated SMLM data. a, Single-molecule emission images before and after denoising. Left, raw data. Middle, SRDTrans denoised data. Right, GT. Magnified views of boxed regions show the emission pattern of a bunch of fluorescent molecules. Scale bars, 2 μm for the whole FOV and 0.5 μm for magnified views. The SNR value of the raw data and denoised are noted. **b**, Intensity profiles along the white dashed line in **a**. **c**, Quantitative evaluation of the denoising performance with SNR and SSIM. Left, image SNR before and after denoising. Right, image SSIM before and after denoising. Each data point is the statistical result of 24000 frames. **d**, Reconstructed super-resolution images of microtubules before and after denoising. Left, the image reconstructed from raw data. Middle, the image reconstructed

from SRDTrans denoised data. Right, GT. Scale bar, 5 μm. **e**, Merged image of the yellow boxed region in **d**. Magenta, the image reconstructed from raw data. Green, the image reconstructed from SRDTrans denoised data. Red, GT. The overlapping positions of red and green appear yellow. Scale bar, 1 μm. **f**, Consistency analysis of the localized fluorescent molecules in raw images (top) and SRDTrans denoised images (bottom) referring to GT. A magnified view of the boxed region is shown at the bottom left of each panel. **g**, Boxplots showing the localization precision quantified with the Jaccard index (left, higher is better) and RMSE (right, lower is better) before and after SRDTrans denoising (N=5000). **h**, Evaluating the performance of single-molecule localization before (blue) and after (yellow) denoising with a more comprehensive metric termed efficiency.

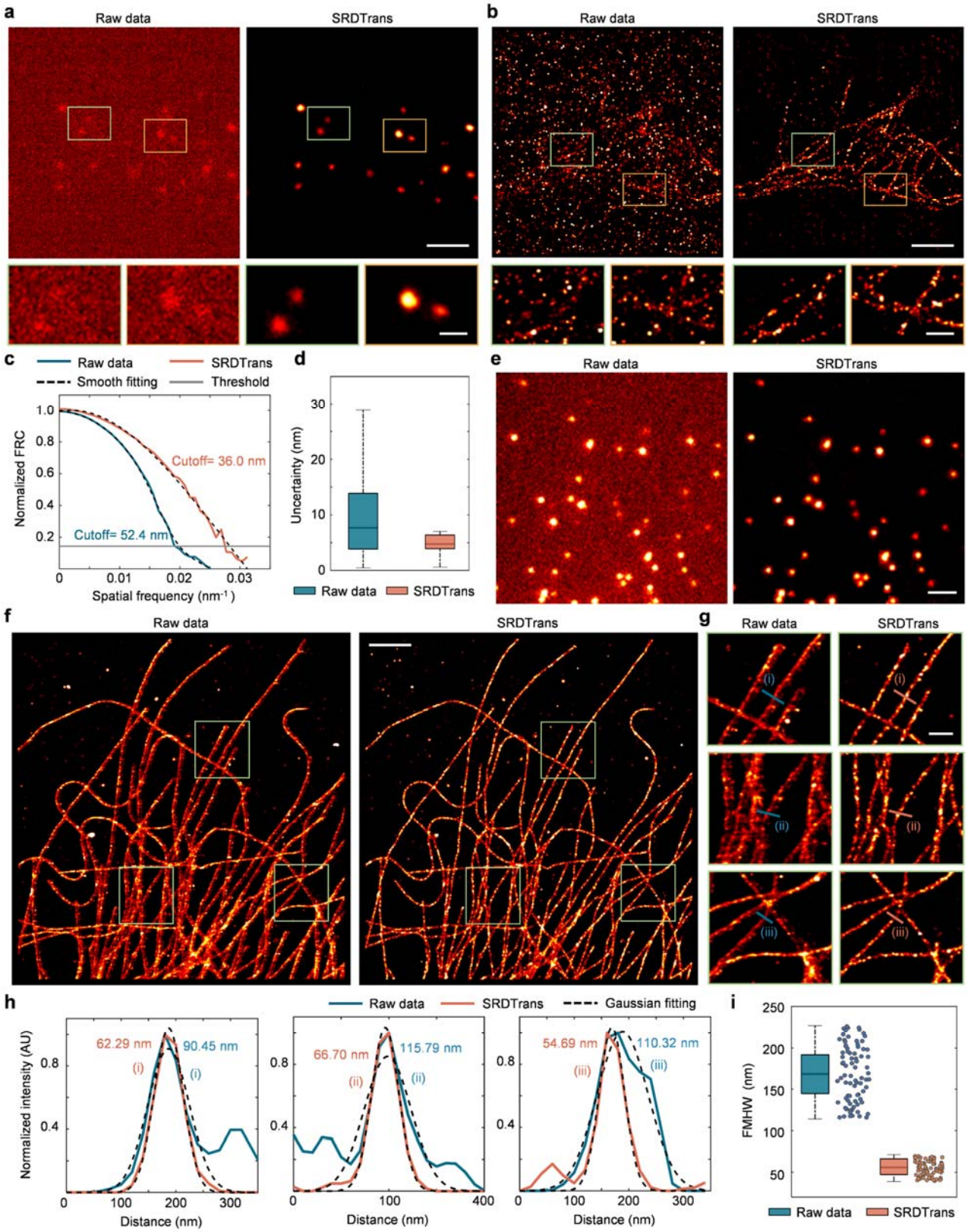
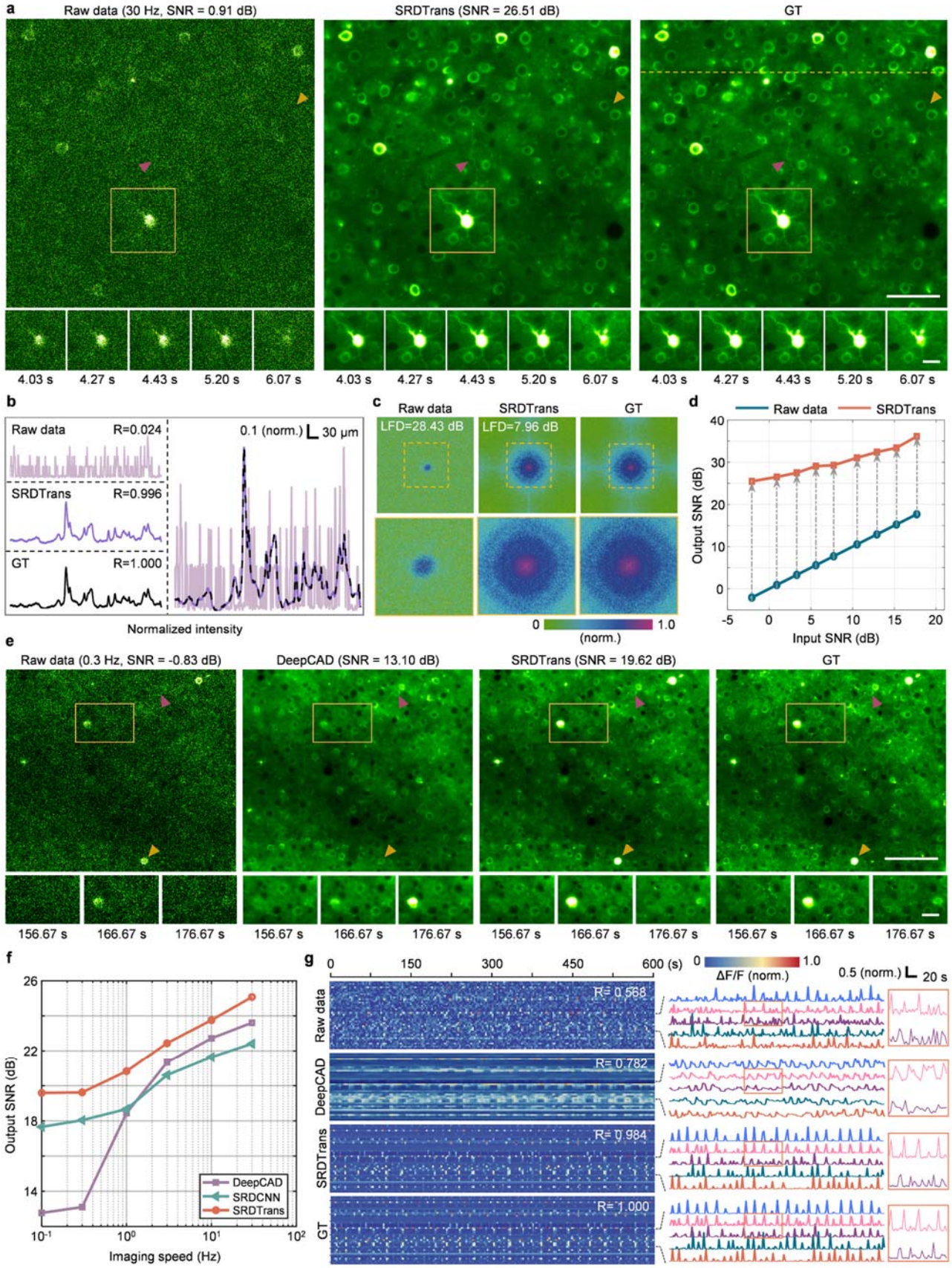


Fig. 3 | SRDTrans enhances experimental SMLM imaging. a, Experimentally obtained single-molecule emission images. Left, raw data. Right, SRDTrans denoised data. The magnified views of two boxed regions are shown below each image. Scale bars, 2 μm for the whole FOV and 0.5 μm for magnified views. **b**, Reconstructed super-resolution images. The microtubules in fixed BSC-1 cells were labeled with Cy5.

Scale bars, 2 μm for the whole FOV and 0.5 μm for magnified views. **c,** Fourier-ring correlation (FRC) curves of the raw reconstructed image (blue) and the SRDTrans enhanced reconstructed image (yellow) in **b**. The estimated resolution (52.4 nm for raw image and 36.0 nm for SRDTrans denoised image) is the inverse of the spatial frequency where the FRC curve drops below the cutoff threshold (~0.143). **d,** Boxplots showing the localizing uncertainty before and after denoising (N=1048575 for raw data, N=395908 for SRDTrans). The uncertainty was calculated by the ThunderSTORM plugin. **e,** Single-molecule emission images from the open-source platform ShareLoc. Left, raw data. Right, SRDTrans denoised data. Scale bars, 2 μm. **f**, Reconstructed super-resolution images of microtubules (immuno-labeled with Alexa 647). Scale bars, 2 μm. **g**, Magnified views of boxed regions in **f**. Scale bars, 0.5 μm. **h**, Intensity profiles perpendicular to the microtubule filaments as indicated in **f** (blue lines: raw data, yellow lines: SRDTrans denoised data, dashed line: the Gaussian fitting result). The corresponding FWHM value is quantified as 2.335σ, where σ denotes the standard deviation of Gaussian fitting results. **i**, Boxplots showing the FWHM of randomly selected filaments from **e** (blue: raw data, yellow: SRDTrans denoised data, N=80).


a
Raw data (30 Hz, SNR = 0.91 dB)
SRDTrans (SNR = 26.51 dB)
GT
4.03 s
4.27 s
4.43 s
5.20 s
6.07 s
b
Raw data
R=0.024
SRDTrans
R=0.996
GT
R=1.000
0.1 (norm.)
30 μm
Normalized intensity
c
Raw data
SRDTrans
GT
LFD=28.43 dB
LFD=7.96 dB
0
1.0
(norm.)
d
Raw data
SRDTrans
Output SNR (dB)
Input SNR (dB)
e
Raw data (0.3 Hz, SNR = -0.83 dB)
DeepCAD (SNR = 13.10 dB)
SRDTrans (SNR = 19.62 dB)
GT
156.67 s
166.67 s
176.67 s
f
Output SNR (dB)
Imaging speed (Hz)
DeepCAD
SRDCNN
SRDTrans
g
0
150
300
450
600 (s)
ΔF/F (norm.)
0.5 (norm.)
20 s
Raw data
DeepCAD
SRDTrans
GT
R= 0.568
R= 0.782
R= 0.984
R= 1.000

765

Fig. 4 | Evaluating the performance of SRDTrans on simulated calcium imaging data. a, SRDTrans denoised calcium imaging data sampled at 30 Hz. Magnified views show the neural activity of yellow boxed regions in a short period (~2 s). Left, the original low-SNR data. Middle, SRDTrans denoised data. Right, GT. Scale bars, 60 μm for the whole FOV and 30 μm for magnified views. The magenta arrowhead indicates a dendritic fiber and the yellow arrowhead indicates two neighboring somas. **b**, Pixel intensity along the yellow dashed line in **a**. Top left, raw data. Middle left, SRDTrans denoised data. Bottom left, GT. Right, plotting the three intensity profiles in one coordinate. The similarity with GT is quantified by Pearson correlation coefficients (R). **c**, Frequency spectrum calculated by discrete Fourier transform before and after denoising. Magnified views of the boxed regions show the frequency components within the optical transfer function. The similarity in the frequency domain is quantified by logarithmic frequency distance (LFD). **d**, The performance of SRDTrans at different SNR levels (averaged on 6000 frames). **e**. Comparing the performance of DeepCAD and SRDTrans on calcium imaging data sampled at 0.3 Hz. Magnified views show the neural activity of yellow boxed regions in a 20-s time window. Scale bars, 100 μm for the whole FOV and 40 μm for magnified views. The yellow and purple arrowheads point to a firing neuron and a resting neuron, respectively. **f**, Ablation experiments to investigate the effects of different sampling strategies and network architectures. A wide range of imaging speeds (from 0.1 Hz to 30 Hz) are surveyed and summarized in a logarithmic coordinate. SRDTrans (yellow) uses spatial redundancy sampling and a lightweight spatiotemporal transformer. DeepCAD (purple) combines temporal redundancy sampling and a CNN (3D-UNet). SRDCNN (green) is the method combining spatial redundancy sampling and a CNN (3D-UNet). The input patch size is 128×128×128 pixels for all methods. **g**, Fluorescence traces extracted from 50 randomly selected neuronal pixels. The similarity with GT is quantified by Pearson correlation coefficients (R). Top, traces extracted from raw data. Middle, traces extracted from DeepCAD denoised data. Middle bottom, traces extracted from SRDTrans denoised data. Bottom, GT.

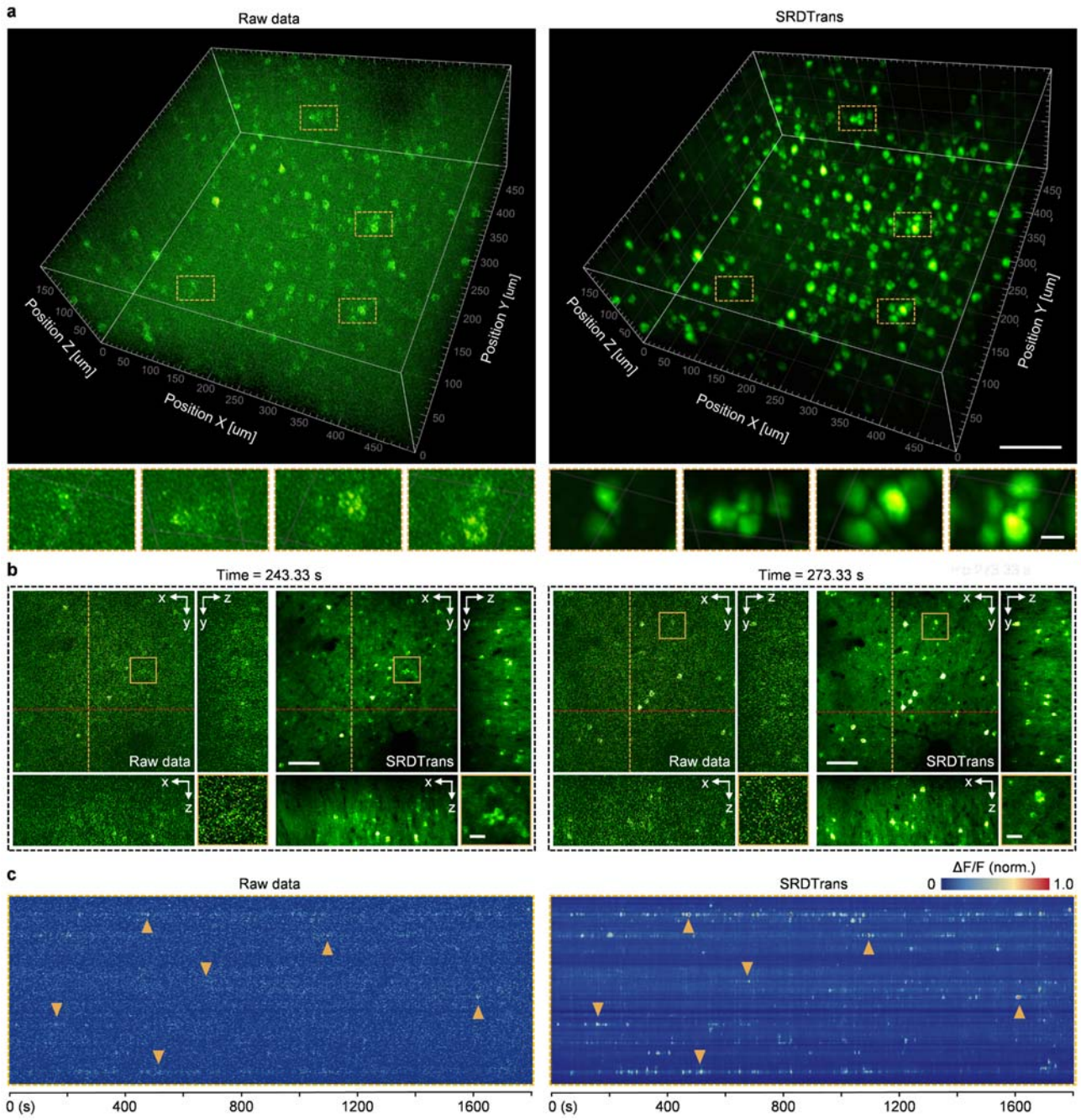
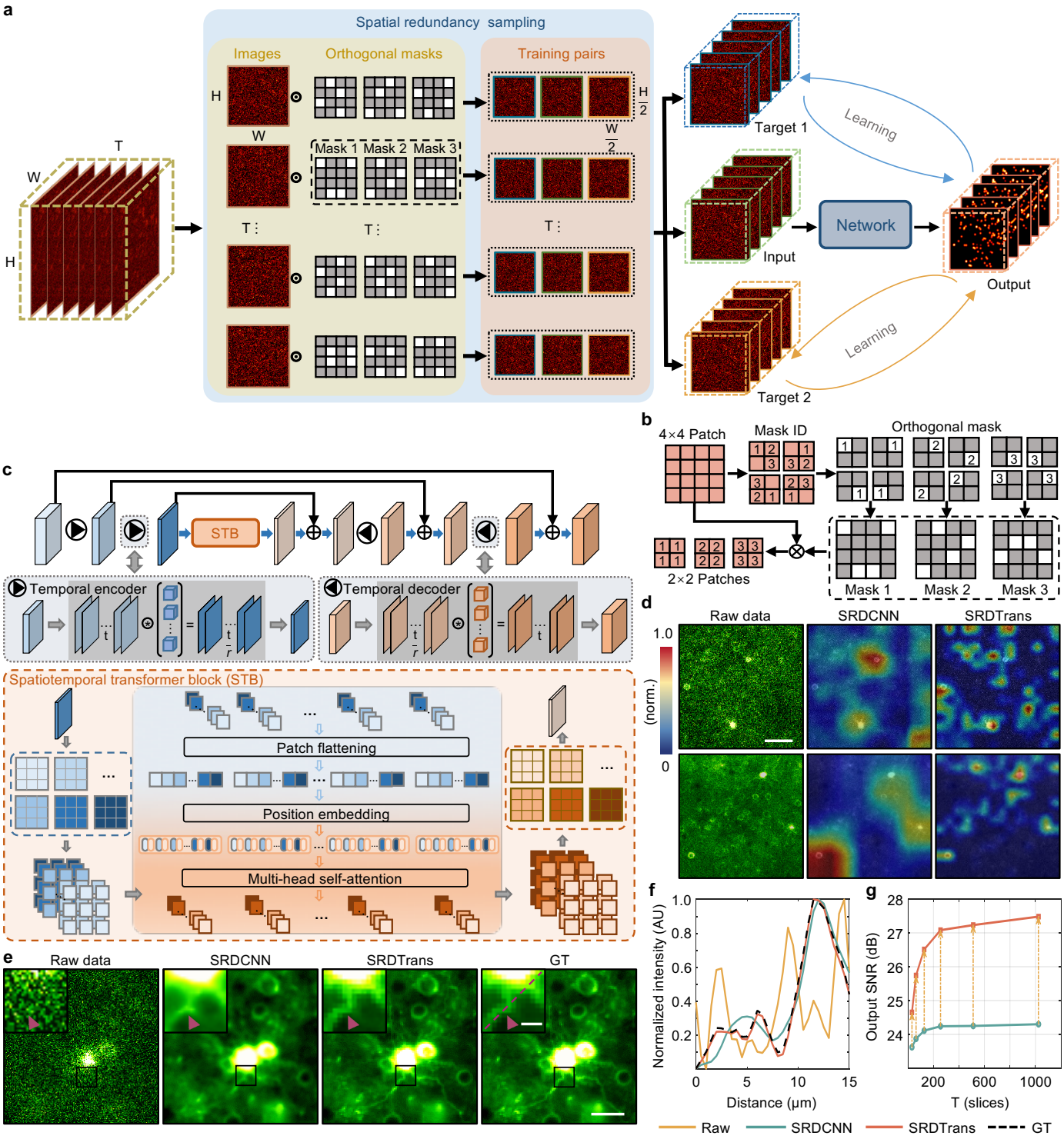
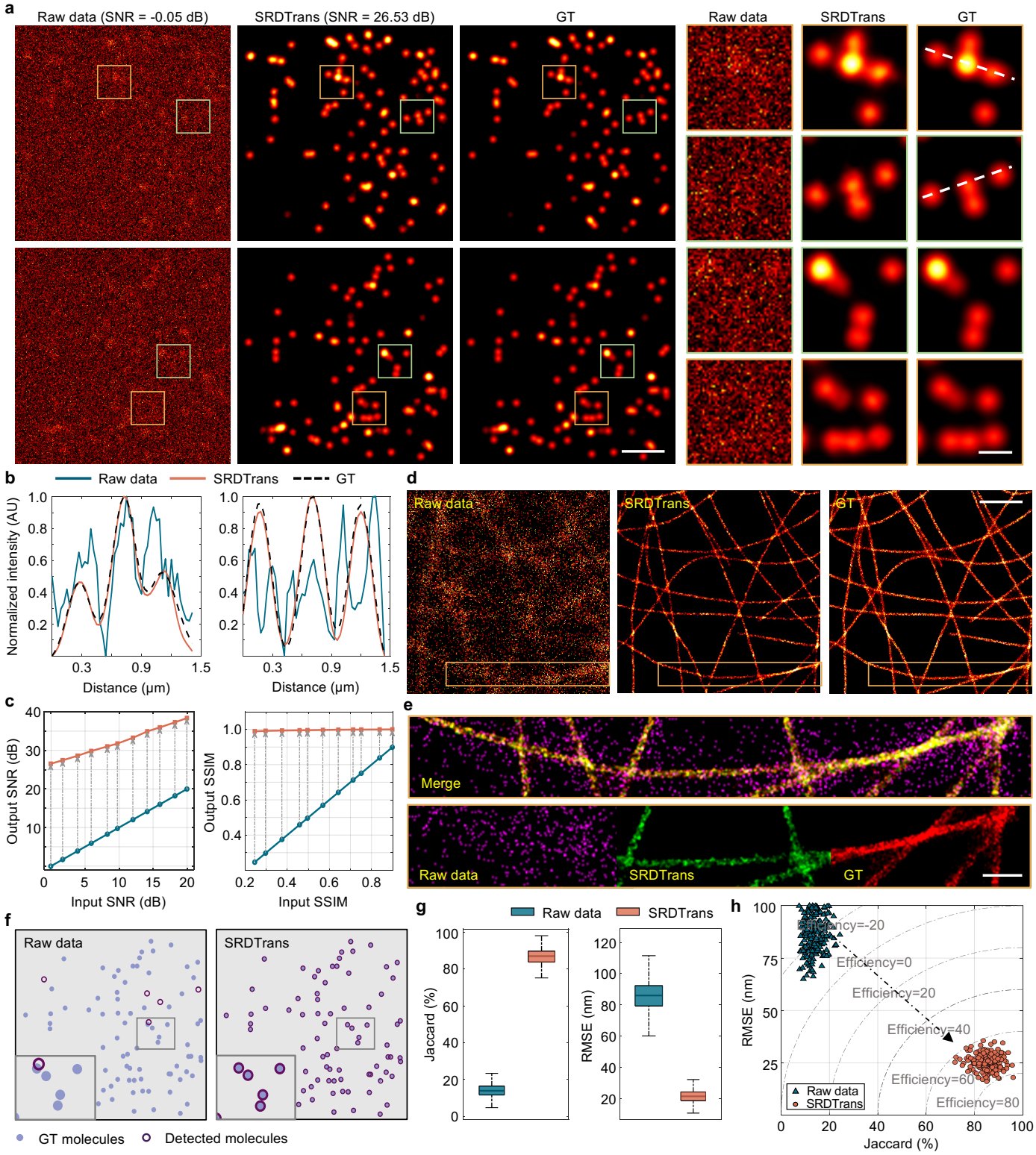
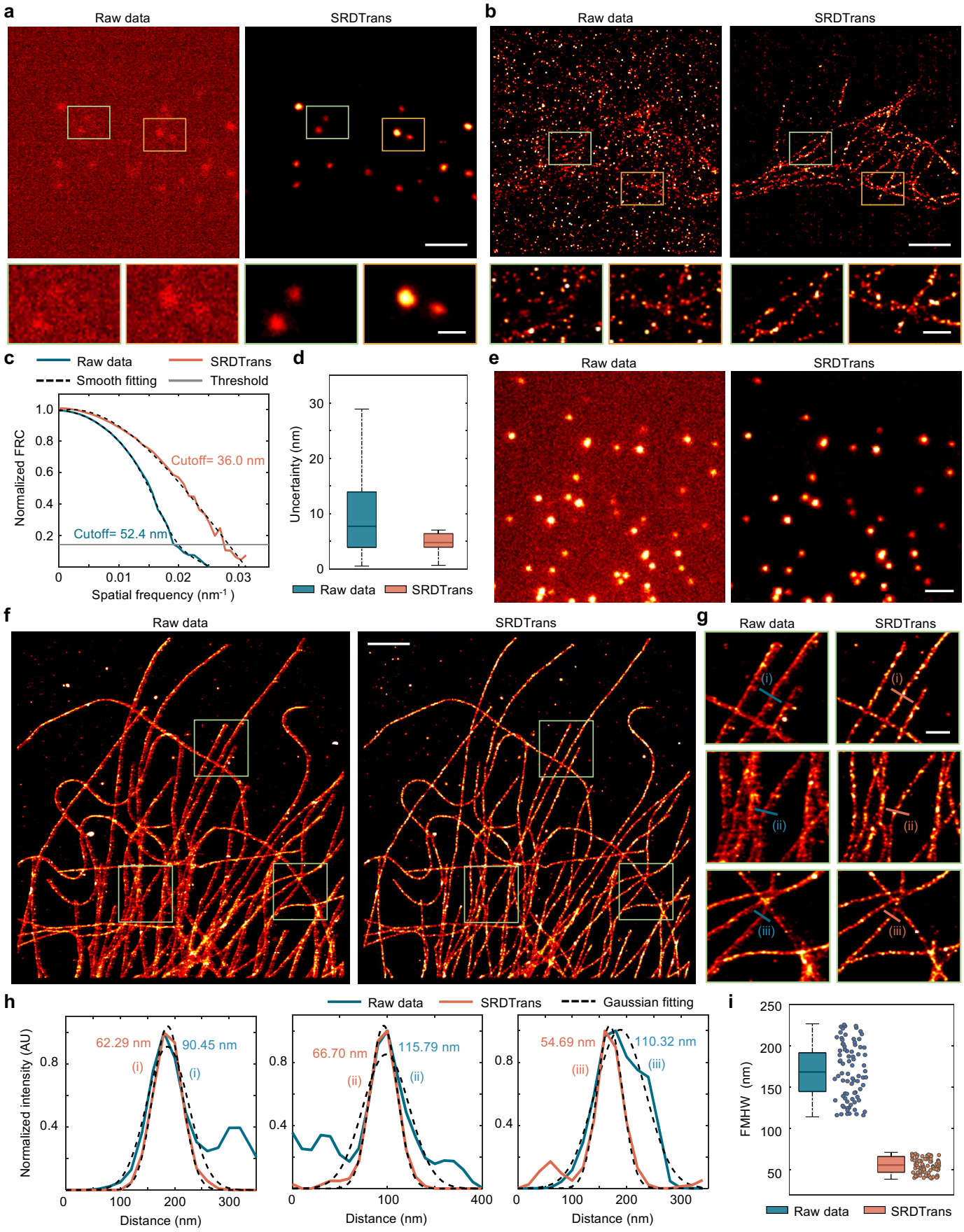
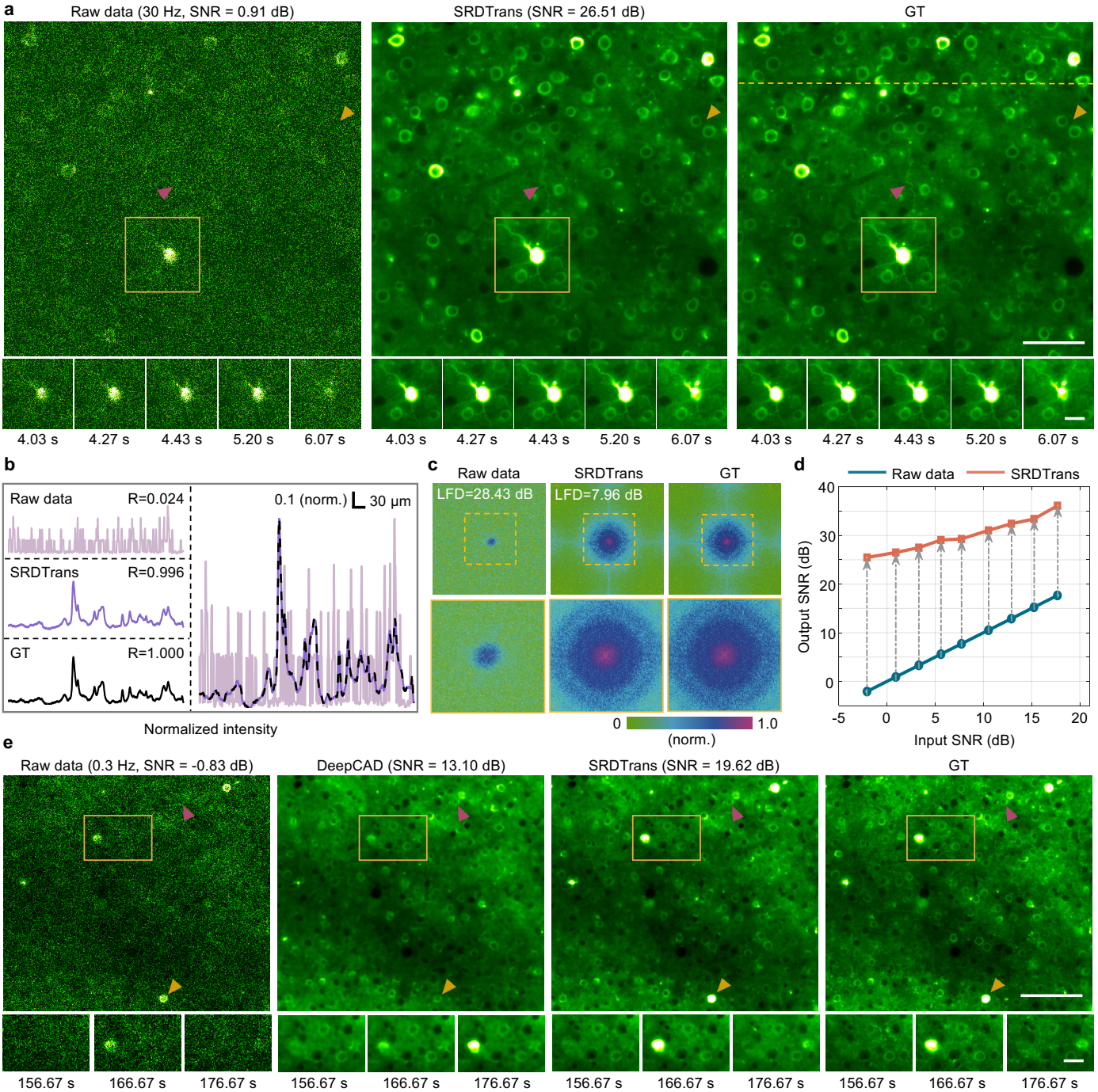


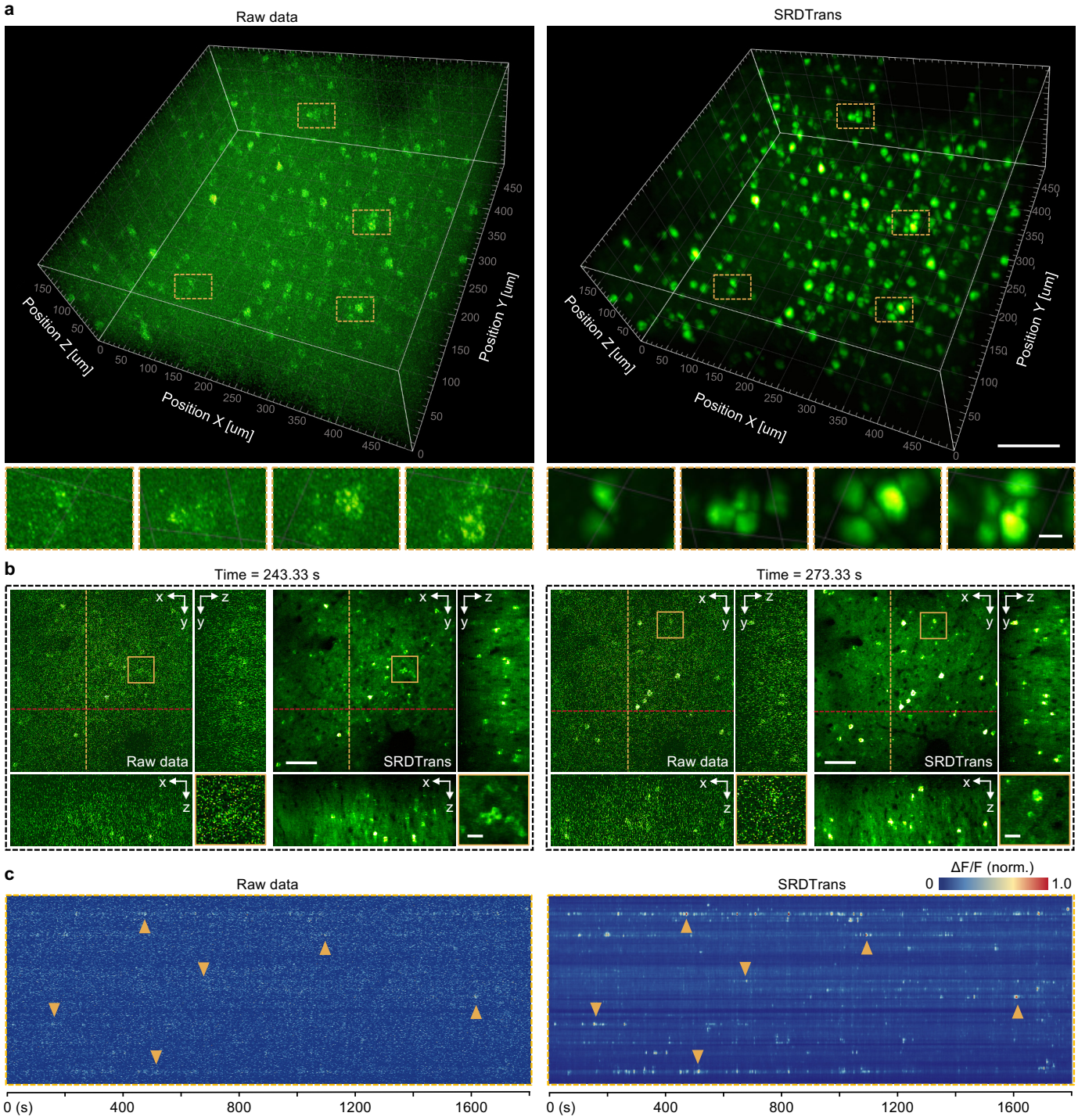
Fig. 5 | High-sensitivity calcium imaging of large neural volumes. a, 3D visualization of the neural activity of a 510×510×200 μm volume (100 planes, 0.3 Hz volume rate) in the mouse cortex. Left, the original low-SNR volume. Right, the same volume denoised with SRDTrans. Magnified views of yellow boxed regions are shown under each snapshot. Scale bars, 100 μm for the whole FOV and 10 μm for magnified views. **b**, Raw frames and SRDTrans denoised counterparts of a single imaging plane at two different moments. The x-z and y-z cross-sections of the volume indicated by the dashed lines are shown beside each x-y plane. Magnified views of yellow boxed regions are shown at the bottom right of the images. Scale bars, 70 μm for the whole FOV and 10 μm for magnified views. **c**, Fluorescence traces extracted from all pixels on the yellow dashed line in **b**, Left, traces extracted from raw data. Right, SRDTrans denoised traces. Yellow arrowheads point to some representative calcium transients.


a
Spatial redundancy sampling
Images
Orthogonal masks
Training pairs
Mask 1
Mask 2
Mask 3
Target 1
Learning
Input
Network
Output
Target 2
Learning
b
4×4 Patch
Mask ID
Orthogonal mask
2×2 Patches
Mask 1
Mask 2
Mask 3
c
STB
Temporal encoder
Temporal decoder
Spatiotemporal transformer block (STB)
Patch flattening
Position embedding
Multi-head self-attention
d
Raw data
SRDCNN
SRDTrans
(norm.)
e
Raw data
SRDCNN
SRDTrans
GT
f
Normalized intensity (AU)
Distance (μm)
g
Output SNR (dB)
T (slices)
Raw
SRDCNN
SRDTrans
GT


a
Raw data (SNR = -0.05 dB)
SRDTrans (SNR = 26.53 dB)
GT
Raw data
SRDTrans
GT
b
Raw data
SRDTrans
GT
Normalized intensity (AU)
Distance (μm)
c
Output SNR (dB)
Input SNR (dB)
Output SSIM
Input SSIM
d
Raw data
SRDTrans
GT
e
Merge
Raw data
SRDTrans
GT
f
Raw data
SRDTrans
GT molecules
Detected molecules
g
Raw data
SRDTrans
Jaccard (%)
RMSE (nm)
h
RMSE (nm)
Jaccard (%)
Efficiency=-20
Efficiency=0
Efficiency=20
Efficiency=40
Efficiency=60
Efficiency=80
Raw data
SRDTrans


a
Raw data
SRDTrans
b
Raw data
SRDTrans
c
Raw data
SRDTrans
Smooth fitting
Threshold
Normalized FRC
Cutoff= 36.0 nm
Cutoff= 52.4 nm
Spatial frequency (nm-1)
d
Uncertainty (nm)
Raw data
SRDTrans
e
Raw data
SRDTrans
f
Raw data
SRDTrans
g
Raw data
SRDTrans
(i)
(ii)
(iii)
h
Raw data
SRDTrans
Gaussian fitting
Normalized intensity (AU)
62.29 nm
90.45 nm
66.70 nm
115.79 nm
54.69 nm
110.32 nm
Distance (nm)
i
FMHW (nm)
Raw data
SRDTrans


a
Raw data (30 Hz, SNR = 0.91 dB)
SRDTrans (SNR = 26.51 dB)
GT
4.03 s
4.27 s
4.43 s
5.20 s
6.07 s
b
Raw data
R=0.024
SRDTrans
R=0.996
GT
R=1.000
0.1 (norm.)
30 μm
Normalized intensity
c
Raw data
SRDTrans
GT
LFD=28.43 dB
LFD=7.96 dB
0
1.0
(norm.)
d
Raw data
SRDTrans
Output SNR (dB)
Input SNR (dB)
e
Raw data (0.3 Hz, SNR = -0.83 dB)
DeepCAD (SNR = 13.10 dB)
SRDTrans (SNR = 19.62 dB)
GT
156.67 s
166.67 s
176.67 s
f
Output SNR (dB)
Imaging speed (Hz)
DeepCAD
SRDCNN
SRDTrans
g
0
150
300
450
600 (s)
ΔF/F (norm.)
0.5 (norm.)
20 s
Raw data
DeepCAD
SRDTrans
GT
R= 0.568
R= 0.782
R= 0.984
R= 1.000


a
Raw data
SRDTrans
Position Z [um]
Position X [um]
Position Y [um]
b
Time = 243.33 s
Time = 273.33 s
Raw data
SRDTrans
c
Raw data
SRDTrans
ΔF/F (norm.)
0
1.0
0 (s)
400
800
1200
1600