Supplementary Information

4D self-supervised learning of cross-dimensional representation enables high-performance denoising of volumetric fluorescence imaging

Yixin Li[1], Tianyu Ma[2], Qi Zhang[2], Zhifeng Zhao[2], Xirou Zhou[3], Yibo Qu[2], Xiaoge Wang[2], Wentao Chen[1], Zhi Lu[4], Ziwei Li[1,5], Xinyang Li[3*], Jiamin Wu[2,6*] & Qionghai Dai[2,6*]

[1]*College of Future Information Technology, Fudan University, Shanghai, China.*

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

[3]*College of AI, Tsinghua University, Beijing, China.*

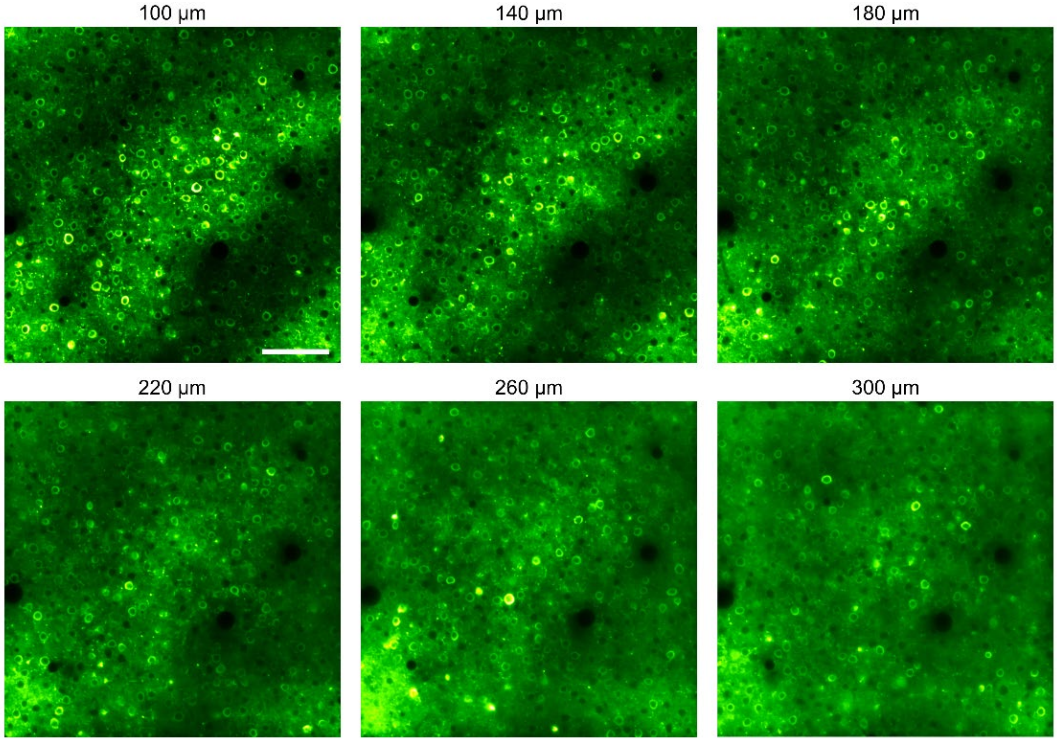
[4]*Department of Psychological and Cognitive Sciences, Tsinghua University, Beijing, China.*

[5]*Key Laboratory for Information Science of Electromagnetic Waves (MoE), Fudan University, Shanghai, China.*

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

[*]*Correspondence: xinyangli@tsinghua.edu.cn; wujiamin@tsinghua.edu.cn; qhdai@tsinghua.edu.cn*

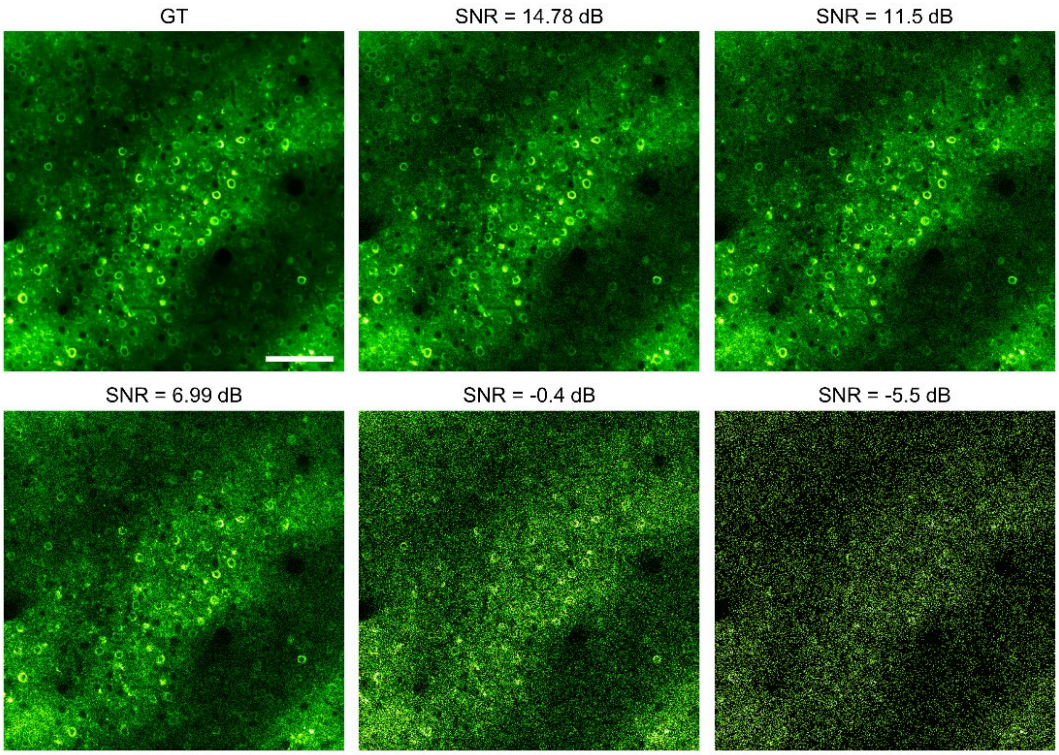
I. Supplementary Figures



Supplementary Figure 1

Examples of simulated calcium imaging data with different depths.

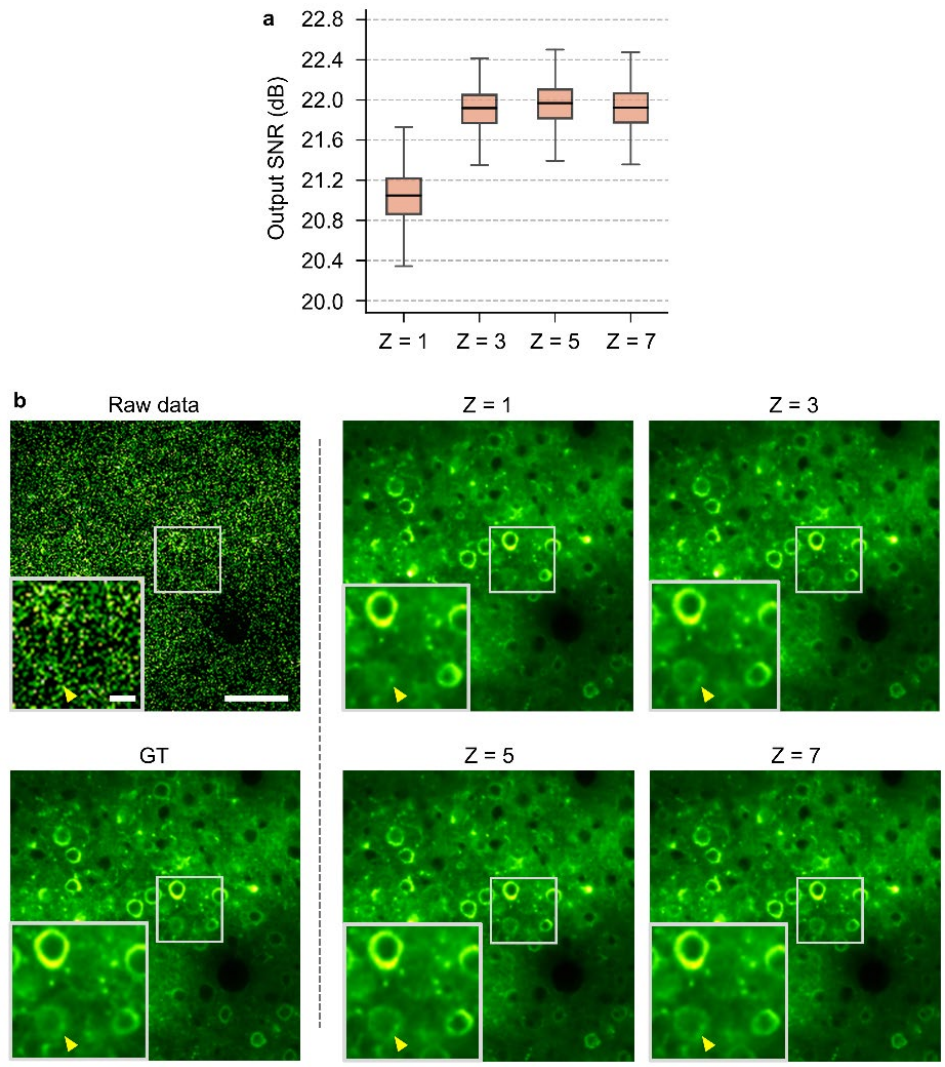
Representative x–y planes from a simulated 4D calcium imaging volume generated using an improved Neural Anatomy and Optical Microscopy (NAOMi)[1]-based simulation framework. The imaging depth is indicated above each image. Scale bar, 100 µm.



Supplementary Figure 2

Examples of simulated calcium imaging data with different SNRs.

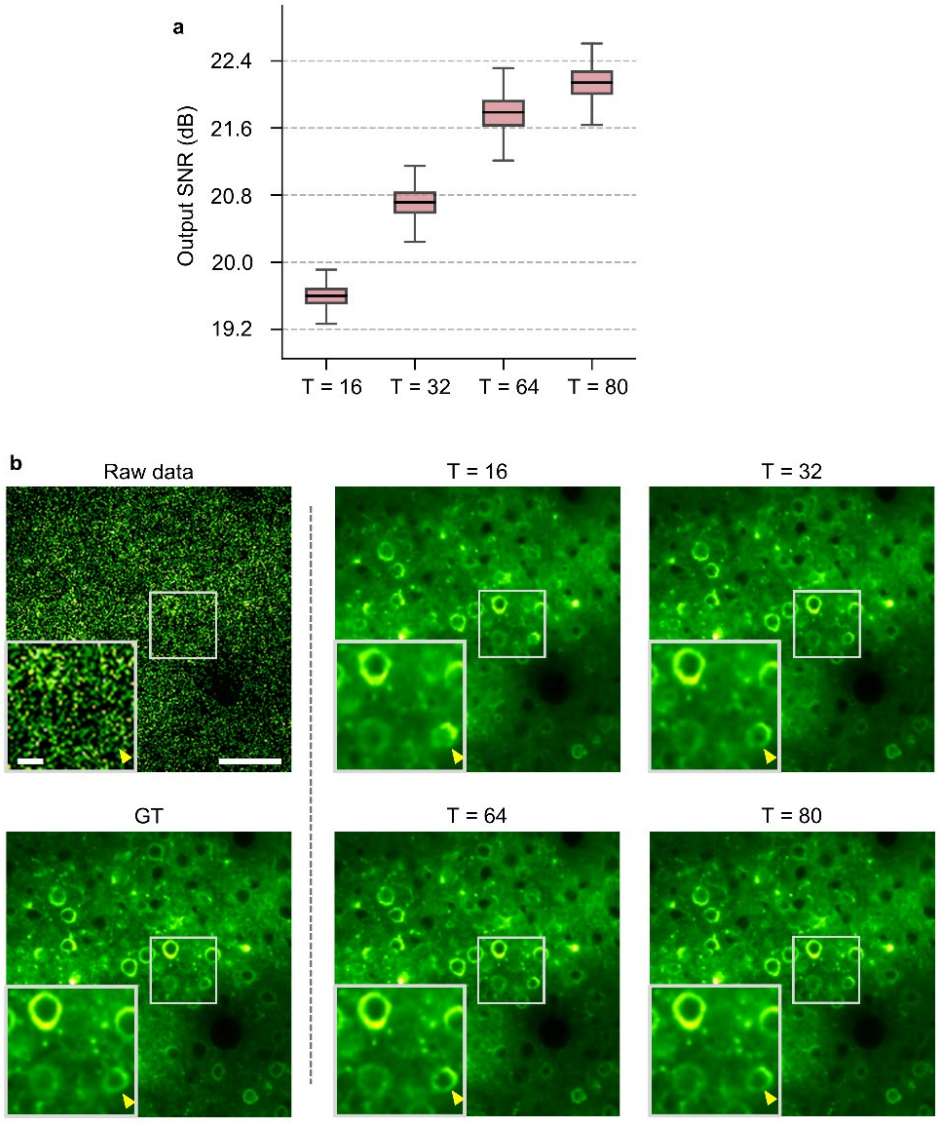
Mixed Poisson–Gaussian (MPG) noise was added to the simulated 4D calcium imaging volume to generate low-SNR observations. Representative images with different SNR levels corresponding to the same ground-truth (GT) data are shown. Scale bar, 100 μm.



Supplementary Figure 3

Performance of SDT-4D with different input axial scales.

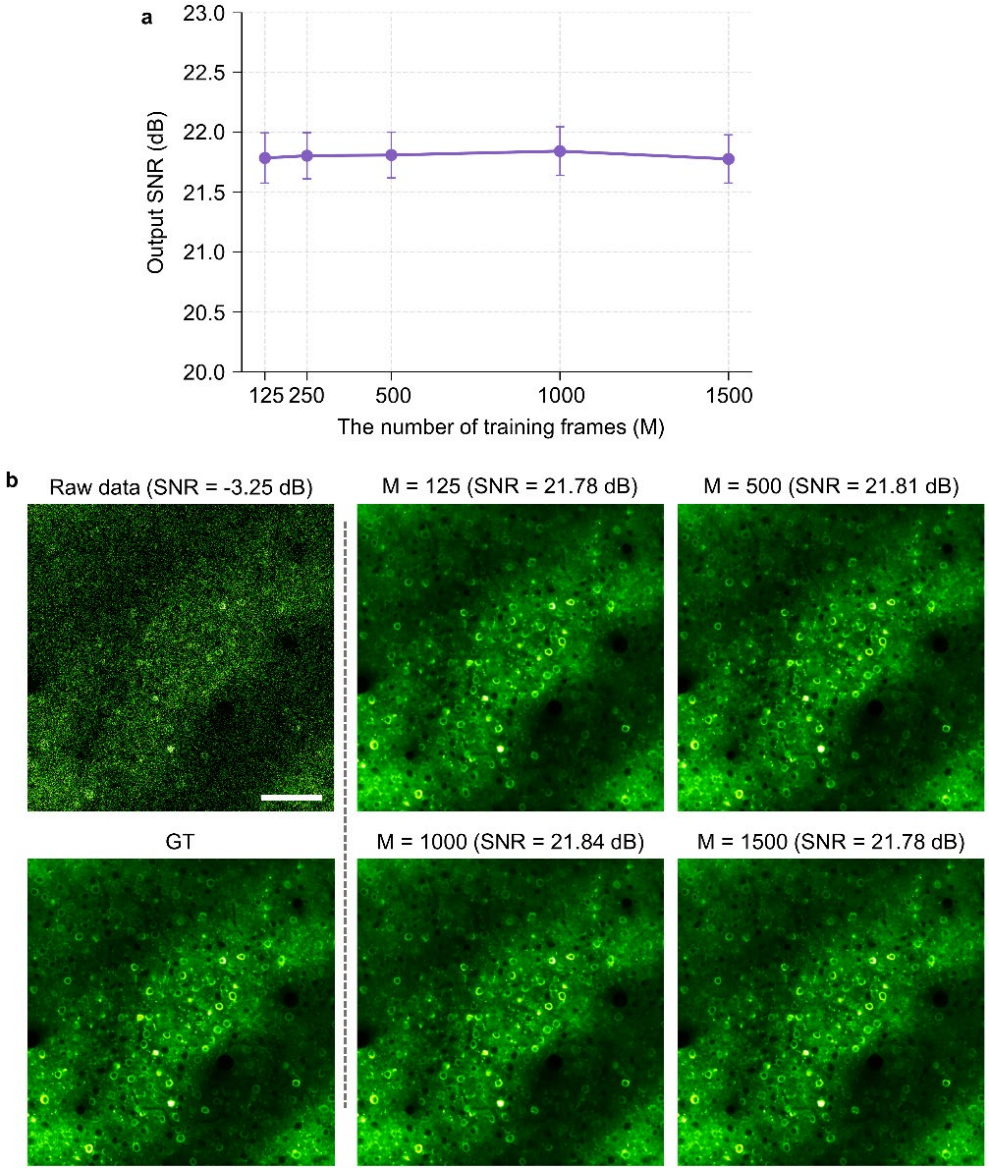
a, Quantitative comparison of denoising performance with different input axial scales. For SDT-4D training, the input patch size was Z × 64 × 64 × 64 pixels along the Z, T, H and W dimensions, where Z = 1, 3, 5 and 7. With an axial sampling interval of 5 μm, Z = 1, 3, 5 and 7 corresponded to axial spans of 0, 10, 20 and 30 μm between the first and last input planes, respectively. The models were trained on simulated volumetric calcium imaging data with a 1 Hz volume rate, an input SNR of −2.52 dB, 41 z-planes and 1,500 volumes. Box plots show the output SNR (N = 1,440 frames). **b**, Representative denoised images obtained with different input axial scales and the corresponding raw data and GT. Magnified views of the boxed regions are shown at the bottom left of each image. Yellow arrowheads mark a representative weak neuronal structure for comparison across conditions. Scale bars, 50 μm for the whole FOV and 10 μm for magnified views.



Supplementary Figure 4

Performance of SDT-4D with different input temporal scales.

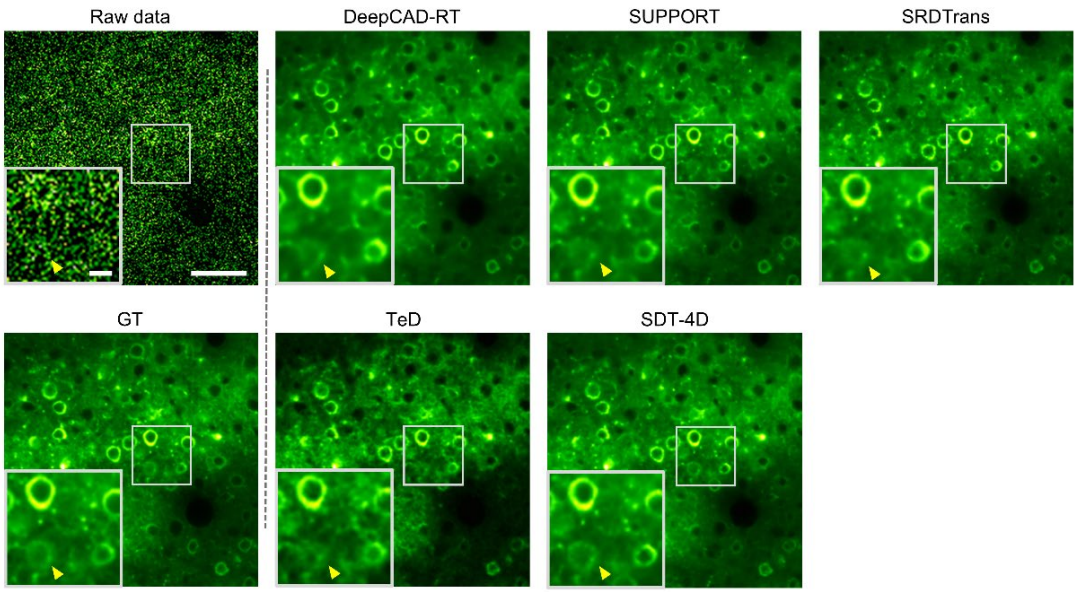
a, Quantitative comparison of denoising performance with different input temporal scales. For SDT-4D training, the input patch size was 5 × T × 64 × 64 pixels along the Z, T, H and W dimensions, where T = 16, 32, 64 and 80. The models were trained on simulated volumetric calcium imaging data with a 1 Hz volume rate, an input SNR of −2.52 dB, 41 z-planes and 1,500 volumes. Box plots show the output SNR (N = 1,440 frames). **b**, Representative denoised images obtained with different input temporal scales and the corresponding raw data and GT. Magnified views of the boxed regions are shown at the bottom left of each image. Yellow arrowheads mark a representative weak neuronal structure for comparison across conditions. Scale bars, 50 μm for the whole FOV and 10 μm for magnified views.



Supplementary Figure 5

Evaluating the data dependency of SDT-4D.

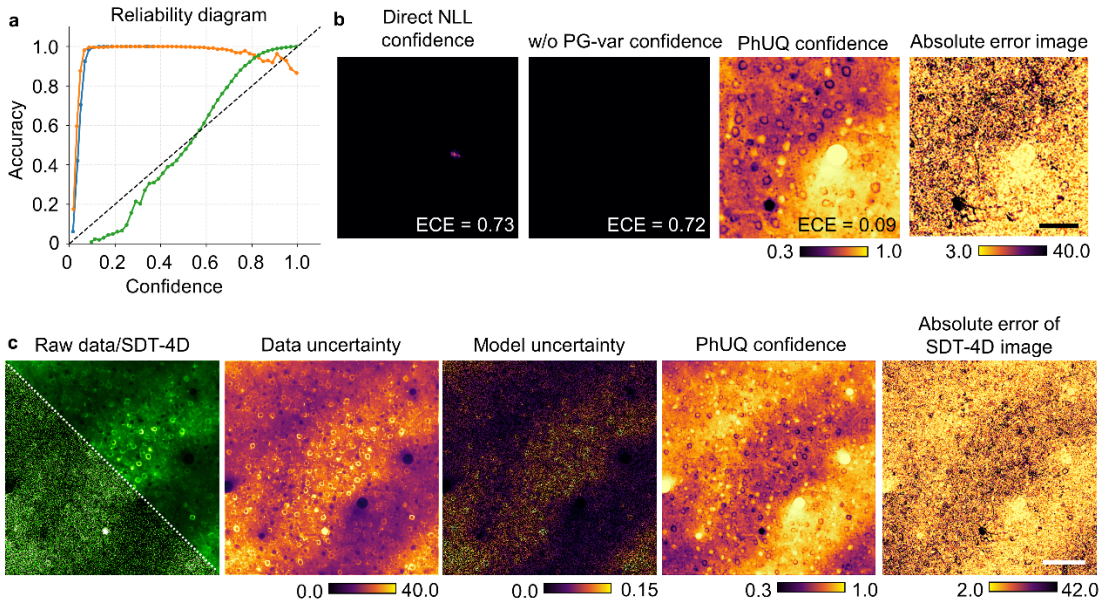
SDT-4D models were trained with simulated volumetric calcium imaging data with a 1 Hz volume rate, an input SNR of −3.25 dB and 41 z-planes. The size of each training image stack was 41× M × 490 × 490 pixels along the Z, T, H and W dimensions, where M denotes the number of training frames selected from {125, 250, 500, 1,000, 1,500}. Each model was trained for 65 epochs. **a**, Quantitative evaluation of denoising performance with different numbers of training frames. Output SNR is plotted as a function of M. The line connects mean values, and error bars represent s.d. (N = 1,440 frames). **b**, Representative denoising results of models trained with different numbers of frames. Raw data and the corresponding GT are shown in the leftmost panels. Scale bar, 100 μm.



Supplementary Figure 6

Comparison of SDT-4D with baseline self-supervised denoising methods.

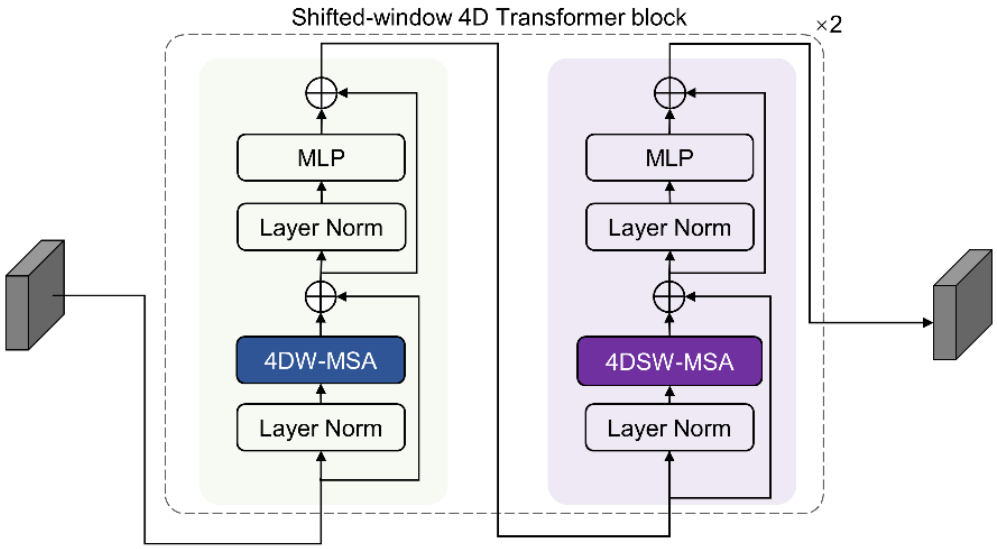
Each method was trained from scratch on simulated volumetric calcium imaging data with a 1 Hz volume rate, an input SNR of −3.25 dB and 41 z-planes. Because the baseline methods operate on xy–t stacks, the 4D volumes were split along the axial dimension for baseline training and inference. Representative denoising results of DeepCAD-RT[2], SUPPORT[3], SRDTrans[4], TeD[5] and SDT-4D are shown together with the raw data and GT. Magnified views of the boxed regions are shown at the bottom left of each image. Yellow arrowheads mark a representative weak neuronal structure for comparison across conditions. Scale bars, 50 μm for the whole FOV and 10 μm for magnified views.



Supplementary Figure 7

Calibration assessment and uncertainty maps of PhUQ.

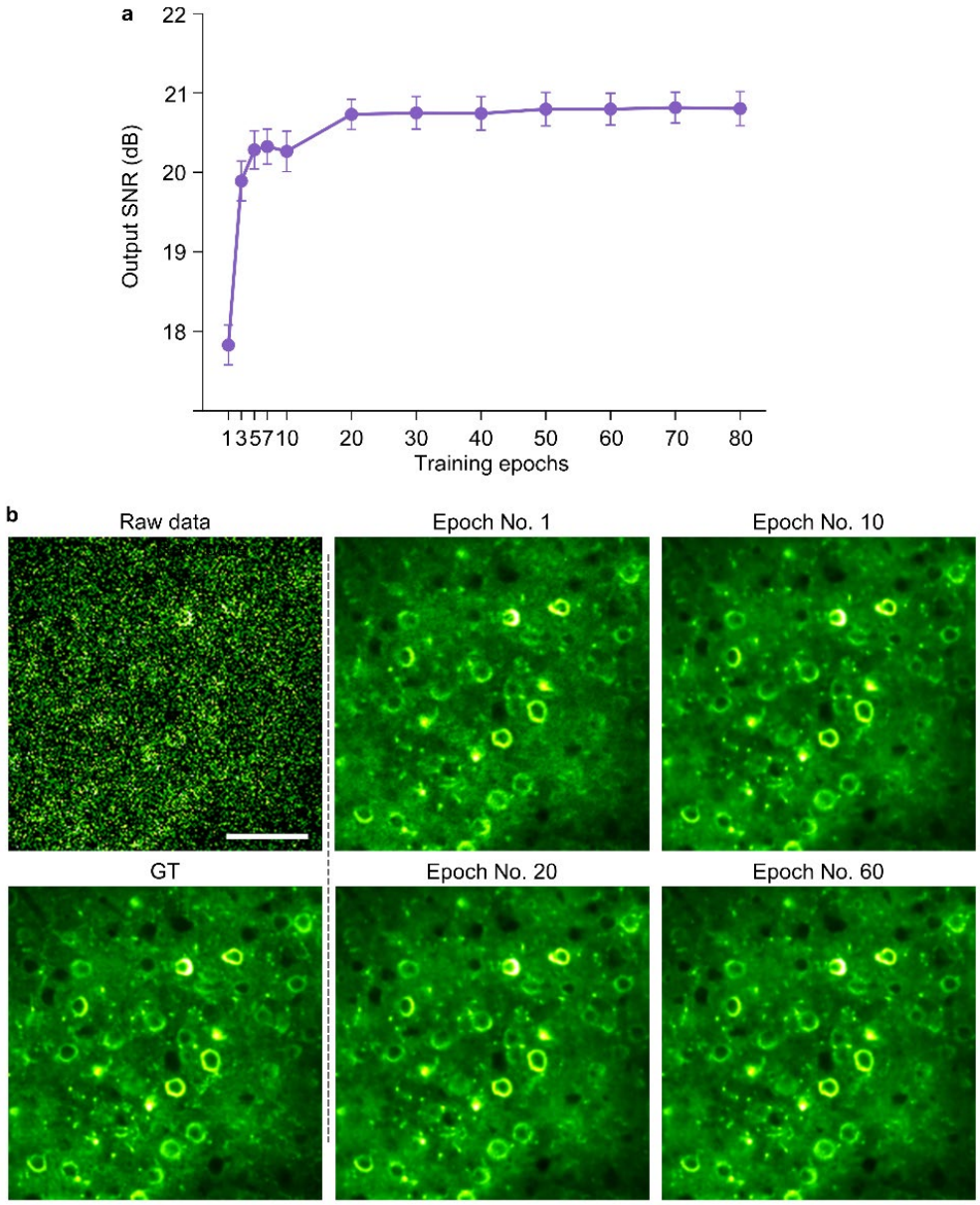
PhUQ was trained on simulated volumetric calcium imaging data with a 1 Hz volume rate, an input SNR of -2.52 dB and 41 z-planes. **a**, Reliability diagrams of Direct NLL, w/o PG-var and PhUQ. Accuracy is plotted against confidence, and the gray dashed line indicates ideal calibration. **b**, Representative confidence maps generated by Direct NLL, w/o PG-var and PhUQ. The absolute error map between the SDT-4D output and GT is shown for reference. The expected calibration error (ECE) is noted in each confidence map. Scale bar, 50 μm. **c**, Representative uncertainty and confidence maps generated by PhUQ. From left to right: raw data and SDT-4D denoised data, data uncertainty, model uncertainty, PhUQ confidence and absolute error of the SDT-4D image. Scale bar, 100 μm.



Supplementary Figure 8

Architecture of the four-dimensional interaction transformer block in SDT-4D.

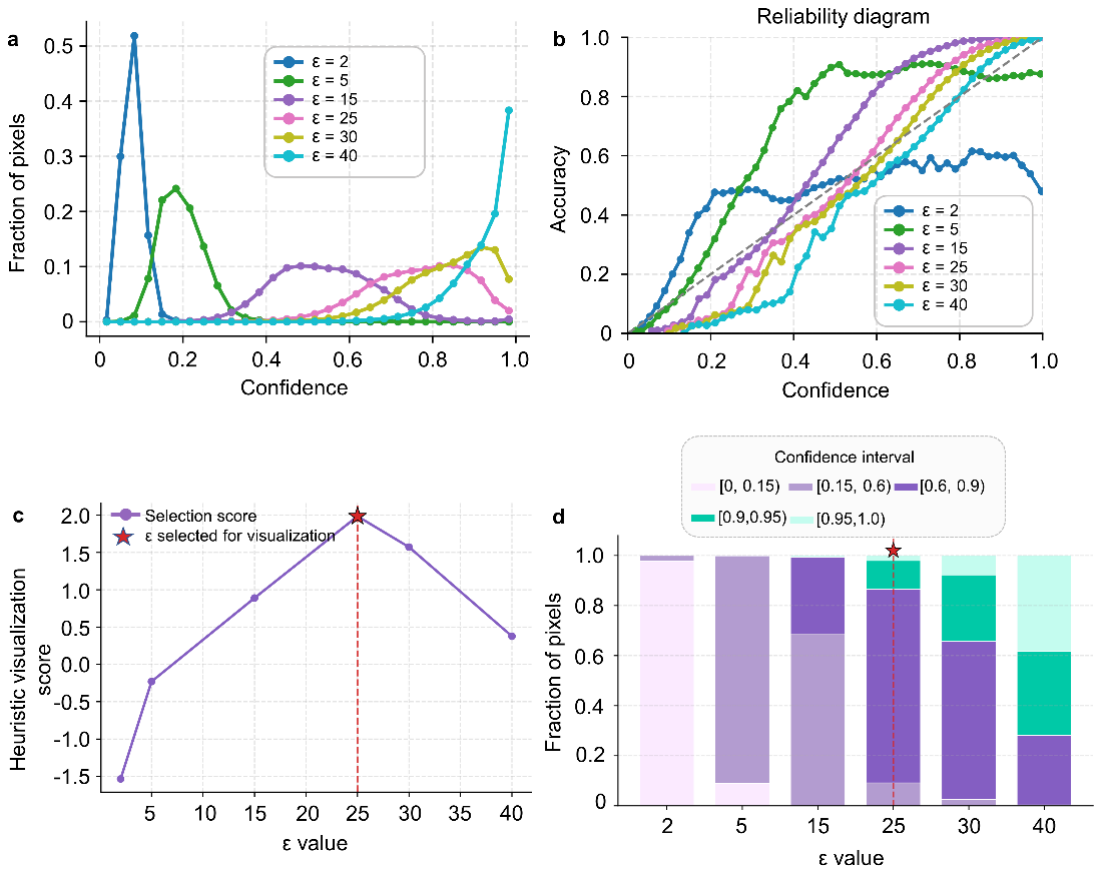
The four-dimensional interaction transformer block (4DITB) contains two cascaded shifted-window 4D Transformer blocks. Each block sequentially applies four-dimensional window multi-head self-attention (4DW-MSA) and four-dimensional shifted-window multi-head self-attention (4DSW-MSA). Each attention module is combined with layer normalization, an MLP and residual connections. The 4DW-MSA module captures interactions within local 4D windows, whereas 4DSW-MSA enables information exchange across neighboring windows by shifting the window partition[6].



Supplementary Figure 9

Training stability of SDT-4D.

Simulated volumetric calcium imaging data with a 1 Hz volume rate, an input SNR of −3.25 dB and 41 z-planes were used for quantitative evaluation. SDT-4D was trained for 80 epochs, and the model of each epoch was tested. **a**, Denoising performance of SDT-4D with increasing training epochs. Output SNR is plotted as a function of the training epoch. All values are shown as mean ± s.d. (N = 1,440 frames). **b**, Representative denoised images obtained at different training epochs. Raw data and the corresponding GT are shown in the leftmost panels. Scale bar, 50 μm.



Supplementary Figure 10

Sensitivity of PhUQ confidence maps to ε.

a, Confidence distributions obtained with different ε values. The fraction of pixels is plotted as a function of confidence. **b**, Reliability diagrams obtained with different ε values. Accuracy is plotted against confidence, and the gray dashed line indicates ideal calibration. **c**, Heuristic visualization score for candidate ε values. The red star indicates the ε selected for confidence-map visualization. **d**, Distribution of confidence mass over different confidence intervals for each ε.

II. Supplementary Videos

Supplementary Video 1	Denoising performance of SDT-4D on simulated volumetric two-photon calcium imaging data.
Supplementary Video 2	SDT-4D restores neuronal structures across multiple axial planes in simulated volumetric two-photon calcium.
Supplementary Video 3	SDT-4D enhances three-dimensional visualization of immune-cell dynamics in vivo.
Supplementary Video 4	SDT-4D reveals dynamic morphological changes of migrating immune cells in vivo.
Supplementary Video 5	SDT-4D facilitates three-dimensional observation of microglial and astrocytic morphology.
Supplementary Video 6	SDT-4D reveals fine morphological features of microglia and astrocytes in vivo.

References

1. 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).
2. Li, X. et al. Real-time denoising enables high-sensitivity fluorescence time-lapse imaging beyond the shot-noise limit. *Nat. Biotechnol.* **41**, 282–292 (2023).
3. Eom, M. et al. Statistically unbiased prediction enables accurate denoising of voltage imaging data. *Nat. Methods* **20**, 1581–1592 (2023).
4. Li, X. et al. Spatial redundancy transformer for self-supervised fluorescence image denoising. *Nat. Comput. Sci.* **3**, 1067–1080 (2023).
5. Lee, W. et al. Self-supervised denoising of dynamic fluorescence images via temporal gradient-empowered deep learning. *PhotoniX* **6**, 15 (2025).
6. Kim, P. et al. SwiFT: Swin 4D fMRI Transformer. In *Adv. Neural Inf. Process. Syst.*, 42015–42037 (2023).