Supplementary Information for

Lattice light sheet activation structured illumination volumetric super-resolution live microscopy

Xue Dong*, Quan Meng*, Xiaoyu Yang*, Haoyu Chen*, Chang Qiao*,
Yuhuan Lin, Siwei Zhang, Xiaohan Geng, Linghui Luan, Tao Jiang, Wenfeng Fu,
Amin Jiang, Wencong Xu, Jiabao Guo, Rongfei Wei, Dong Li†

*These authors contributed equally
†Correspondence: li-dong@tsinghua.edu.cn

This PDF file includes:

Contents

Supplementary Notes

Supplementary Note 1 | rsFPs properties in live-cell imaging

Reversibly photo-switchable fluorescent proteins (rsFPs)[1,2] are essential to our method, as imaging performance depends closely on the switching kinetics of these proteins. Conventional rsFPs are typically classified by their switching mechanisms into two primary modes: positive and negative. In the negative mode, the wavelength that induces fluorescence also switches the rsFPs from the on-state to the off-state. In contrast, in the positive mode, the light that excites fluorescence transfers the protein from the off-state to the on-state. The inherent characteristics of rsFPs can significantly affect imaging performance. For live-cell imaging, there are four key characteristics, including the brightness of the protein in its on-state, the on-off switching speed, the fluorescence contrast ratio of on/off states and the switching fatigue, which influence image intensity, imaging speed, image contrast and imaging duration, respectively.

Skylan-NS[3], a truly monomeric rsFP, was first developed by Xi Zhang et al. and tailored for patterned activation nonlinear structured illumination microscopy (PA NL-SIM)[4]. Comparing with other rsFPs, such as Dronpa[5] and rsEGFP2[6], Skylan-NS offers a superior number of switching cycles, higher photon output per cycle, and a more favorable on/off contrast ratio[3]. In live-cell experiments, Skylan-NS effectively labels various cellular components without inducing artificial aggregation, demonstrating its monomeric nature and suitability for cellular labeling.

Supplementary Note 2 | Detailed optical layout

The schematic of the optical system is shown in Supplementary Fig. 1. Here for simplicity, we only show the activation light (405 nm) and excitation light (488 nm) in this figure's main part. Actually, in our LA-SIM system, five lasers with wavelengths of 405 nm (365 mW, Cobolt, 06-MLD-405 nm), 445 nm (400 mW, Cobolt, 06-MLD-445 nm), 488 nm (500 mW, Coherent, Sapphire 488-500 LPX), 560 nm (500 mW, MPB Communications, 2RU-VFL-P-500-560-B1R) and 642 nm (500 mW, MPB Communications, 2RU-VFL-P-500-642-B1R) are expanded to a $1/e^2$ diameter of 2.5 mm using two lenses and then combined by a reflecting mirror and four dichroic mirrors (see inset in Supplementary Fig. 1). After combination, the lasers pass through an acousto-optic tunable filter (AOTF, AA Quanta Tech, AOTFnC-400.650-TN) which is employed to select the desired wavelength and control the laser intensity. Following the AOTF, a half wave plate (Bolder Vision Optik, BVO AHWP3) and a polarization beam splitter (Edmund, #49002) split the laser beam into two illumination paths. In the light sheet generation path, the laser passes through an achromatic doublet lens (L1, 20 mm FL/12.5 mm dia.) and a pair of cylindrical lenses (CL1, 50 mm FL/25.4 mm dia, Thorlabs, ACY254-50A; CL2, 250 mm FL/25.4 mm dia, Thorlabs, ACY254-250A) to shape the beam into a rectangular profile. This shaped beam is then modulated by patterns displayed on the spatial light modulator (SLM, Forth Dimension, QXGA-3DM). The SLM features a resolution of 2048×1536 ferroelectric liquid crystal pixels. When paired with a polarizing beam splitter cube (Edmund, #49002) and an achromatic half-wave plate (Bolder Vision Optik, BVO AHWP3), it enables a phase retardance of 0 or π in the diffracted beam, depending on the on/off status of the pixels. The diffracted light from the SLM is then focused through a 450 mm focal length lens (L2, 450 mm FL/40 mm dia, Edmund 49-282) onto a customized annular mask. This annular mask has a series of different size constraints that can filter out unwanted diffraction orders corresponding to different lattice patterns. After passing through the annular mask, the selected diffraction orders are magnified by a pair of relay lenses (L3, 70 mm FL/30 mm dia, Optosigma DLB 30-70PM, L4, 75 mm FL/30 mm dia, Optosigma DLB 30-75PM) and then conjugated to a scanning module that includes two galvanometers

(Cambridge Technology, 6210H) and a pair of relay lenses with the same focal length (L5, L6, 70 mm FL/25 mm dia, Optosigma DLB 25-70PM) configured in a 4f arrangement. Each galvanometer is conjugated to the back pupil plane of the illumination objective (Thorlabs, TL20X-MPL, 0.6 NA, 5.5 mm WD), allowing the system to scan along both the *x*-axis and *z*-axis of the sample. The image of the annular mask is further magnified by a factor of ~2.29 times using a relay lens (L7, 175 mm FL/25 mm dia, Edmund, 47-644; L8, 400 mm FL/25.4 mm dia, Thorlabs, AC254-400-A) and conjugated to the back focal plane of the illumination objective. The illumination objective then transforms the image to create the desired light sheet at its front focal plane, illuminating the sample. The emitted fluorescence is collected by the detection objective (Nikon, CFI Apo LWD 25XW, 1.1 NA, 2 mm WD) and imaged by a tube lens (L13, 400 mm FL / 50 mm dia, Thorlabs, AC508-400-A). The image is then magnified by another pair of lenses (L12, 120 mm FL/30 mm dia, Optosigma, DLB 30-120PM, L14, 170 mm FL/30 mm dia, Optosigma, DLB 30-170PM) with a total magnification of ~70× from the sample plane to the camera plane and recorded by an sCMOS camera (Hamamatsu, Orca Flash 4.0 v3 sCMOS) after passing through an emission filter.

In the structured excitation arm of the system, stripe patterns are sequentially displayed on the SLM, synchronized with the generation of the activation light sheet. Before projection onto the SLM, the laser beam is magnified 11.4× through a pair of relay lenses (L9, 17.5 mm FL/12.5 mm dia, Edmund 49-928; L10, 200 mm FL/30 mm dia, Optosigma, DLB-30-200PM). The reflected laser beam from the SLM passes through a 350 mm focal length lens (L11, 350 mm FL/30 mm dia, Optosigma, DLB-30-350PM) and focuses onto the mask to filter out unwanted diffraction orders. The laser beam then passes through the mask and a customized six/ten-section half-wave plate to alter its polarization state. After passing through the dichroic mirror (Chroma), the beam is magnified 3.33× through relay lenses (L12, L13) and directed into the detection objective to form SIM patterns on the sample. In this setup, the SLM is conjugated to the focal plane of the detection objective while the mask is conjugated to the back focal plane of the detection objective.

A wide-field imaging setup is also established to facilitate sample localization. When the flip mirror is positioned in the optical path, the collimated light beam is expanded by a pair of relay lenses (L15, 10 mm FL/8 mm dia, Thorlabs, AC080-010-A, L16, 100 mm FL/25 mm dia, Thorlabs, AC254-100-A), then demagnified by a lens (L17, 100 mm FL/25 mm dia, Optosigma, DLB-25-100PM) and the epifluorescence objective (Nikon, MRD07420, 40X/0.8 NA, 3.5 mm WD) to illuminate the sample. The fluorescence signal is collected by the epifluorescence objective, passes through lens L17, and is filtered before being imaged onto an sCMOS camera (Excelitas Technologies, pco.panda 4.2 bi sCMOS).

Supplementary Note 3 | Simulation of LA-NLSIM

We wrote MATLAB codes following the equations described below to visualize the nonlinear activation process of LA-NLSIM in Extended Data Fig. 5 and Supplementary Video 8.

In the first step of the activation process, we uniformly activate the fluorescent molecules as time progresses. The switching speed from the on-state to the off-state depends on the intensity of the excitation light. Assuming that after irradiation with laser intensity I_0 for time τ_0, the activated fluorescent molecules can be completely returned to the off-state, then the distribution of fluorescent molecules remaining in the on-state after an exposure time t is given by

$$S_{on}(x) = e^{-\frac{t}{\tau_0 I_0} I_{off}(x)} S(x) \tag{1}$$

where $S(x)$ is the sample distribution, $I_{off}(x)$ is the intensity distribution of the turn off light that can be assumed as a one-dimensional sinusoidally varying pattern

$$I_{off} = \frac{I_0}{2}(1 - \cos(2\pi k_0 x + \varphi)) \tag{2}$$

Here we define the term, saturation factor (SF), to be the ratio of the exposure time t to the off time τ_0. Eq. (1) then becomes

$$S_{on}(x) = e^{-\frac{SF}{2}(1-\cos(2\pi k_0 x+\varphi))} S(x) \tag{3}$$

In the last step of patterned read-out, we collect the remaining fluorescence from the on molecules by shifting the pattern by π phase. The intensity distribution of the readout illumination can be described as

$$I_{readout} = \frac{I_0}{2}(1 - \cos(2\pi k_0 x + \varphi + \pi)) \tag{4}$$

In this case, the fluorescent molecules that remain in the on-state will be

$$S'_{on}(x) = I_{readout} S_{on}(x) e^{-\frac{SF}{2}(1-\cos(2\pi k_0 x+\varphi+\pi))} \tag{5}$$

Supplementary Note 4 | Conceptual design of SRFormer

a) Architecture design of SRFormer

In the regime of volumetric biological data restoration, 3D residual channel attention network (3D RCAN)[7,8] is considered as one of the most powerful yet simple models. However, due to its relatively shallow architecture constituted based on traditional 3D convolutional layers and the channel attention mechanism, the expansibility and feature extraction capability of 3D RCAN are limited. Recently, transformer-based image super-resolution models for processing natural images such as the Swin-transformer image restoration model (SwinIR)[9] and dual aggregation transformer (DAT) model[10] have emerged, featuring a larger model scale and a better performance in various natural 2D image restoration tasks. However, whether the transformer-based image SR model outperforms conventional convolutional neural networks in volumetric data super-resolution reconstruction has not been explored.

To this end, our preliminary experiments started with comparing the performance of SwinIR[9], DAT[10] and 3D RCAN[7] following the network configuration of the original papers. Interestingly, we found that though with a shallower and small-scale, the 3D RCAN generated better high-frequency details of biological structures especially in the axial dimension than the more complex SwinIR and DAT (Extended Data Fig. 6a-c and Supplementary Fig. 7a-c). We speculated that the underlined reason is the backbone of SwinIR and DAT focuses on 2D feature extraction, preventing effective utilization of axial structural continuity of the volumetric LLSM data. Therefore, we then tried replacing the original 2D convolutional layers and 2D Swin-transformer blocks with 3D convolutional layer and the video Swin-transformer blocks[11], respectively, in the original DAT while not changing its depth and overall architecture, which is denoted as 3D DAT hereafter. As a result, a notable improvement in both axial resolution and reconstruction fidelity in terms of peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) is observed after such 3D modification (Supplementary Fig. 7). These results indicate that 3D feature extraction capability is of vital importance to volumetric SR models.

Moreover, our previous research[12] and other existing literature[13,14] have demonstrated that incorporating frequency feature manipulation and the pyramid network architecture generally benefits image SR capability of neural network models. Inspired by this, we further equipped the aforementioned 3D DAT with the U-shaped architecture for hierarchical feature extraction and spatial-frequency fusion block (SFFB) for non-local spectral information aggregation (Supplementary Fig. 6 and Supplementary Note 4b). Particularly, we explored integrating various depths of U-shaped structures, i.e., with different numbers of down-sampling and up-sampling modules, and found that the one-stage U-shaped architecture, i.e., incorporating only one down-sampling and up-sampling operations in each forward propagation, yielded the best SR reconstruction performance.

Most deep learning approaches for microscopy image denoising and super-resolution require training specific models for each type of biological specimens. This is because a general model trained with multiple biological specimens usually performs worse than the models trained on a specific type of specimens, due to limitations in neural network scalability. On the other hand, recent explorations of large language models suggest that a larger model with more trainable parameters is more robust in processing multi-tasks parallelly. Therefore, following the scaling law[15], we scaled up the network to 62,767,512 parameters by increasing feature channels and repeating feature extraction blocks in SRFormer. To our best knowledge, this is the largest model reported for biological image restoration. We trained the SRFormer model on two Nvidia A800 GPUs, and our experiments revealed that with all abovementioned advancements, a well-trained SRFormer enabled volumetric SR reconstruction of multiple biological specimens with the performance comparable to that achieved by individual 3D RCAN models trained on datasets of each specific biological structures (Extended Data Fig. 6).

b) Detailed network architecture of SRFormer

As shown in Supplementary Fig. 6, SRFormer comprises three modules: shallow feature extraction, deep feature extraction, and high-quality image reconstruction. Initially, the low-resolution (LR) input image stack $X \in \mathbb{R}^{Z\times H\times W}$, where Z, H and

W denote the size along three dimensions of the data.

First, the LR image stack is processed through two 3D convolutional layers to obtain the shallow feature $F_{s1} \in \mathbb{R}^{Z\times H\times W\times C_1}$ and $F_{s2} \in \mathbb{R}^{Z\times H\times W\times C_2}$. The dimensions of the feature, denoted as C_1 and C_2, are set to 90 and 180, respectively, in our experiments. Next, the shallow feature F_{s1} is fed into four dual aggregation transformer block (DATB) groups that are arranged following a U-shaped structure to generate hierarchical deep features. Specifically, the first group of feature channels $F_{d1} \in \mathbb{R}^{Z\times H\times W\times C_1}$ are obtained by sending F_{s1} to the first DATB group, then F_{d1} are sequentially passed through a down-sampling layer, the second DATB group, and an up-sampling layer (realized by pixel shuffle), generating a group of deep feature channels $F_{d2} \in \mathbb{R}^{Z\times H\times W\times C_1}$. The F_{d1} and F_{d2} are concatenated to constitute hierarchical deep feature $F_{hd} = concat(F_{d1}, F_{d2}) \in \mathbb{R}^{Z\times H\times W\times C_2}$, which is then passed through the last two DATB groups to obtain the refined feature $F_d \in \mathbb{R}^{Z\times H\times W\times C_2}$. Each DATB group consists of 6 consecutive 3D DATB, and one 3D spatial-frequency fusion block (SFFB). Each DATB is constructed by one 3D dual spatial transformer blocks (3D DSTB) and one 3D dual channel transformer blocks (3D DCTB).

Finally, the combined feature of the shallow feature F_{s2} and the refined feature F_d is fed into an up-sampling block and a 3D convolutional layer to generate the final super-resolution image $\hat{Y} \in \mathbb{R}^{Z'\times H'\times W'}$, where the parameters Z', H' and W' are set to $3\times Z$, $2\times H$ and $2\times W$, respectively.

Supplementary Note 5 | Characterization of SRFormer LA-SIM

a) Calculation of assessment metrics

To quantitatively evaluate the images reconstructed by SRFormer LA-SIM and other methods, we used the PSNR and SSIM between the SR image $\hat{Y}$ and the ground truth (GT) image Y. Since the signal intensity of the SR and GT images differs in dynamic range, we applied a linear transformation[12] to the SR image as follows

$$\tilde{Y} = \alpha\hat{Y} + \beta \tag{6}$$

where α and β represent the transformation coefficients aimed at minimizing the squared error between the transformed image and the normalized GT image, $\tilde{Y}$ denotes the linear transformed SR image. This problem can be formulated as a linear regression problem:

$$\min_{\alpha,\beta} \left|\left|\alpha\hat{Y} + \beta - Y\right|\right|_2^2 \tag{7}$$

The closed-form solution to this problem is given by:

$$\hat{\alpha} = \frac{\sum_{i=1}^{N} Y_i \times \left(\tilde{Y}_i - \mathrm{mean}(\tilde{Y})\right)}{\sum_{i=1}^{N} \tilde{Y}_i^2 - N \times \mathrm{mean}(\tilde{Y})^2} \tag{8}$$

$$\hat{\beta} = N \times \sum_{i=1}^{N} \left(Y_{\mathrm{i}} - \hat{\alpha} \times \tilde{Y}_i\right) \tag{9}$$

where $\hat{\alpha}$ and $\hat{\beta}$ denote the optimal values of the transformation coefficients α and β, respectively.

The final PSNR and SSIM are calculated as follows:

$$\mathrm{PSNR}(\tilde{Y}, Y) = 10 \times \log_{10}\left(\frac{\mathrm{N}}{\sum_{i=1}^{N}(Y_i - \tilde{Y}_i)^2}\right) \tag{10}$$

$$\mathrm{SSIM}(\tilde{Y}, Y) = \frac{(2\mu_{\tilde{Y}}\mu_Y + \mathrm{c}_1)(2\sigma_{\tilde{Y}Y} + \mathrm{c}_2)}{(\mu_{\tilde{Y}}^2 + \mu_Y^2 + \mathrm{c}_1)(\sigma_{\tilde{Y}}^2 + \sigma_Y^2 + \mathrm{c}_2)} \tag{11}$$

where$\mu_{\tilde{Y}}$, μ_Y, $\sigma_{\tilde{Y}}$ and σ_Y denote the mean values and standard deviations of the SR image $\tilde{Y}$ and the GT image Y, respectively. $\sigma_{\tilde{Y}Y}$ denotes the cross-covariance between the SR image $\tilde{Y}$ and GT image Y. The constants c_1 and c_2 used in this paper are 0.01^2 and 0.03^2, respectively.

b) Ablation study of SRFormer

As is discussed in the Methods section of the main text and Supplementary Note 4, there are three key advancements in SRFormer over existing transformer-based image SR neural network models such as SwinIR[9] and DAT[10]. First, instead of using 2D convolutional layers and 2D shifted window (Swin) transformer blocks[16], we used 3D convolutional layers and Swin transformer blocks designed for 3D data[11] in SRFormer for volumetric feature extraction. Second, we designed a U-shaped architecture for the cascading DATB groups to enable hierarchical feature manipulation within the network. Third, by incorporating Fourier space learning[12,13], we endowed SRFormer with an extended receptive field across the whole image, further strengthening the representation capability of the model.

To demonstrate the effectiveness of above modifications, we conducted ablation experiments for SRFormer using LLSM and rDL LA-SIM image pairs of three different biological specimens including outer mitochondrial membrane (Mito), microtubules (MTs), and F-actin. In detail, we trained several image SR models including the original DAT models with 2D feature extraction modules (i.e., 2D convolution and 2D Swin transformer blocks), denoted as 2D DAT, modified DAT models with 3D feature extraction modules (i.e., 3D convolution and 3D Swin transformer blocks), denoted as 3D DAT, modified SRFormer model without SFFB blocks (denoted as SRFormer w/o SFFB), modified SRFormer model without U-shaped architecture (denoted as SRFormer w/o U-shape), in which four DATB groups are connected sequentially, and the proposed SRFormer model with all reinforcements mentioned above.

Typical results of these models and quantitative comparisons in terms of PSNR and SSIM are shown in Supplementary Fig. 7, from which we draw such conclusions: (1) By comparing 2D DAT to other four models with 3D adaptation, we found that the 3D feature extraction modules substantially enhance axial resolution in the inferred volumetric SR results. (2) By comparing 3D DAT with SRFormer w/o SFFB and SRFormer w/o U-shape, we found that the incorporation of either SFFB or one-stage U-shaped architecture is able to slightly improve the SR resolvability both qualitatively and quantitatively. However, these two models struggle with reconstructing the hollow reconstruction of the outer mitochondrial membrane (Supplementary Fig. 7a) and are

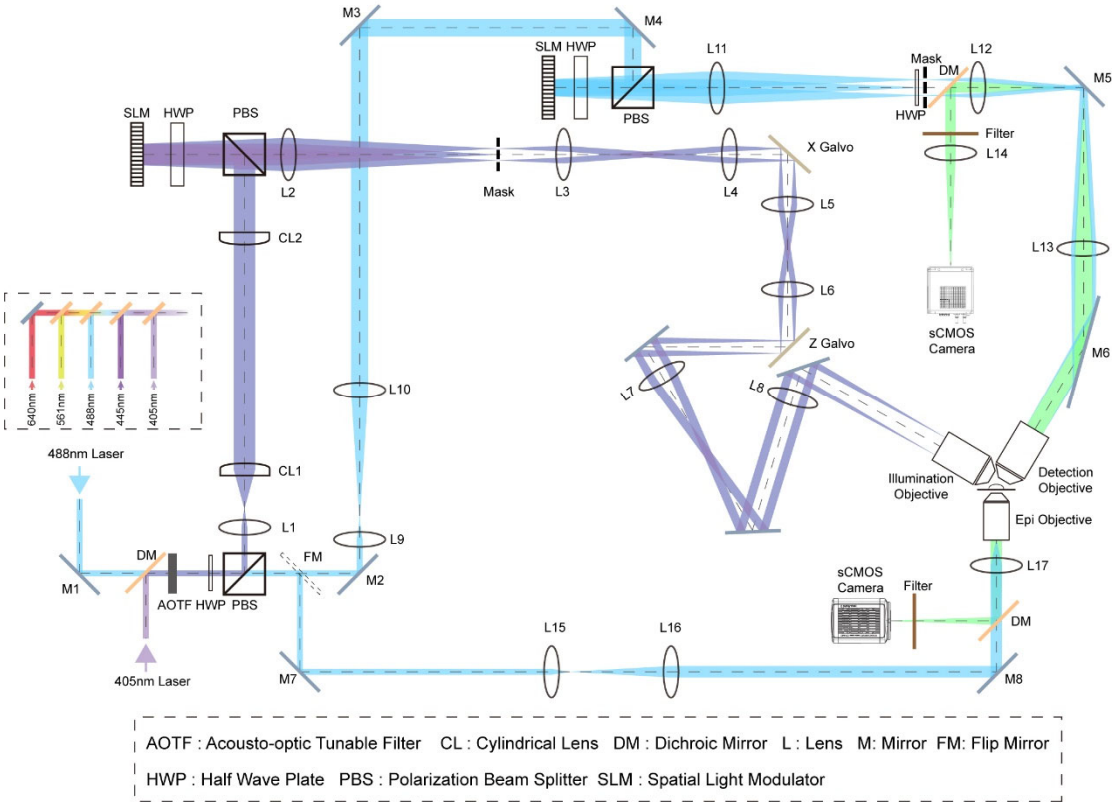
prone to generating unnatural structure in MTs and F-actin images (Supplementary Fig. 7b, c). (3) Benefiting from volumetric and hierarchical feature extraction capability provided by 3D adaptation, SFFB, and U-shaped architecture, SRFormer can distinguish the hollow structure of outer mitochondrial membrane in both lateral and axial (Supplementary Fig. 7a), and reconstruct tubular structure with similar morphology and comparable resolution with rDL LA-SIM, i.e., the GT used in training. Meanwhile, SRFormer achieves the highest reconstruction PSNR and SSIM compared to other models.

c) Comparison of SRFormer with 3D RCAN and SwinIR

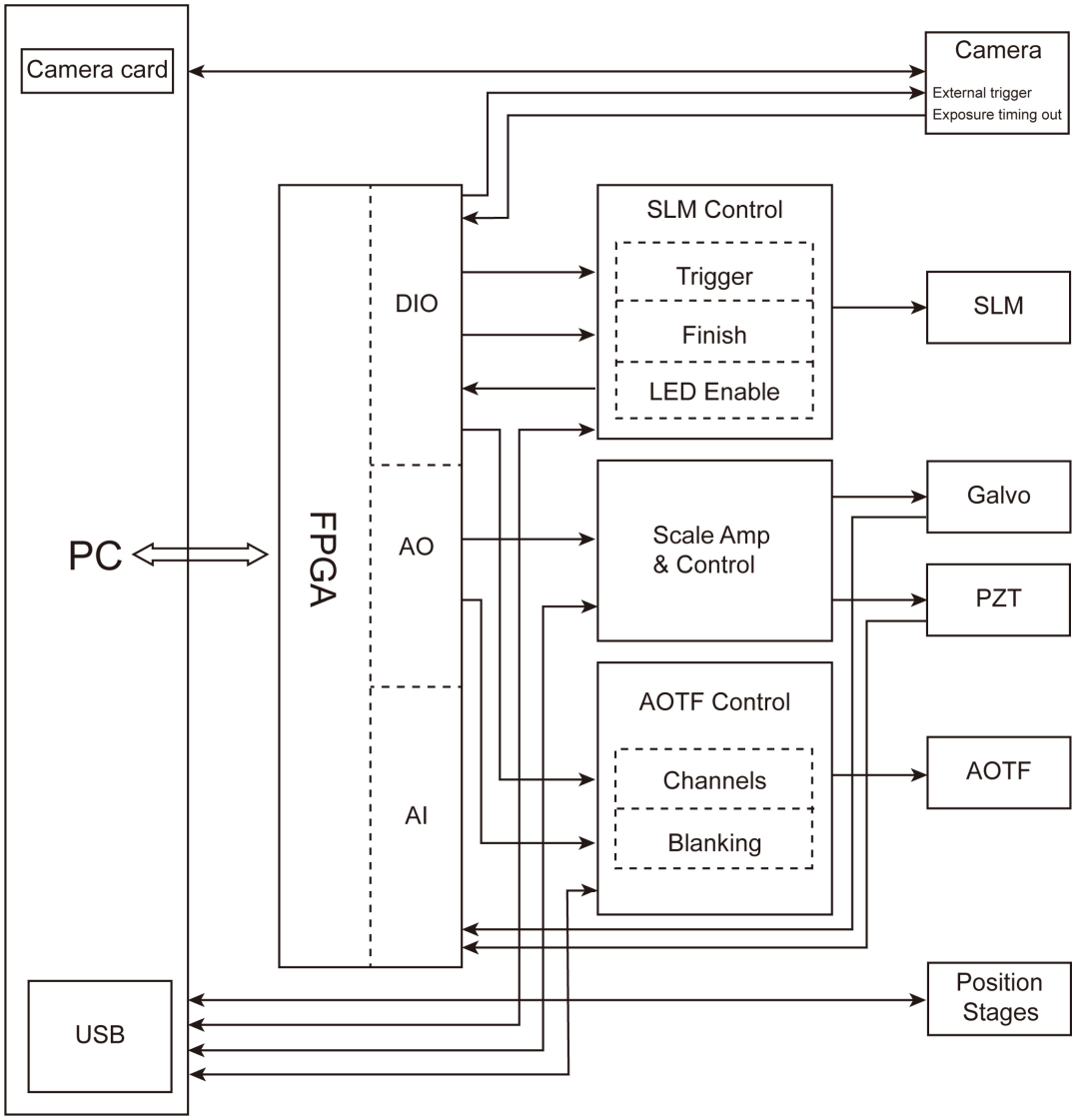
To further validate the superior performance of SRFormer in volumetric image super-resolution tasks, we compared it with two state-of-the-art deep learning algorithms: three-dimensional residual channel attention networks (3D RCAN)[7], a commonly adopted neural network model for 3D biological data denoising and super-resolution, and SwinIR[9], a well-recognized neural network for image restoration which is constructed based on the Swin Transformer. We first independently trained three 3D RCAN models using three datasets of Mito, MTs and F-actin, respectively, referred to as 3D RCAN specific models. Next, we trained 3D RCAN, SwinIR (following its original configuration[9] with 2D feature extraction modules), and SRFormer models using a mixed dataset of all three biological structures, referred to as general models. As such, the "3D RCAN Specific" models were trained for processing data of a single type of specimen, and "3D RCAN General", "SwinIR General", and "SRFormer General" models each was trained to process data of various types of biological structures. Extended Data Fig. 6 presents representative SR images reconstructed using these models, which indicate several conclusions, including: (1) Although built based on conventional convolutional architecture, the 3D RCAN general model outperformed the transformer-based SwinIR model due to its volumetric feature extraction and aggregation capability; (2) the 3D RCAN specific model trained for a certain type of specimen achieved relatively better accuracy compared to the 3D RCAN general model trained with all relative datasets. This result is consistent with findings in other literature[12] that independent image SR neural network models should be trained for

each biological specimen to achieve optimal SR performance; (3) Consistent with the scaling law[15], with over 60 million trainable parameters and sufficient training data, SRFormer trained with datasets of all types of specimens achieved better performance than the 3D RCAN general model and even the three 3D RCAN specific models. It successfully reconstructed the hollow outer membrane of mitochondria in three dimensions (Supplementary Fig. 7a), as well as tubular structures of MTs and F-actin (Supplementary Fig. 7b, c). Additionally, the quantitative comparison between SRFormer and other models, shown in Supplementary Fig. 7d-f, also demonstrates the superior reconstruction fidelity and robustness to variations in signal-to-noise ratio and structural differences among various cell types.

Supplementary Figures

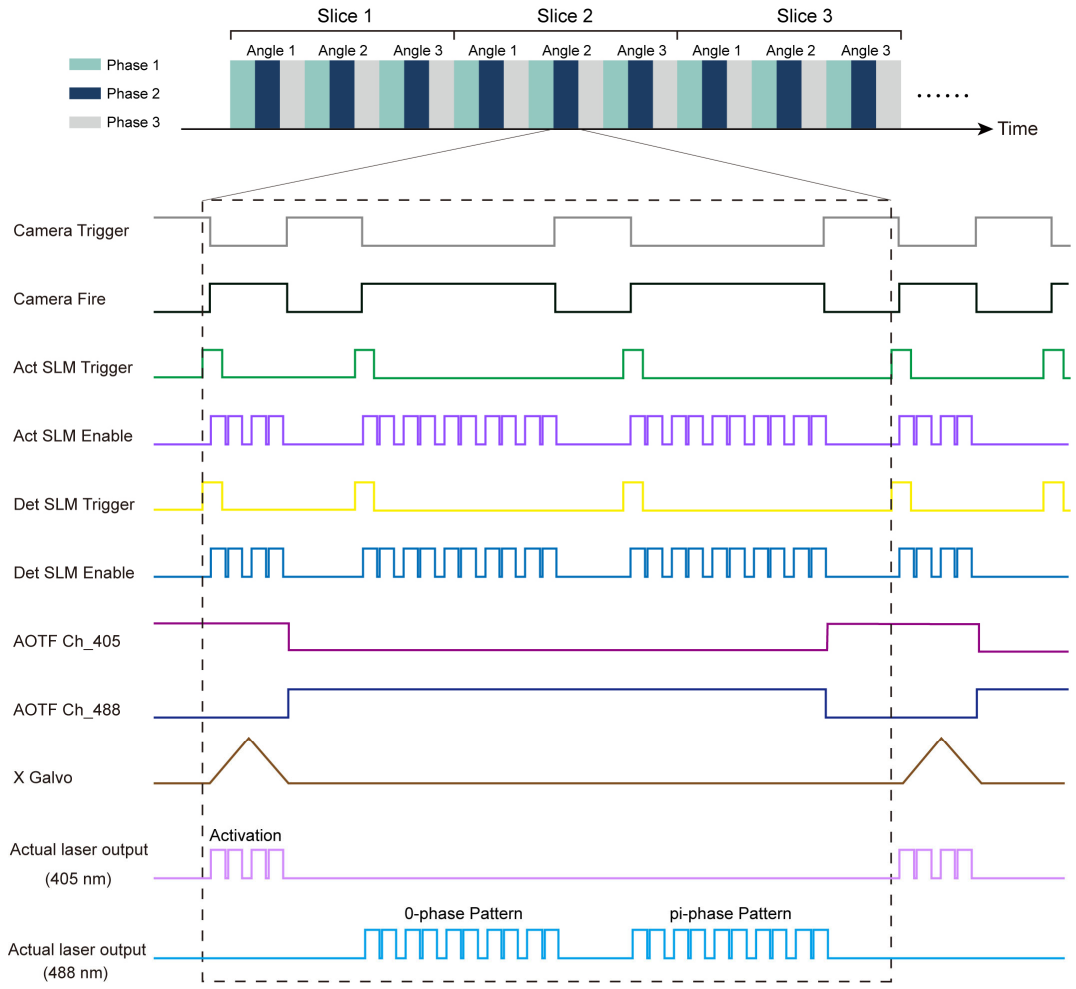


Supplementary Fig. 1 | Schematic of LA-SIM system. See Methods and Supplementary Note 2 for more information.

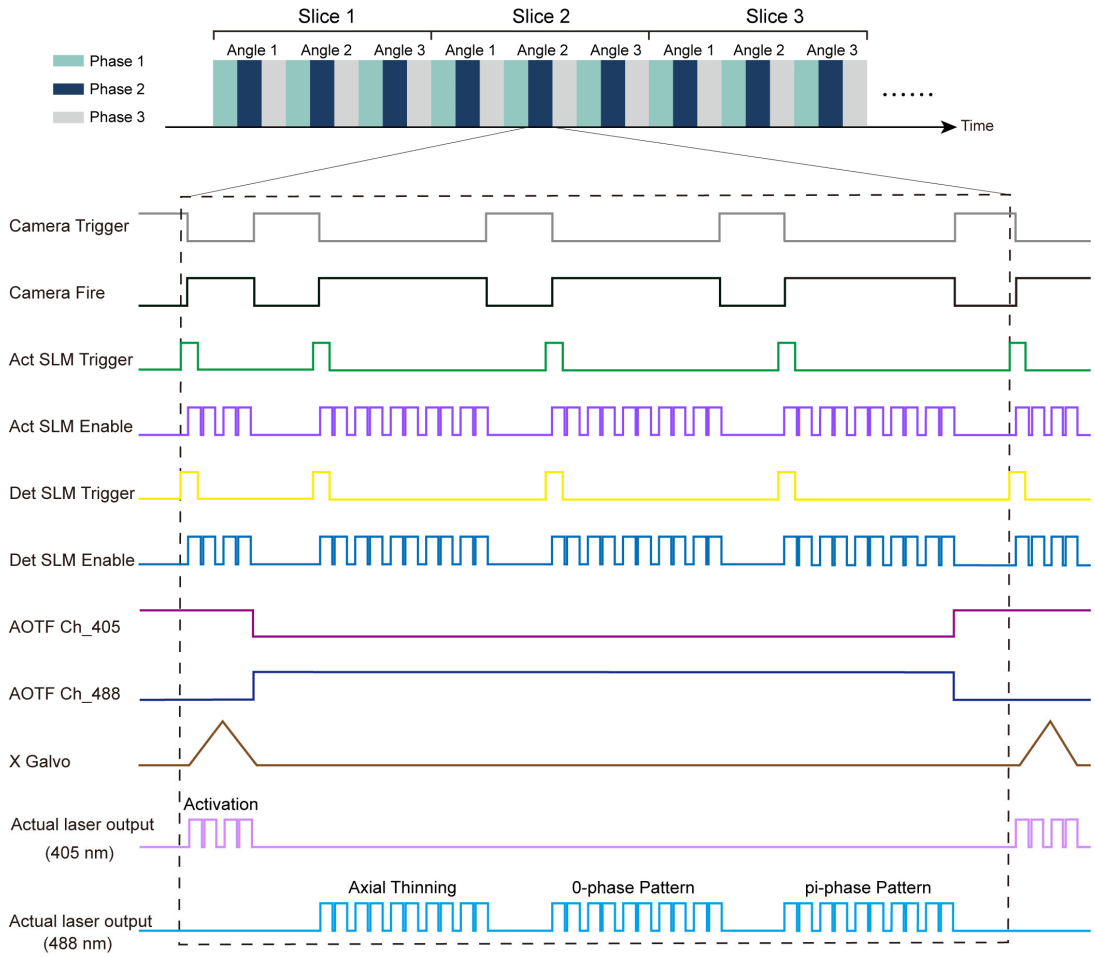


Supplementary Fig. 2 | Hardware control schematic of LA-SIM system.

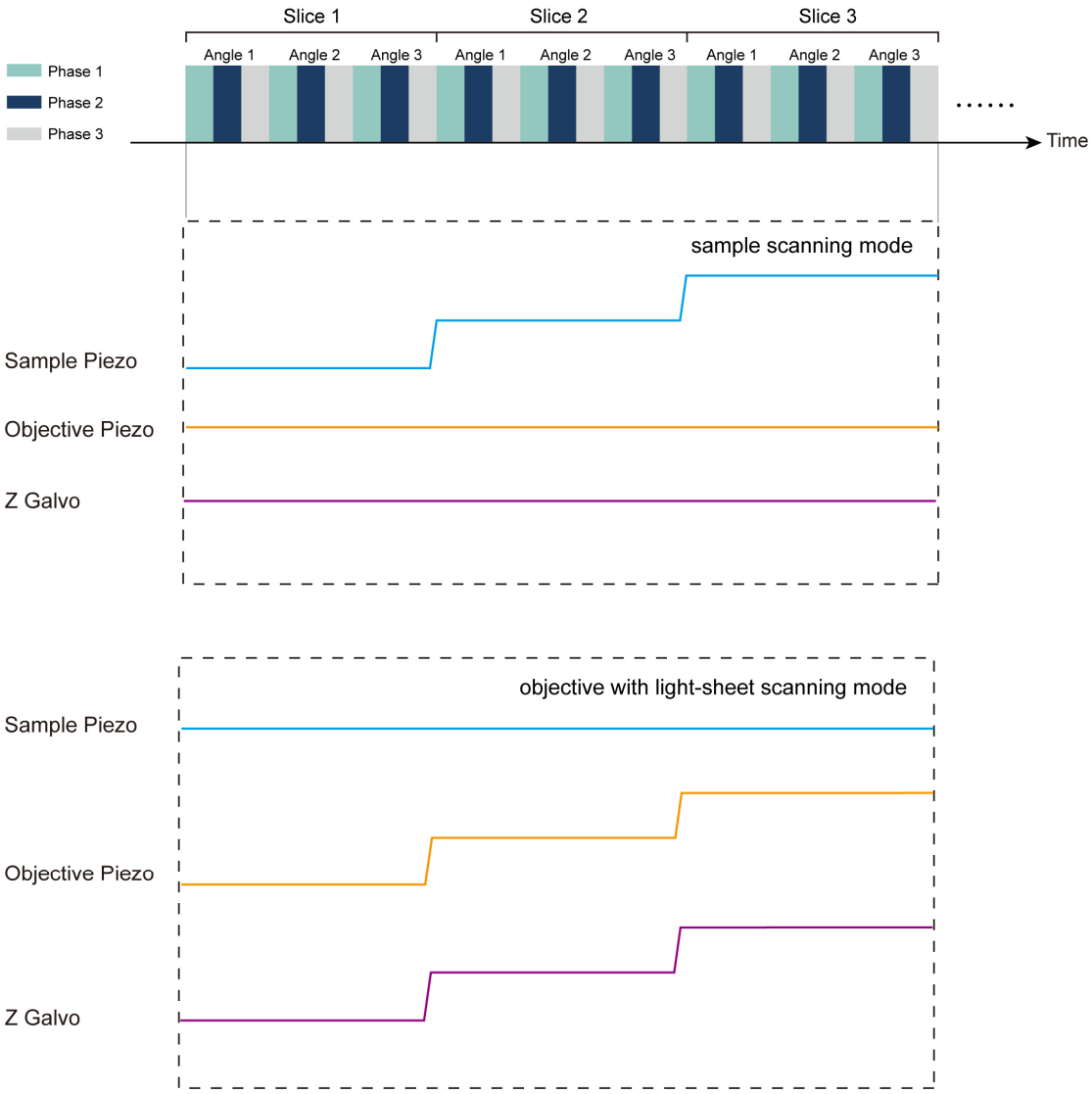
FPGA: field-programmable gate array; DIO: digital input and output; AO: analog output; AI: analog input; SLM: spatial light modulator; Galvo: galvanometer; PZT: piezoelectric transducer; AOTF: acousto-optic tunable filter. FPGA provides the analog and digital outputs to control the essential electronics for image acquisition. The AO ports connect and control devices requiring analog voltage modulation, including Galvo, PZT and AOTF (for controlling power output). The DO ports connect and control devices requiring highly synchronized operation, including Camera, SLM and AOTF (for controlling switch state). The AI and DI ports receive feedback signals from Galvo and PZT. The camera acquisition card in the computer receives data from the camera. Different acquisition timing is designed based on imaging mode requirements. Refer to the Methods section and Supplementary Figs. 3-5 for additional details.



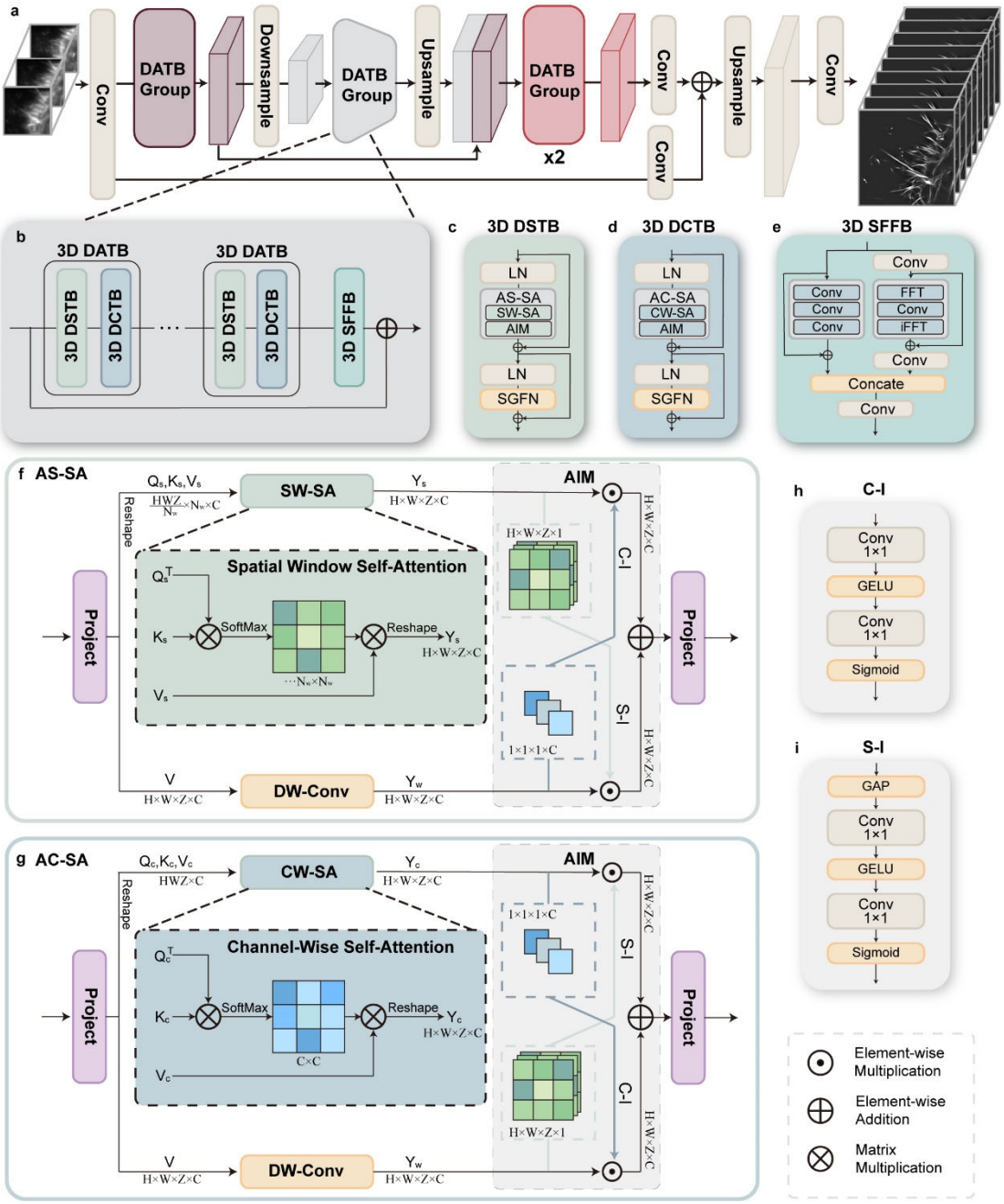
Supplementary Fig. 3 | Timing diagrams for hardware control, LA-LSIM without axial thinning acquisition for one orientation one phase.



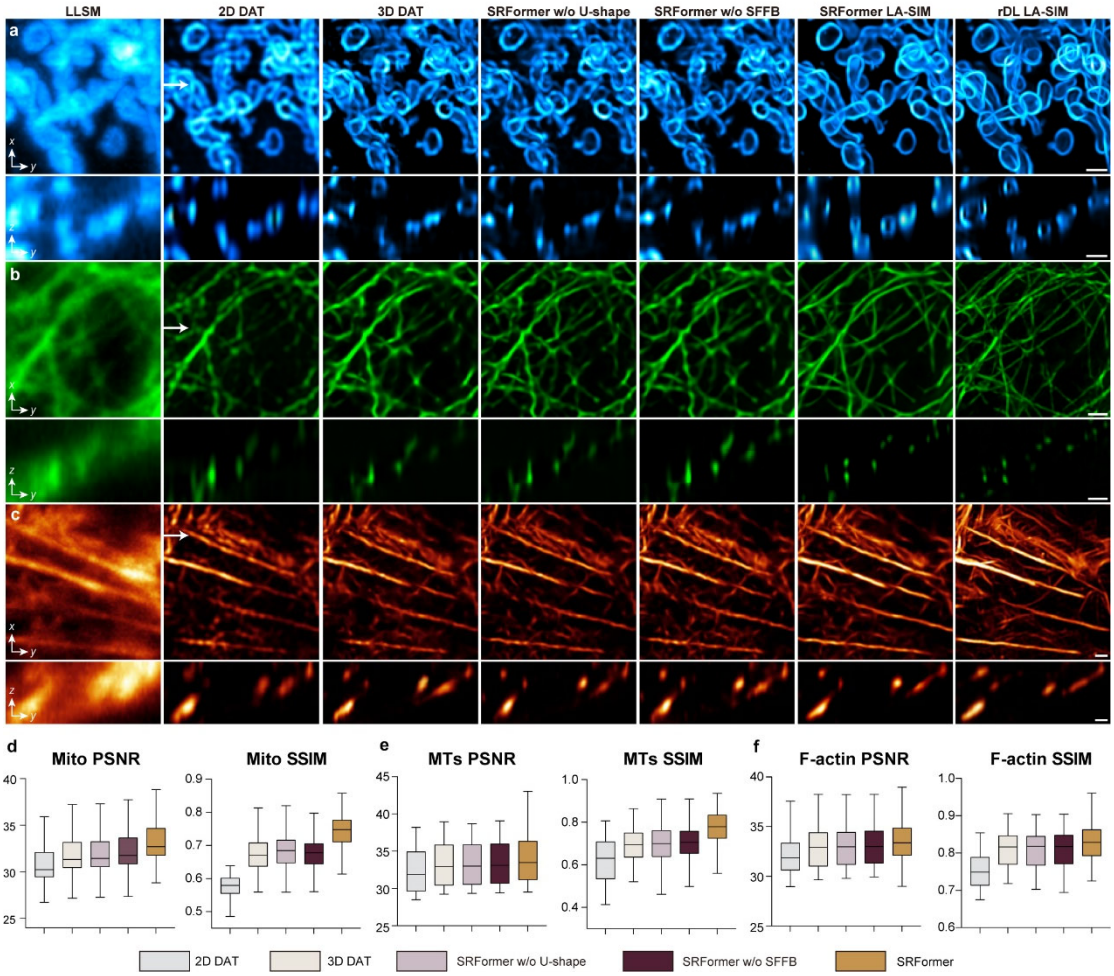
Supplementary Fig. 4 | Timing diagrams for hardware control, LA-LSIM with axial thinning acquisition for one orientation one phase.



Supplementary Fig. 5 | Timing diagrams for hardware control, volume acquisition for two scanning modes. Sample scanning mode is the volumetric acquisition mode in which the specimen is translated with a high-precision piezo stage through the stationary light sheet. Objective with light-sheet scanning mode is achieved by moving the light sheet and detection objective together through the specimen.



Supplementary Fig. 6 | Network architecture of SRFormer. a, The schematic of the inference phase of SRFormer. **b,** The architecture of dual aggregation transformer block (DATB) group. **c,** The architecture of 3D dual spatial transformer block (3D DSTB). **d,** The architecture of 3D dual channel transformer block (3D DCTB). **e,** The architecture of spatial-frequency fusion block (SFFB). **f,** The architecture of adaptive spatial self-attention (AS-SA). **g,** The architecture of adaptive channel self-attention (AC-SA). **h,** The architecture of channel-interaction (C-I). **i,** The architecture of spatial-interaction (S-I).



Supplementary Fig. 7 | Ablation study of SRFormer. a-c, Representative maximum intensity projections (MIP, *xy*-plane) and *yz*-slices of LLSM image stacks (first column) and SR images of Mito (a), MTs (b), and F-actin (c) reconstructed by 2D DAT (second column), 3D DAT (third column), SRFormer w/o U-shape (fourth column), SRFormer w/o SFFB (fifth column), and SRFormer LA-SIM (sixth column). Super-resolution rDL LA-SIM MIP images are provided for reference in the seventh column. Arrows indicate the *x* position for the *yz*-slices shown below. Scale bar, 2 μm. **d,** Statistical comparison of PSNR and SSIM values for the output SR images produced by 2D DAT, 3D DAT, SRFormer w/o U-shape, SRFormer w/o SFFB, and SRFormer on test datasets of Mito (d), MTs (e), and F-actin (f) (n=1000).

Supplementary Tables

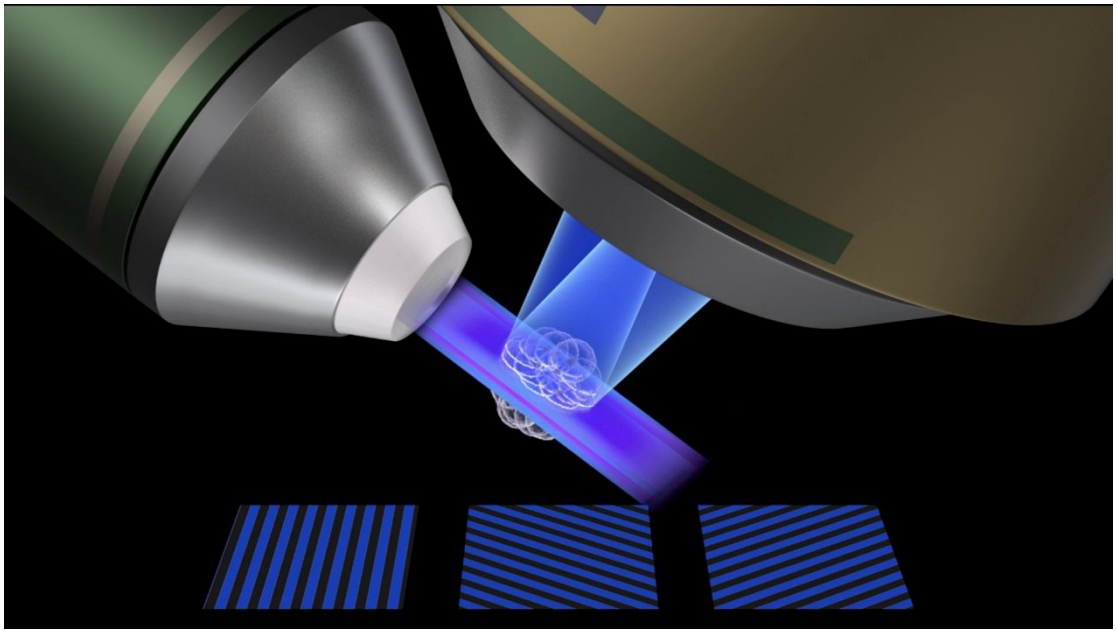
Supplementary Table 1 | Imaging parameters of LA-SIM

Data	Imaging mode	Sample (situation)	Label	Volume size of raw data (Width×Height ×Z-slice)	Exposure time for one phase one orientation (ms)				NA		Time points
					activation	axial thinning	0-phase	pi-phase	Activation	Excitation	
Fig. 1e-g, k, l Supplementary Video 3	LA-LSIM-*z*	COS-7 (fixed)	Skylan-NS-Ensconsin	640×640×445	5	25	15	15	0.35 0.14	1.0	/
Fig. 2a-e, g Supplementary Video 5	LA-LSIM	Hela (live)	Skylan-NS-Lifeact	512×512×91	1	/	5	5	0.35 0.14	1.0	150
Fig. 2i-l Supplementary Video 7	LA-LSIM-*z*	COS-7 (live)	Skylan-NS-Tomm20	448×672×401	2	5	5	5	0.35 0.14	1.0	45
Fig. 3a-c Supplementary Video 9	LA-NLSIM- *z*	COS-7 (fixed)	Skylan-NS-Ensconsin	800×800×491	5	25	20	10	0.35 0.14	1.0	/
Extended Data Fig. 3a Supplementary Video 4	LA-LSIM-*z*	COS-7 (fixed)	Skylan-NS-Tomm20	512×512×242	7	20	15	15	0.35 0.14	1.0	/
Extended Data Fig. 3b Supplementary Video 6	LA-LSIM	COS-7 (live)	Skylan-NS-Tomm20	512×512×65	2	/	10	10	0.35 0.14	1.0	/
Extended Data Fig. 4b-e	LA-LSIM-*z*	Hela (fixed)	Skylan-NS-Lifeact	512×512×221	3	15	15	15	0.35 0.14	1.0	/
	3D-SIM	Hela (fixed)	Skylan-NS-Lifeact	512×512×101	5	/	20 (excitation)		/	1.0	/

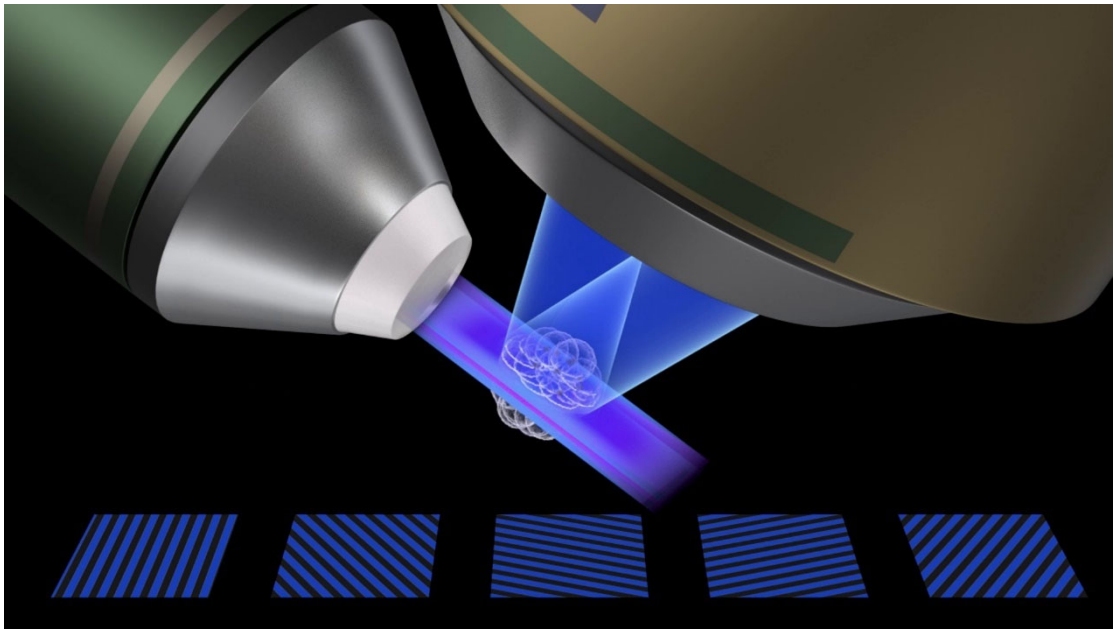
Supplementary Table 2 | Imaging parameters of SRFormer LA-SIM

Data	Imaging method (Acquisition mode)	Sample	Label	Excitation NA	Excitation λ (nm)	Exposure time per raw image (ms)	Volume size of raw data (Width×Height ×Z-slice×Channel)	Cycle time (Acquisition + resting time) (s)	Time Points (Video)
Fig. 4c-f Supplementary Video 10	LLSM (sheet-scan mode)	COS-7	Ensconsin-mStayGold SKL-mCherry LAMP1-Halo	0.35,0.14	488 560 642	10 10 10	320×832×191×3	6.42	690
Fig. 5a-c Supplementary Video 12	LLSM (sheet-scan mode)	COS-7	G3BP1-mStayGold LAMP1-Halo	0.35,0.14	488 560	10 10	352×768×181×2	4.11	500
Fig. 5d Supplementary Video 14	LLSM (sheet-scan mode)	COS-7	G3BP1-mStayGold LAMP1-Halo	0.35,0.14	488 560	10 10	512×512×101×2	2.44	55
Fig. 5e Extended Data Fig. 8 Supplementary Video 15	LLSM (slit-scan mode)	Mouse embryo	LAMP1-mStayGold	0.07	488	10	1024×1024×401×1	30	300
Extended Data Fig. 7 Supplementary Video 11	LLSM (sheet-scan mode)	COS-7	Ensconsin-3×mStaygold Tomm20-mCherry	0.35, 0.14	488 560	10 10	288×768×151×2	3.33	400
Extended Data Fig. 9	LLSM (slit-scan mode)	Mouse embryo	LAMP1-mStayGold	0.07	488	10	1024×1024×301×1	30	/
	3D-SIM	Mouse embryo	LAMP1-mStayGold	1.49	488	30	512×512×37×1	32.69	/

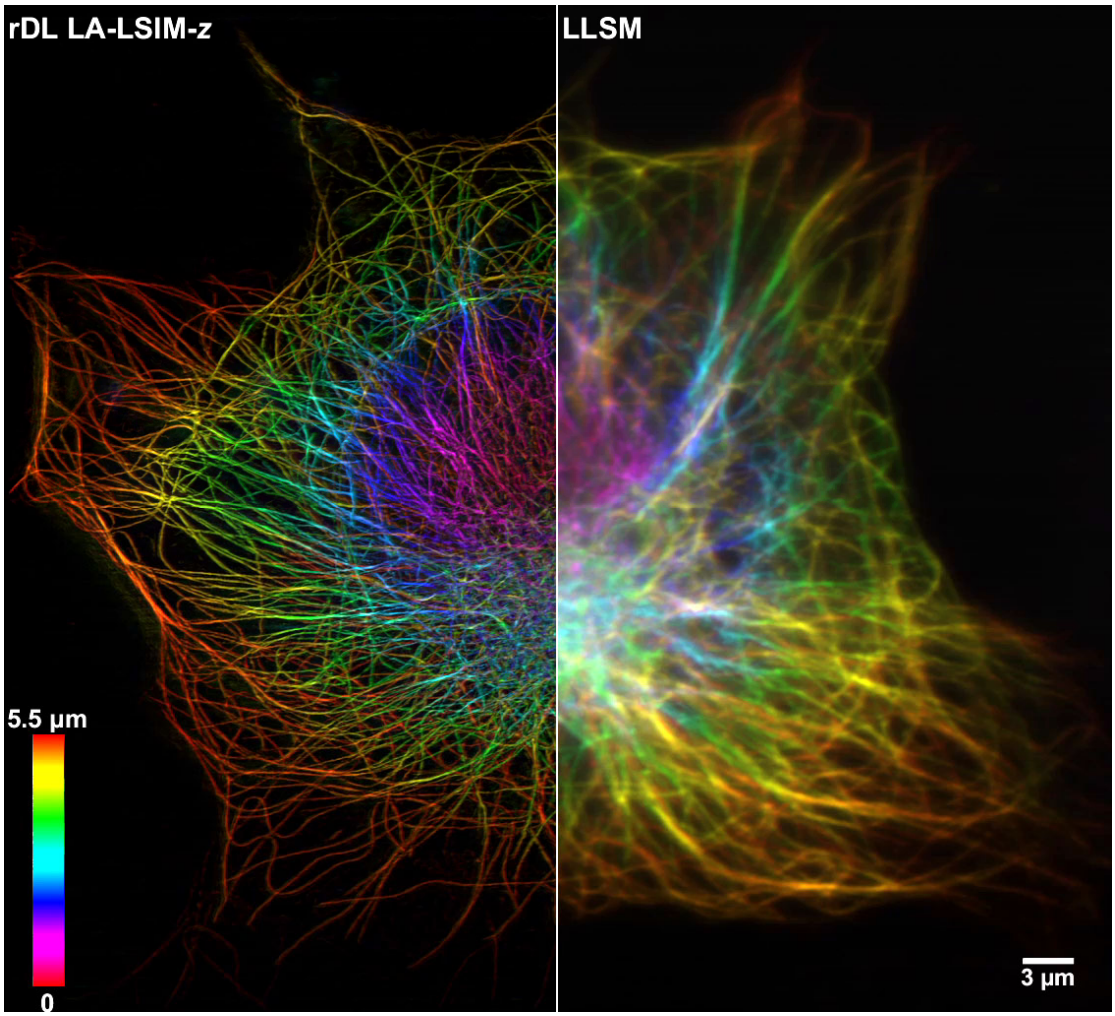
Captions for Supplementary Videos



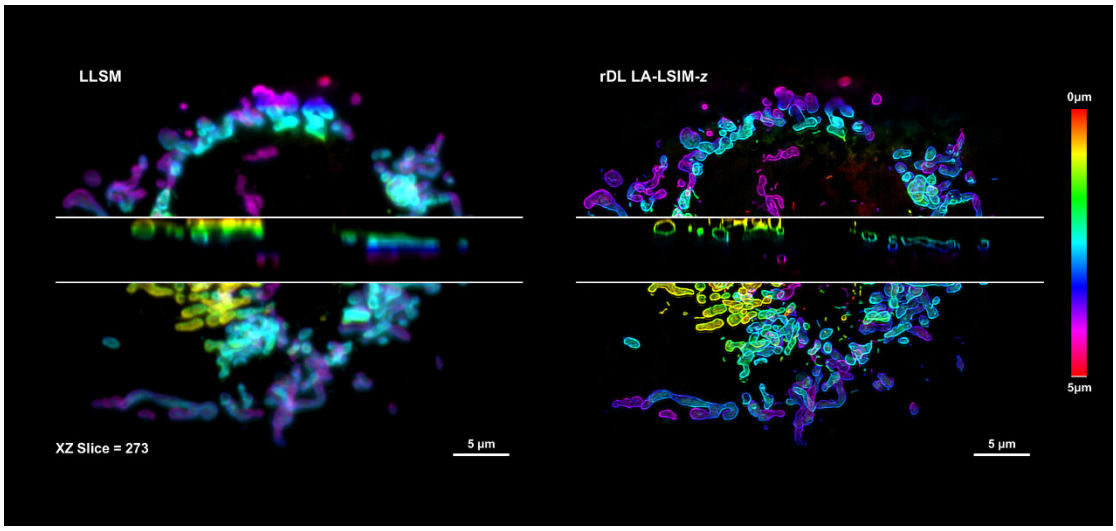
Supplementary Video 1 | Animation of LA-LSIM illumination and acquisition steps without (part I) and with (part II) sandwiched axial thinning.



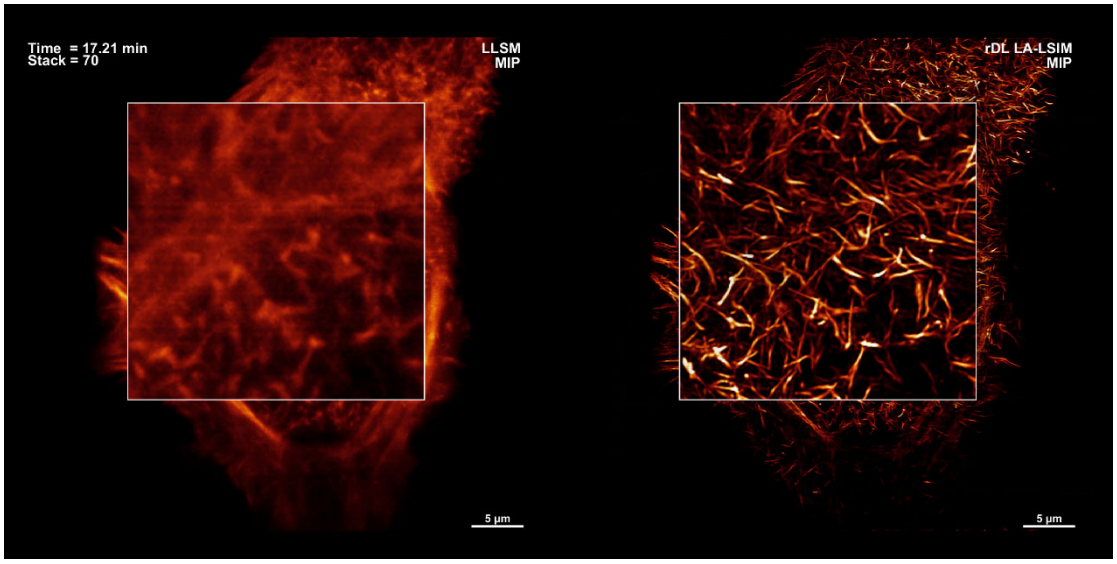
Supplementary Video 2 | Animation of LA-NLSIM illumination and acquisition steps without (part I) and with (part II) sandwiched axial thinning.



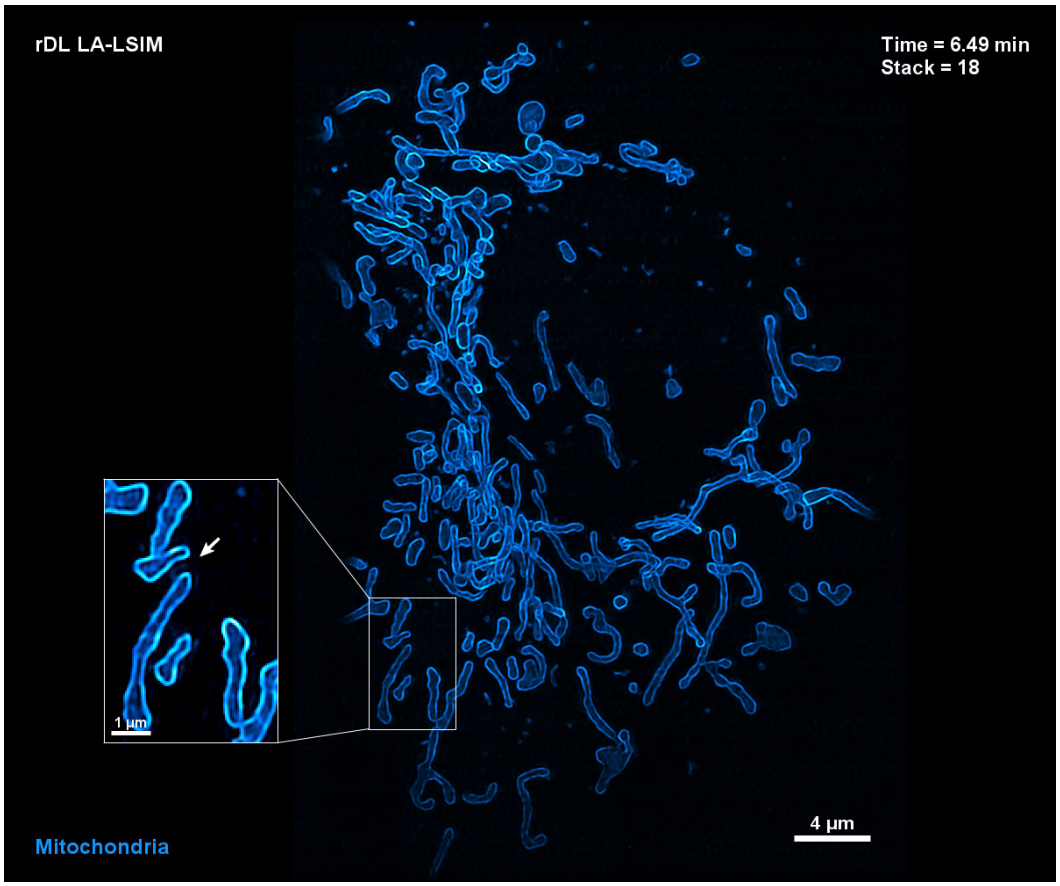
Supplementary Video 3 | Volume rendering of LA-LSIM-*z* image acquired from fixed COS-7 cell expressing Ensconsin-Skylan-NS, showing the progressive resolution enhancement from LLSM, LLSM with axial thinning, to LA-LSIM-*z* and LA-LSIM-*z* with rDL denoising. See also **Fig. 1**.



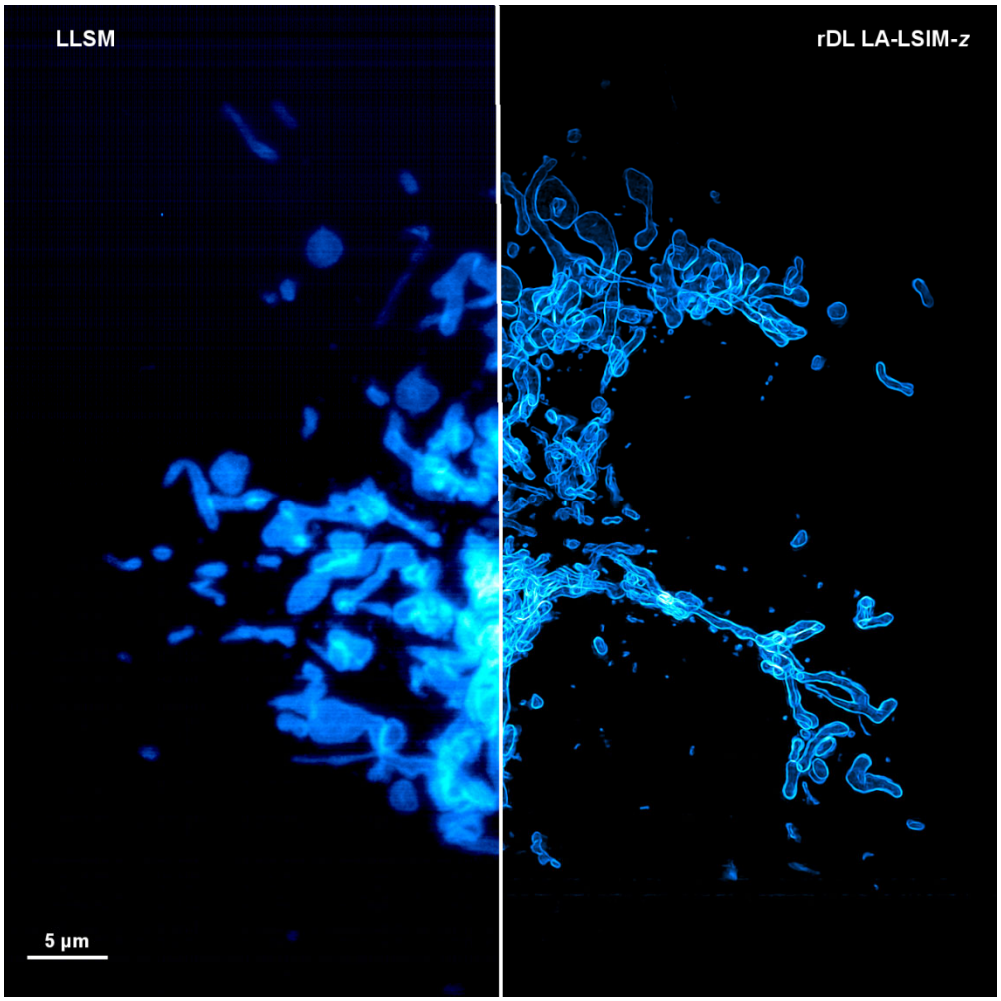
Supplementary Video 4 | Volume rendering and 3D projection of LLSM (left) and rDL LA-LSIM-*z* images from fixed COS-7 cell expressing Tomm20-Skylan-NS. The *x*-*z* scrolling views present resolution improvements in both lateral and axial dimensions via rDL LA-LSIM-*z*. See also **Extended Data Fig. 2.**



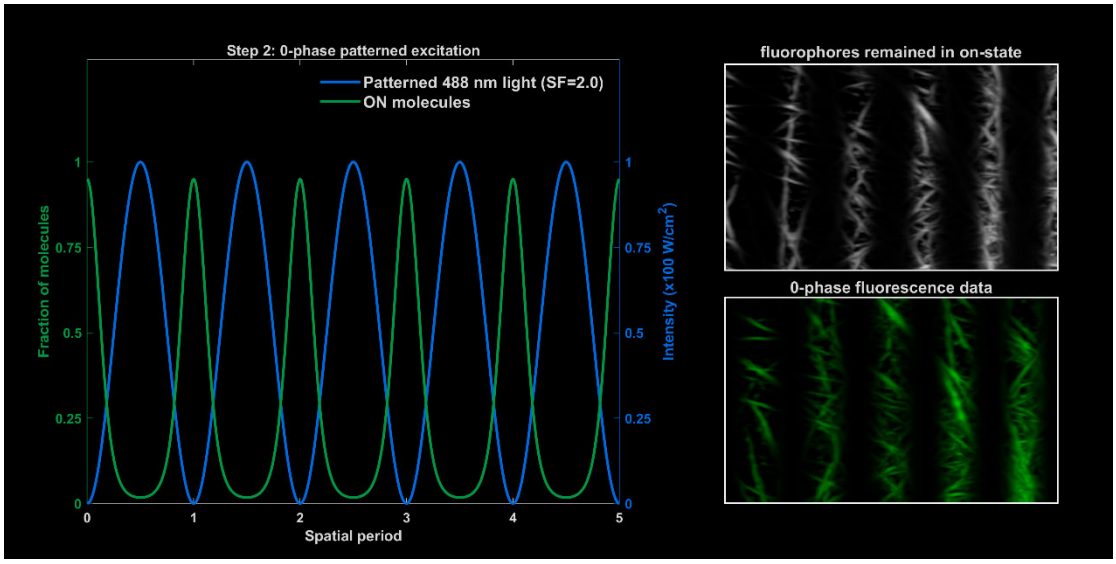
Supplementary Video 5 | 3D projections of LLSM (left) and rDL LA-LSIM (right) imaging of a live Hela cell expressing Lifeact-Skylan-NS, showing the F-actin cytoskeleton dynamics over the whole cell volume for 150 time points lasting ~37 mins. See also **Fig 2.**



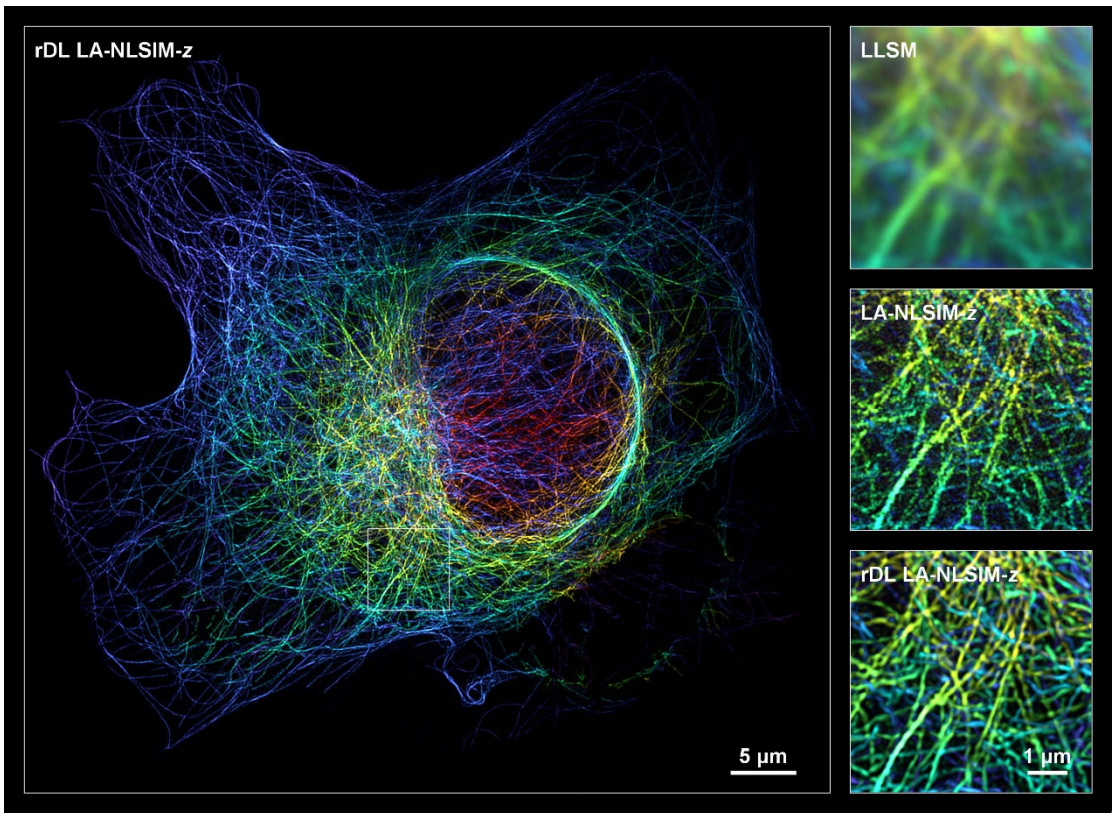
Supplementary Video 6 | 3D projections of rDL LA-LSIM imaging of live COS-7 cell expressing Tomm20-Skylan-NS, showing the mitochondrial fission and fusion membrane dynamics over the whole cell volume for 50 time points lasting ~18 mins. See also **Extended Data Fig. 2**.



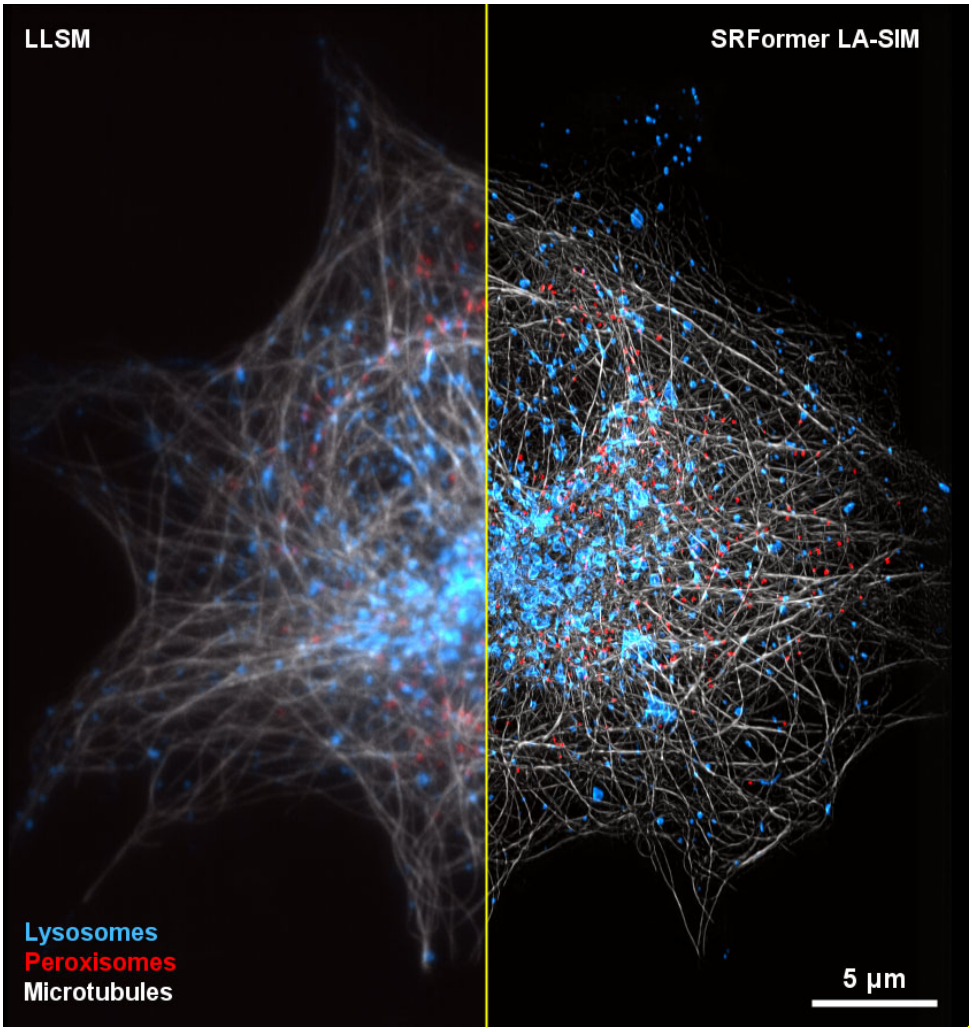
Supplementary Video 7 | 3D projections and surface rendering of rDL LA-LSIM-*z* imaging of live COS-7 cell expressing Tomm20-Skylan-NS for 45 time points lasting ~84 mins. See also **Fig.2**.



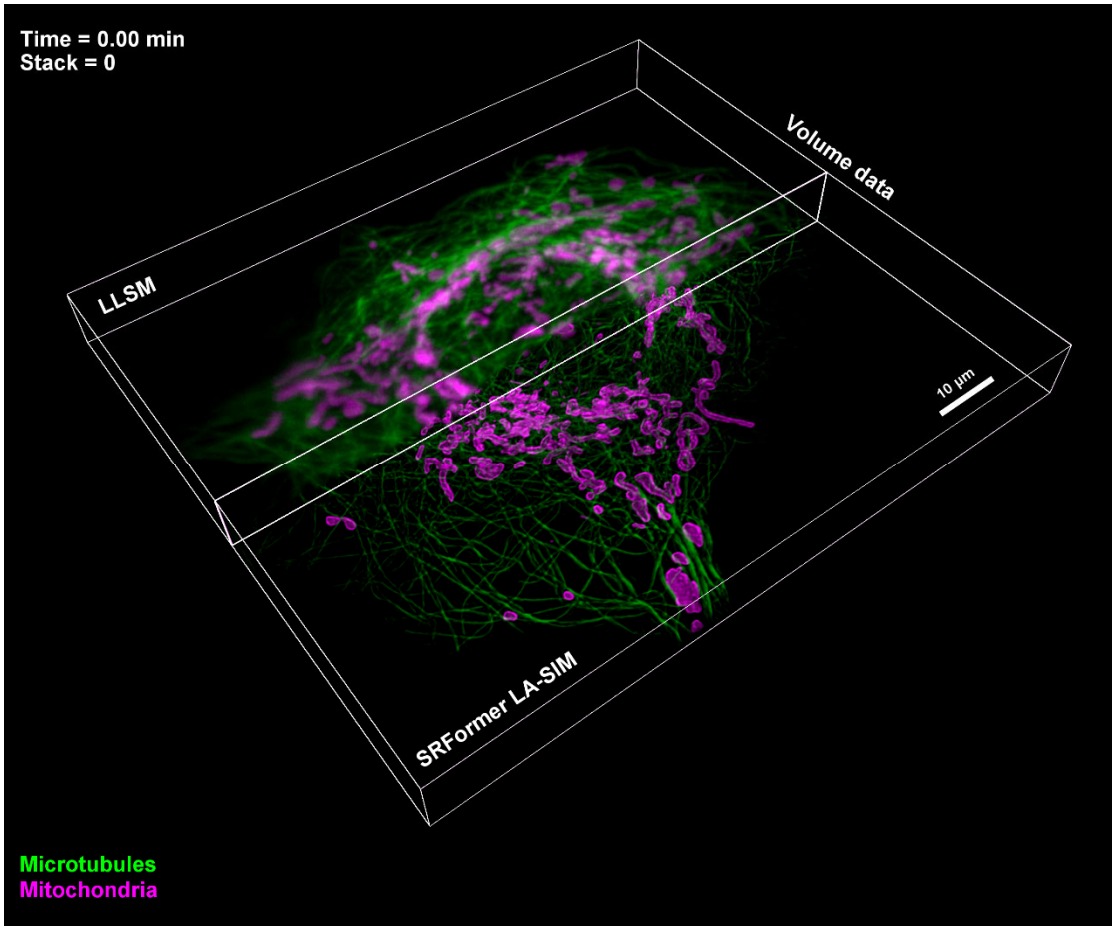
Supplementary Video 8 | Illustration of the sequential steps of LLS activation, 0-phase patterned excitation, and pi-phase patterned excitation in the illumination procedure of LA-NLSIM.



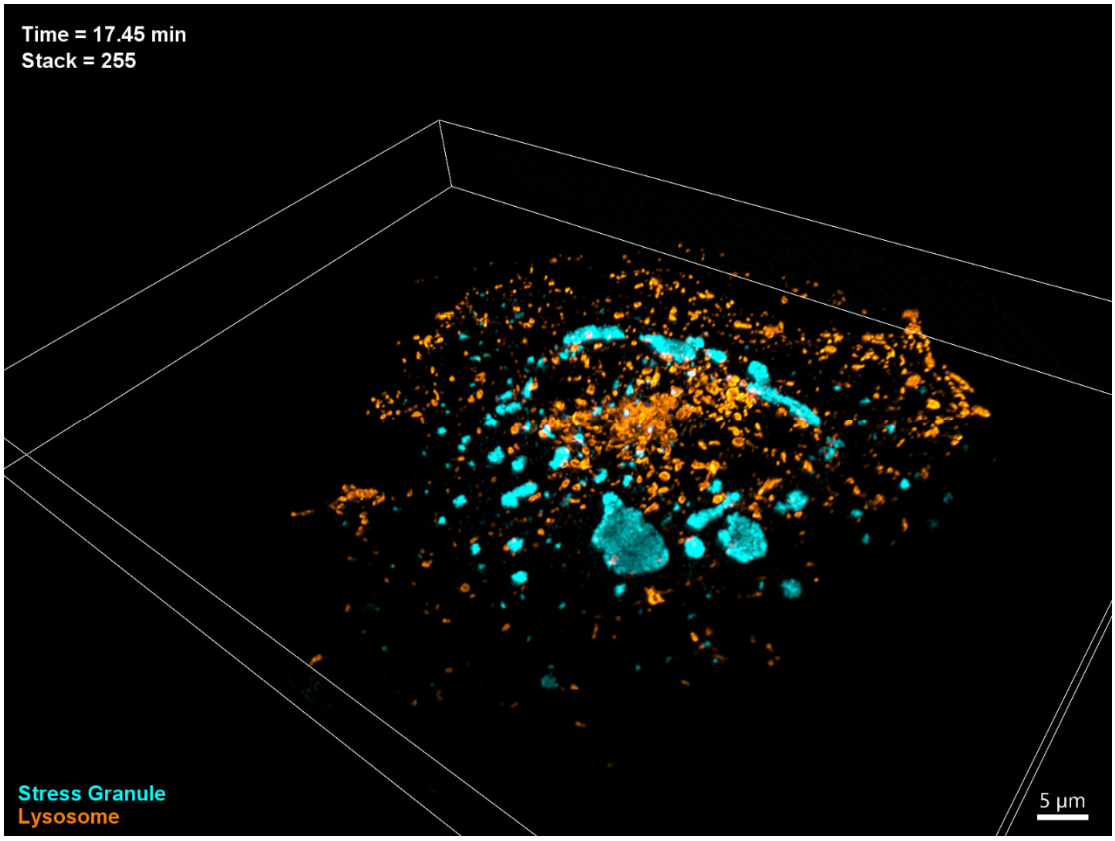
Supplementary Video 9 | 3-D projection of rDL LA-NLSIM-*z* image acquired from fixed COS-7 cell expressing Ensconsin-Skylan-NS. The magnified views show the resolution and SNR comparison of LLSM (top), LA-NLSIM-*z* (middle), and rDL LA-NLSIM-*z* (bottom). See also **Fig. 3.**



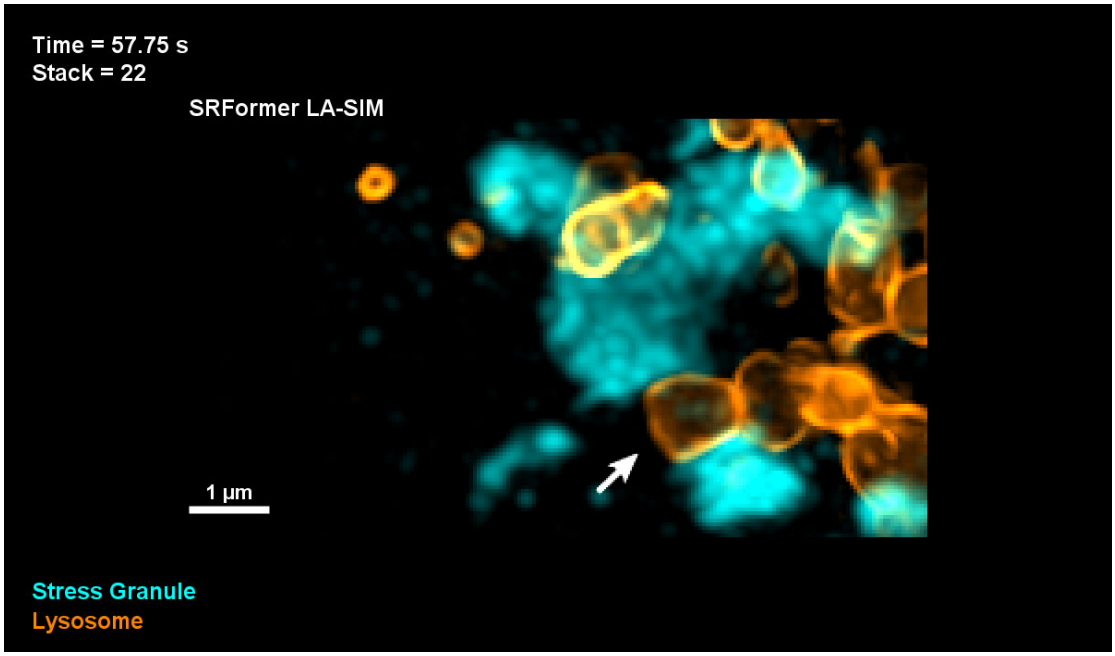
Supplementary Video 10 | Long-term three-color SRFormer LA-SIM imaging of COS-7 cell expressing Ensconsin-mStayGold (gray), SKL-mCherry (red) and LAMP1-HaloTag (blue), revealing the dynamic interactions among lysosomes, peroxisomes and microtubules over the whole cell volume for 690 time points lasting ~74 mins. See also **Fig. 4**.



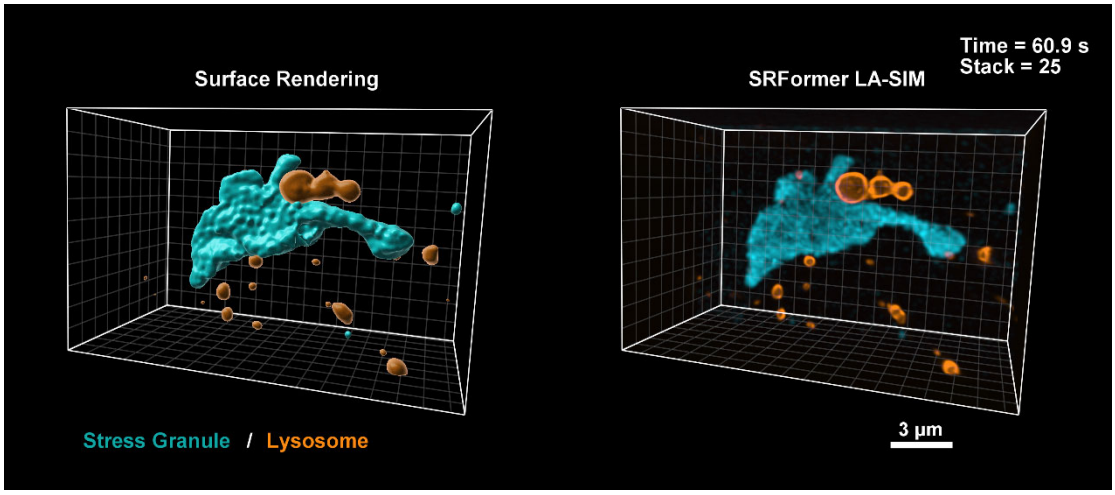
Supplementary Video 11 | Long-term two-color SRFormer LA-SIM imaging of COS-7 cell expressing Ensconsin-3×mStayGold and Tomm20-mCherry, showing the mitochondrial membrane dynamics and their translocation along microtubule tracks. See also **Extended Data Fig. 7**.



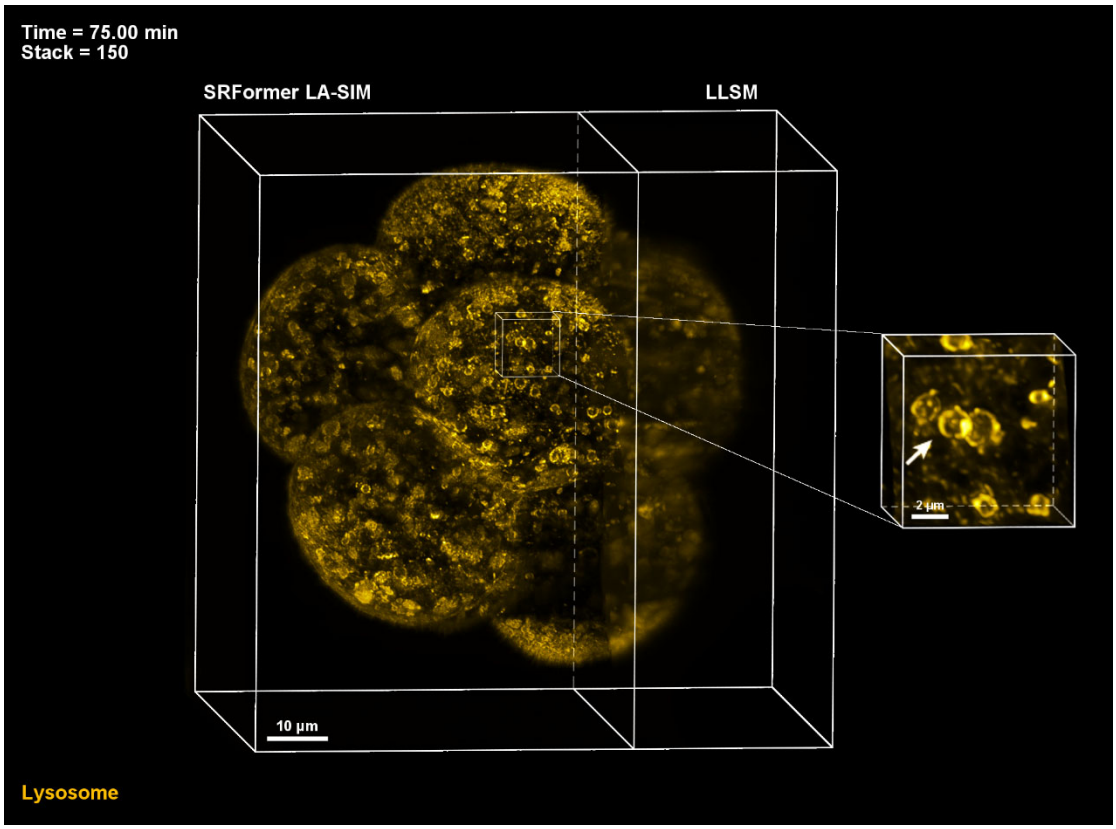
Supplementary Video 12 | Long-term two-color SRFormer LA-SIM imaging of COS-7 cell expressing G3BP1-mStayGold and LAMP1-Halo after being exposed to 500 µM $NaAsO_2$ for 30 min, showing the common dynamic interactions between lysosomes and stress granules over the whole cell volume for 500 time points lasting ~34 mins. See also **Fig. 5**.



Supplementary Video 13 | Two additional examples showing the lysosome movements mediate the fission of stress granule condensates.



Supplementary Video 14 | Two-color SRFormer LA-SIM imaging of COS-7 cell expressing G3BP1-mStayGold and LAMP1-Halo after being exposed to 500 µM $NaAsO_2$ for 30 min, showing that a moving lysosome mediates the fission of large stress granule condensates. See also **Fig. 5**.



Supplementary Video 15 | Long-term SRFormer LA-SIM imaging of mouse early embryo labeled with LAMP1-mStayGold, revealing the dynamics of each individual lysosome over the whole embryo range for 300 time points lasting 2.5 hours. See also **Fig**. **5**, **extended Data Fig. 8**.

Supplementary References

1. Jensen, N. A., Jansen, I., Kamper, M. & Jakobs, S. Reversibly switchable fluorescent proteins for RESOLFT nanoscopy. *Nanoscale Photonic Imaging* 241–261 (2020).
2. Chang, H. *et al.* A unique series of reversibly switchable fluorescent proteins with beneficial properties for various applications. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 4455–4460 (2012).
3. Zhang, X. *et al.* Highly photostable, reversibly photoswitchable fluorescent protein with high contrast ratio for live-cell superresolution microscopy. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 10364–10369 (2016).
4. Li, D. *et al.* Extended-resolution structured illumination imaging of endocytic and cytoskeletal dynamics. *Science* **349**, (2015).
5. Habuchi, S. *et al.* Reversible single-molecule photoswitching in the GFP-like fluorescent protein Dronpa. *Proceedings of the National Academy of Sciences* **102**, 9511–9516 (2005).
6. Grotjohann, T. *et al.* rsEGFP2 enables fast RESOLFT nanoscopy of living cells. *eLife* **1**, e00248 (2012).
7. Chen, J. *et al.* Three-dimensional residual channel attention networks denoise and sharpen fluorescence microscopy image volumes. *Nat Methods* **18**, 678–687 (2021).
8. Zhang, Y. *et al.* Image Super-Resolution Using Very Deep Residual Channel Attention Networks. in *Computer Vision – ECCV 2018* (eds. Ferrari, V., Hebert, M., Sminchisescu, C. & Weiss, Y.) vol. 11211 294–310 (Springer International Publishing, Cham, 2018).
9. Liang, J. *et al.* SwinIR: Image Restoration Using Swin Transformer. Preprint at https://doi.org/10.48550/arXiv.2108.10257 (2021).
10. Chen, Z. *et al.* Dual Aggregation Transformer for Image Super-Resolution. Preprint at

https://doi.org/10.48550/arXiv.2308.03364 (2023).

11. Liu, Z. *et al.* Video Swin Transformer. in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 3192–3201 (IEEE, New Orleans, LA, USA, 2022). doi:10.1109/CVPR52688.2022.00320.
12. Qiao, C. *et al.* Evaluation and development of deep neural networks for image super-resolution in optical microscopy. *Nat Methods* **18**, 194–202 (2021).
13. Zhu, Q., Li, P. & Li, Q. Attention Retractable Frequency Fusion Transformer for Image Super Resolution. in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 1756–1763 (IEEE, Vancouver, BC, Canada, 2023). doi:10.1109/CVPRW59228.2023.00176.
14. Chen, X. *et al.* A Comparative Study of Image Restoration Networks for General Backbone Network Design. Preprint at https://doi.org/10.48550/arXiv.2310.11881 (2024).
15. Kaplan, J. *et al.* Scaling Laws for Neural Language Models. Preprint at https://doi.org/10.48550/arXiv.2001.08361 (2020).
16. Liu, Z. *et al.* Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 9992–10002 (IEEE, Montreal, QC, Canada, 2021). doi:10.1109/ICCV48922.2021.00986.