Supplementary Materials

Deep Learning-Enabled Transformation of Scanning Superlens Microscopy Images into Scanning Electron Microscopy-like Large Depth-of-Field Images

Hui SUN[1, #], Hao LUO[2,3, #], Feifei WANG[4,#], Qingjiu CHEN[1], Meng CHEN[1], Xiaoduo WANG[2,3], Haibo YU[2,3], Guanglie ZHANG[1,5,*], Lianqing LIU[2,3,*], Jianping WANG[5,6], Dapeng WU[5,6,*], Wen Jung LI[1,2,3,5,*]

1. Department of Mechanical Engineering, City University of Hong Kong, Hong Kong SAR, China.
2. State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China.
3. Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169, China.
4. Department of Electrical and Electronics Engineering, The University of Hong Kong, Hong Kong SAR, China.
5. City University of Hong Kong Shenzhen Research Institute (CityUSRI), Shenzhen 518000, China.
6. Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China.
# These authors have contributed equally to this work.
* CORRESPONDING AUTHORS: Guanglie Zhang; Lianqing Liu; Dapeng Wu; Wen Jung Li (e-mail: gl.zhang@cityu.edu.hk; lqliu@sia.cn; dapengwu@cityu.edu.hk; wenjli@cityu.edu.hk)

Quantification of performance for each network. Image quality refers to the visual attributes of images, and focuses on the perceptual assessment of viewers. In general, image quality assessment (IQA) methods include subjective methods based on human perception (i.e., how realistic the image looks) and objective computational methods. The former methods are more in line with this study's needs but as they are often time-consuming and expensive, the latter methods are currently the mainstream. However, objective methods do not necessarily produce consistent assessments, because they are often unable to capture human visual perception very accurately, which can lead to large variation in IQA results[1].

The PSNR and SSIM are currently the most commonly used super-resolution metrics. The PSNR is used to compare the difference in gray value of the corresponding pixels of the target image and the reference image to determine image quality. However, the PSNR is restricted in its capacity to capture perceptually relevant differences, such as high texture detail, because it is defined by pixel-wise image differences. A higher PSNR does not always imply a perceptually superior super-resolution performance.

SSIM is a subjective measure of structural similarity between two images based on three relatively independent evaluations of brightness, contrast, and structure. The SSIM can be

computed individually and presented as a weighted product of brightness, contrast, and structure comparisons. The higher the SSIM value, the less image distortion there is, and the better is the image quality. However, SSIM is still not completely accurate in assessing perceptual quality. We therefore use four metrics to assess the performance of our networks throughout the image testing process, namely grayscale variation of images in the spatial domain, variation of images in the frequency domain, PSNR, and SSIM. The image evaluation metrics of PSNR and SSIM are calculated as shown in Equation 1 and Equation 2:

$$PSNR = 20 \times \log_{10}\left(\frac{255}{\sqrt{\sum_{i=1}^{W}\sum_{j=1}^{H}(U_x(i,j)-U_{x'}(i,j))^2/(W\times H)}}\right) \tag{1}$$

$$SSIM = \left(\frac{2U_xU_{x'}+C_1}{U_x{}^2+U_{x'}{}^2+C_1}\right)\times\left(\frac{2\delta_{xx'}+C_2}{\delta_x^2+\delta_{x'}^2+C_2}\right) \tag{2}$$

where W and H represent the width and height of the reference image in the training step (W = 256, H = 256 across all networks); U_x denotes the average grayscale value of the original image, and δ_x denotes the variance of the original image; $U_{x'}$ represents the average grayscale value of the reconstructed high-resolution image, and $\delta_{x'}$ represents the variance of the reconstructed high-resolution image; $\delta_{xx'}$ is the covariance of U_x and $U_{x'}$; C_1 and C_2 are constants. For the image evaluations in this paper, we randomly select 12 super-resolution images generated by CycleGAN with multi-residual blocks in the test datasets and evaluate their image quality, with the results shown in **Fig. S1** and **Table S1**. In terms of PSNR and SSIM values, the deep learning-based generated image performs the best on all the test datasets, suggesting that the generated super-resolution images provide more detailed information than the LR images and bring the generated results closer to the HR images. These comparisons firmly demonstrate the effectiveness of the multi-residual-block CycleGAN super-resolution image reconstruction model and indicate that using an optical super-resolution imaging system combined with deep learning improves performance. However, the SSIM values are low, which is mainly a result of the inconsistent structural information of the objects in the scene generating a small deviation in the image registration step in the image pre-processing stage. This generates a large difference in the pixel-wise values between the two images when calculating the SSIM, and thus gives low SSIM values. Another reason for the smaller SSIM values is that some noise is introduced during the imaging phase of optical super-resolution, and after the training of the model this noise generates erroneous features, which directly leads to a reduction in the SSIM value. In future work, we will increase the SSIM value in terms of image registration precision

and consequently improve the quality of the produced images.

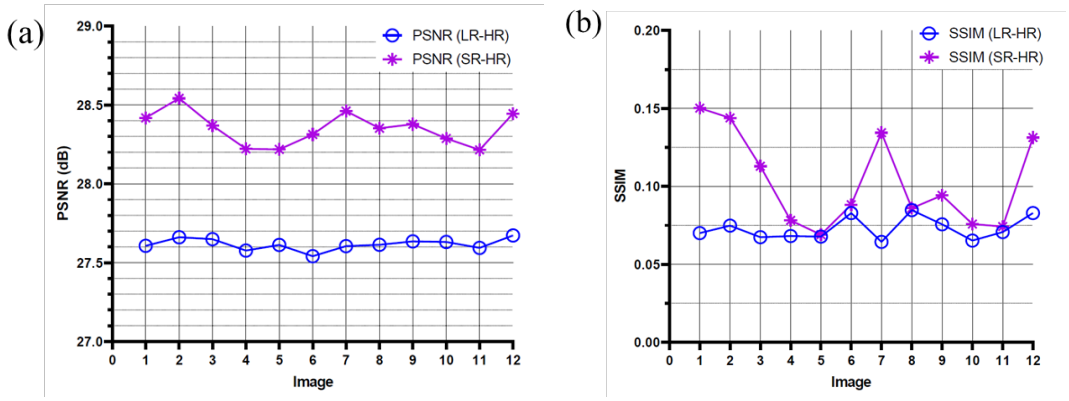

Figure S1. Quantitative performance indicators for super-resolution image quality. Results of evaluations of optical super-resolution images and deep learning-based generated super-resolution images for **a** PSNR and **b** SSIM.

Table S1. Quantitative image quality results for the optical super-resolution images and super-resolution reconstructed images based on multi-residual-block CycleGAN model compared with SEM images.

Tested Image	PSNR (dB) (LR-HR)	SSIM (LR-HR)	PSNR (dB) (SR-HR)	SSIM (SR-HR)
1	27.60745	0.06996	28.41808	0.15016
2	27.66148	0.07481	28.54237	0.14391
3	27.64884	0.06733	28.36969	0.11288
4	27.57721	0.06815	28.22173	0.07817
5	27.61247	0.06776	28.21874	0.06900
6	27.54138	0.08290	28.31287	0.08825
7	27.60484	0.06440	28.46176	0.13442
8	27.61344	0.08473	28.35220	0.08594
9	27.63534	0.07564	28.37810	0.09422
10	27.63165	0.06518	28.28794	0.07580
11	27.59391	0.07064	28.21596	0.07420
12	27.67237	0.08287	28.44538	0.13130

Model robustness validation tests. To verify the robustness of the trained super-resolution model, another type of IC chip is used as test data. In this model validation trial, the image texture in the test set contains many line segments with large curvature, and most of these texture features are not present in the training set. The same data augmentation methods are used during the training process: image flips, image rotation at different degrees, image scaling As shown in **Fig. S2 (a1)–(c1)**, some of the features never appear in the training dataset but the resulting super-resolution images still provide a relatively clear and accurate picture of the texture of the IC chip. Although there are some minor differences in the presentation of some details compared with SEM, this can be addressed by expanding the richness of the IC chip texture in the training dataset, thus extending the usefulness and increasing the application value of the method.

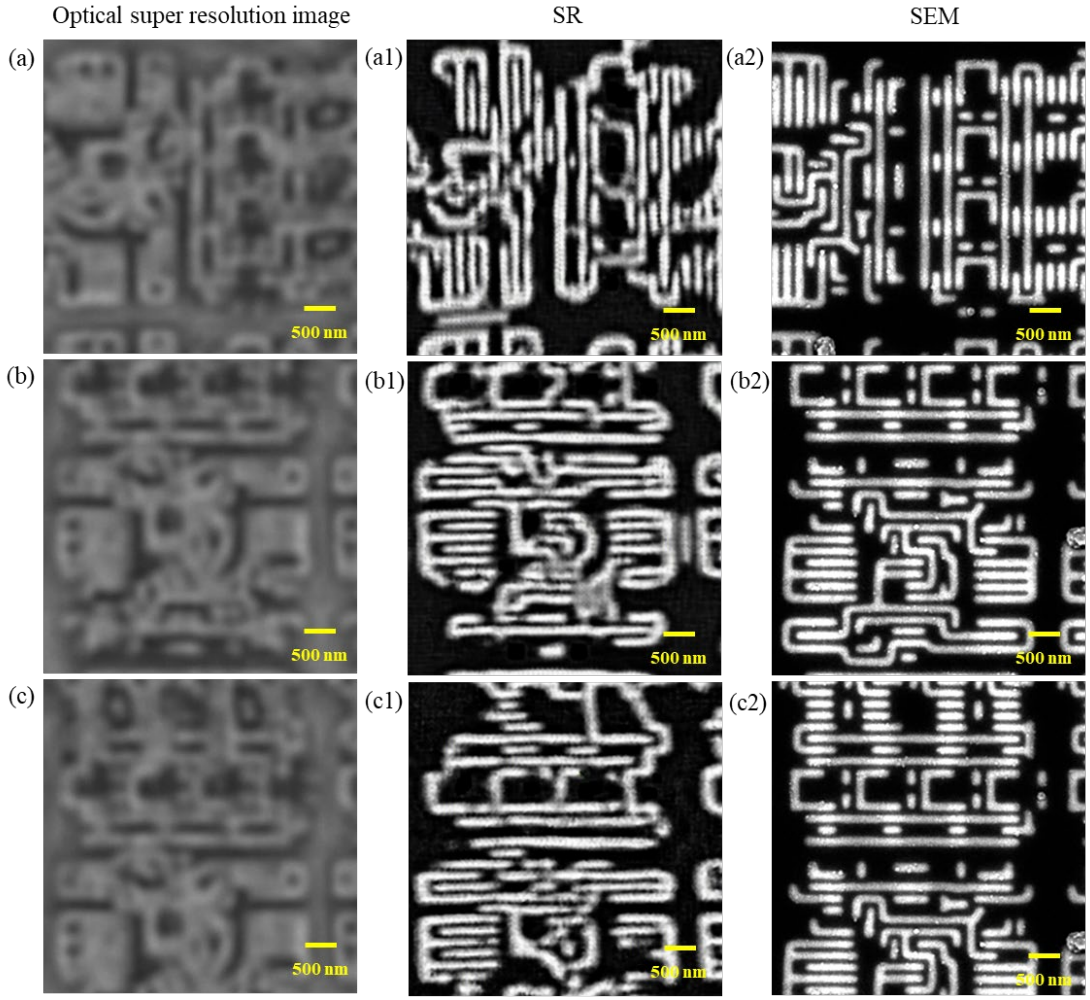

Figure S2. Super-resolution model robustness validation tests. a–c: SSUM imaging; **a1**–**c1**: Super-resolution imaging results based on CycleGAN with multi-residual blocks; **a2–c2**: SEM imaging (reference). Scale bar: 500 nm.

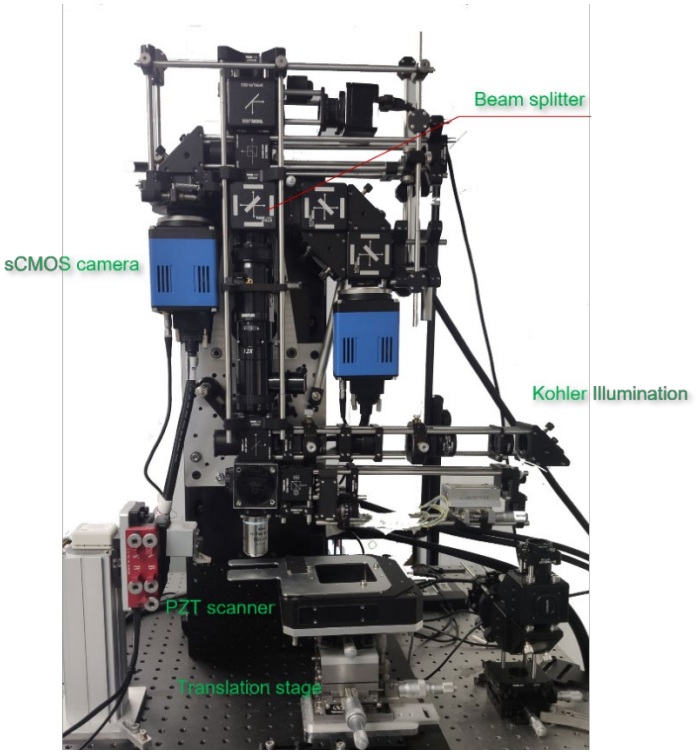

Figure S3. Physical and optical schematic diagram of the optical scanning superlens microscopy system (SSUM).

Super-resolution microsphere scanning microscope system components 3D piezoelectric ceramic (PZT) scanner (P-733.3CL, Physik Instrumente, Germany). Three-dimensional translation stage (MAXYZ-60L). AFM cantilever (TESP probe, Bruker). 100× objective lens (Nikon LU Plan EPI ELWD). UV-curable adhesive (NOA63, Edmund Optics). High-speed scientific complementary metal oxide semiconductor (sCMOS) camera (PCO.Edge 5.5). Intensity-controlled light source (C-HGFI, Nikon, Japan).

Scanning electron microscope parameters. A field-emission gun scanning electron microscope (Quanta 450 FEG-SEM) was used to perform SEM imaging of the silicon wafer samples. The beam voltage was 5 keV and the spot size was 3.0 for high-resolution imaging.

Technical characteristics. The micro/nano optical SSUM 3D imaging platform was constructed to enable observation and manipulation at micro- and nano-scales. In addition, using the fabricated microsphere lens with a diameter of 10–200 μm for optical super-resolution imaging, the 2D (X/Y plane) resolution of the image in bright field can reach 18.4 nm per pixel, far below the image resolution of 500 nm. Finally, the deep learning super-resolution model performs further image reconstruction on the optical image, which can achieve super-resolution imaging with a large field of view, and we use a 3D translation stage (MAXYZ-60L, with a range of movement on the X and Y axes of 13 mm) to realize the scanning of the sample, which can easily achieve an imaging range of greater than 10 mm × 10 mm.

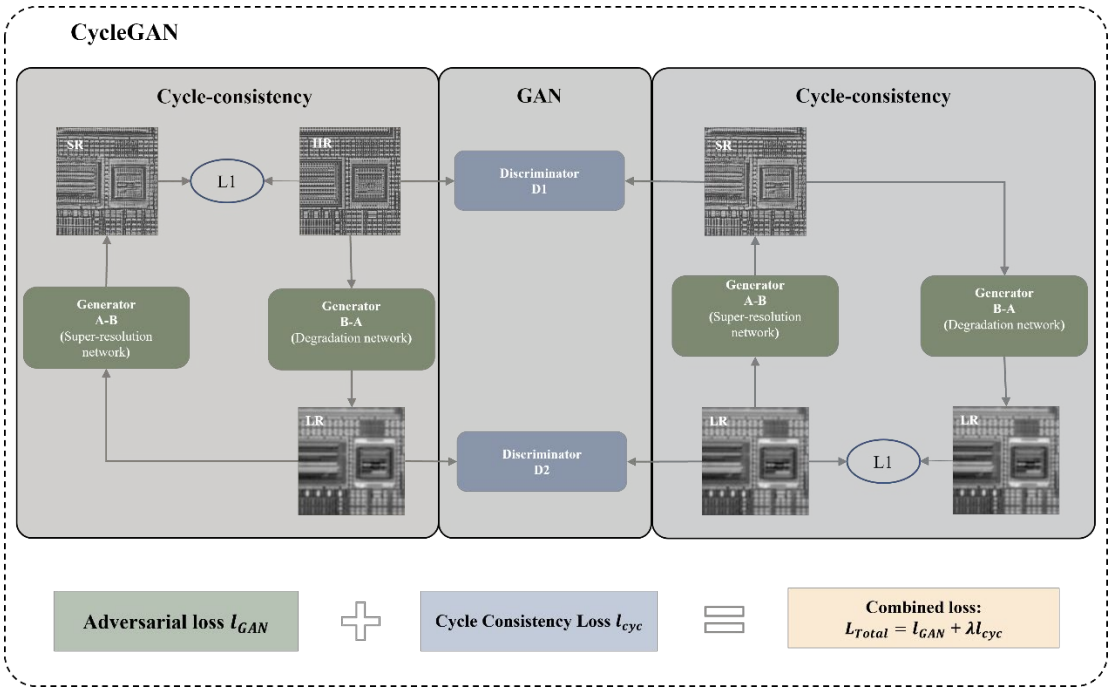

Figure S4. Cycle-consistency-based image super-resolution network structure.

Experimental Setup. The super-resolution results were evaluated using the PSNR and SSIM indices on the Y channel in the YCbCr space. We rescaled the training images to the same physical size to train all the models. The experiments were conducted on a PC equipped with an Intel Core i7-10700F CPU, 32G RAM, and a single Nvidia RTX2060S GPU (8G).

Training Details and Parameters. We trained all networks on an Nvidia RTX2060S GPU (8G) using a random sample of 81 images from the database: 50 for training and 31 for testing. The testing images were distinct from the training images. We obtained the LR and HR images using microscope superlens arrays and SEM, respectively. Our generator network has nine identical residual blocks. For the experiment, we set $\lambda = 10$ in Equation 5. For optimization, Adam was used to improve the network settings with $\beta_1=0.9$, $\beta_2=0.99$ and a batch size of 1. Each actual high-quality image of each batch was cropped into numerous 512×512 pixels sub-images and resized into 256×256 pixels for training. The high-quality images and the corresponding low-quality images produced from the preceding processes were then utilized as inputs for iterative training in the discriminator and generator networks, respectively. The CycleGAN with multi-residual-block networks was trained from scratch with a learning rate of 2×10^{-4}, and we kept the same learning rate for the first 500 epochs and linearly decayed the rate to zero over the next 1,000 epochs. To avoid undesired local optima, we used the trained MSE-based ResNet network for the initialization of the generator when training the actual CycleGAN with multi-residual blocks.

Loss functions. In the super-resolution field, loss functions are used to measure reconstruction error and guide the model optimization.

CycleGAN Loss Function. Our goal is to learn mapping functions between two domains: LR and HR. As illustrated in **Fig. S4**, our model includes two mappings, $G_{LR-HR}: LR \to HR$ and $G_{HR-LR}: HR \to LR$. In addition, we introduce two adversarial discriminators $D_{I^{LR}}$ and $D_{I^{HR}}$, where $D_{I^{LR}}$ aims to distinguish between SEM images and translated images; in the same way, $D_{I^{HR}}$ aims to distinguish between optical super-resolution images and translated images. Our objective contains two kinds of terms: adversarial losses for matching the distribution of generated images to the data distribution in the target domain, and a cycle consistency loss to prevent the learned mappings G_{LR-HR} and G_{HR-LR} from contradicting each other.

Adversarial Loss. The mapping of optical super-resolution images to SEM images using CycleGAN is represented by $G_{LR-HR}: LR \to HR$. The discriminator $D_{I^{HR}}$ is used to distinguish whether the generated image is real or not, and the objective function is shown below:

$$\begin{aligned} \ell_{GAN}(G_{LR-HR}, D_{I^{HR}}, I^{LR}, I^{HR}) = & \mathrm{E}_{I^{HR} \sim P_{data}(I^{HR})}[\log D_{I^{HR}}(I^{HR})] \\ & + \mathrm{E}_{I^{LR} \sim P_{data}(I^{LR})}[\log(1 - D_{I^{HR}}(G_{LR-HR}(I^{LR})))] \end{aligned} \tag{3}$$

The generator G_{LR-HR} converts the optical super-resolution image to an SEM image $G_{LR-HR}(I^{LR})$ in the target domain to the maximum extent possible, and then $D_{I^{HR}}$ is used to distinguish between the generated SEM image and the real SEM image, which is a generative adversarial process. In contrast, the CycleGAN training process requires converting the SEM image into an optical super-resolution image, i.e. $G_{HR-LR}: I^{HR} \to I^{LR}$, to determine the truth of the generated image by the discriminator $D_{I^{LR}}$ with the loss function: $\ell_{GAN}(G_{HR-LR}, D_{I^{LR}}, I^{HR}, I^{LR})$.

Cycle Consistency Loss. Cycle consistency loss is an important part of CycleGAN that is used to overcome the problem of inconsistencies between the output distribution and the target distribution. The inputs in this paper are the optical super-resolution images and SEM-registered image pairs, and using only an adversarial loss function is not sufficient to ensure that the input optical super-resolution image can be mapped to the desired output SEM image. Therefore, this paper introduces forward consistency loss and backward consistency loss for each image from the LR and HR domains, which can be transformed back to the original image according to the cycle consistency principle. $G_{HR-LR}(G_{LR-HR}(I^{LR}))$ cycles $G_{LR-HR}(I^{LR})$ to I^{LR}. At the same time, $G_{LR-HR}(G_{HR-LR}(I^{HR}))$ recycles $G_{HR-LR}(I^{HR})$ to I^{HR}. There are two constraints in the training: $G_{HR-LR}(G_{LR-HR}(I^{LR})) \approx I^{LR}$, $G_{LR-HR}(G_{HR-LR}(I^{HR})) \approx I^{HR}$. We can incentivize this behavior using a cycle consistency loss:

$$\ell_{cyc}(G_{LR-HR}, G_{HR-LR}) = \mathrm{E}_{I^{LR}\sim P_{data}(I^{LR})}[\left\|G_{HR-LR}(G_{LR-HR}(I^{LR})) - I^{LR}\right\|_1] + \mathrm{E}_{I^{HR}\sim P_{data}(I^{HR})}[\left\|G_{LR-HR}(G_{HR-LR}(I^{HR})) - I^{HR}\right\|_1] \quad (4)$$

Our full objective is:

$$\ell(G_{LR-HR}, G_{HR-LR}, D_{I^{LR}}, D_{I^{HR}}) = \ell_{GAN}(G_{LR-HR}, D_{I^{HR}}, I^{LR}, I^{HR}) + \ell_{GAN}(G_{HR-LR}, D_{I^{LR}}, I^{HR}, I^{LR}) + \lambda\ell_{cyc}(G_{LR-HR}, G_{HR-LR}) \quad (5)$$

where λ is the weighting factor for the cycle consistency loss, which is set to 10 according to ref [2]. Finally, the optimization process of CycleGAN is represented by the following equation:

$$G_{LR-HR}^{*}, G_{HR-LR}^{*} = \arg \min_{G_{LR-HR}, G_{HR-LR}} \max_{D_{I^{HR}}, D_{I^{LR}}} \ell(G_{LR-HR}, G_{HR-LR}, D_{I^{LR}}, D_{I^{HR}}) \quad (6)$$

Fig. S5 shows the error of the generator and both discriminators for the proposed CycleGAN with multi-residual blocks after 1,500 training epochs. As the generator error approaches zero, the reconstructed image becomes closer to the real image. Meanwhile, as the discriminator error approaches 0.5, the discriminators lose the ability to determine whether the generated image is real; that is, the reconstructed image becomes more similar to the original image. According to the results in **Fig. S5**, the super-resolution reconstruction of a low-resolution image progressively produces robust results as the number of epochs increases.

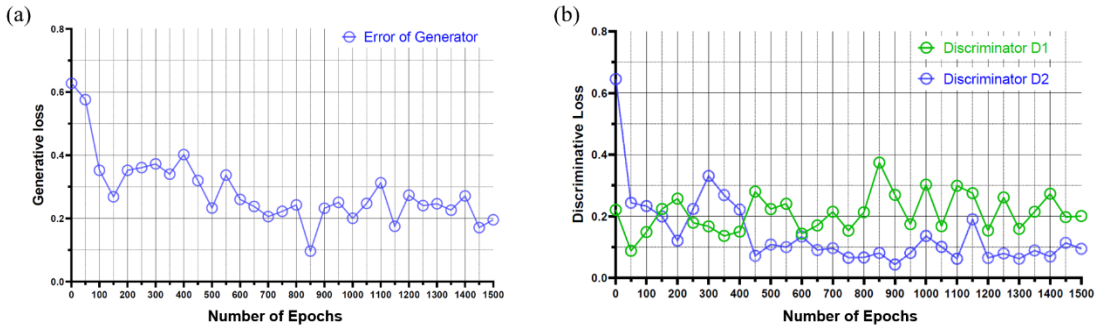

Figure S5. Errors of CycleGAN with multi-residual blocks versus the number of iterations. a Error of generator; **b** Error of discriminators.

References

1. Wang, Z., Chen, J. & Hoi, S. C. H. Deep Learning for Image Super-resolution: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell*. **43**, 3365-3387 (2020).
2. Zhu, J., Park, T., Efros, A. A., Ai, B. & Berkeley, U. C. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *in Proceedings of the IEEE international conference on computer vision*. 2223-2232 (2017).