Sun et al. *Microsystems & Nanoengineering* (2025)11:252
https://doi.org/10.1038/s41378-025-01060-9

Microsystems & Nanoengineering
www.nature.com/micronano

ARTICLE Open Access

Deep learning-driven conversion of scanning superlens microscopy to high depth-of-field SEM-like imaging

Hui Sun[1,2], Hao Luo[3,4], Feifei Wang[5], Qingjiu Chen[1], Meng Chen[1,2], Xiaoduo Wang[3,4], Haibo Yu[3,4], Guanglie Zhang[1,2]✉, Lianqing Liu[3,4]✉, Jianping Wang[2,6], Dapeng Wu[2,6]✉ and Wen Jung Li[1,2,3,4]✉

Abstract

Scanning electron microscopy (SEM) enables nanoscale imaging but requires vacuum environments and coating samples with conductive films. We present a deep learning approach to transform optical super-resolution (OSR) microscopy images into high-resolution images resembling SEM, specifically optimized for chip samples. Utilizing our custom-designed scanning superlens microscopy (SSUM) system, we acquire OSR images with a resolution down to ~80 nm without the need for coatings or vacuum conditions. Notably, the SSUM system achieves an effective depth-of-field (DoF) of approximately 2 μm through Z-stack scanning, enabling clear visualization of multilayer chip structures across a larger axial range than conventional optical imaging. Our algorithm further enhances the nanoscale microstructures observed with the SSUM platform, significantly improving the visibility of structures that are otherwise less distinct. A cycle-consistent generative adversarial network (CycleGAN) model is trained on paired OSR and SEM images to learn the mapping between these imaging modalities. The model is then applied to unseen OSR test images from silicon wafer samples. Quantitative analysis shows that the reconstructed images achieve a mean peak signal-to-noise ratio (PSNR) 1.64 dB higher than the input OSR images. Qualitative assessment further demonstrates the model's ability to generate results with high structural detail, specifically in chip-level applications. This technique overcomes key SEM constraints while preserving nanoscale resolution, offering the potential for advanced chip manufacturing and inspection tasks where traditional SEM requirements pose challenges.

Introduction

The resolution of conventional optical microscopes is fundamentally constrained by diffraction, limiting their ability to resolve nanoscale features smaller than approximately 200 nm. Overcoming this diffraction barrier has been a critical focus for advancing both fundamental research and nanoscale manufacturing. As a result, over the past two decades, various super-resolution microscopy techniques have emerged to surpass this limit. These techniques can be broadly categorized into four main groups based on their underlying principles: fluorescence-based super-resolution microscopy[1], surface plasmon polariton microscopy[2], structured illumination microscopy[3], and microsphere lens-based super-resolution microscopy[4]. These innovative approaches have significantly extended the capabilities of optical microscopy, enabling more detailed studies of nanoscale structures.

Fluorescence-based super-resolution microscopy has significantly advanced beyond the diffraction limit of conventional optical microscopy, with notable techniques including stimulated emission depletion microscopy (STED)[5], structured illumination microscopy (SIM)[6], photoactivated localization microscopy (PALM)[7],

Correspondence: Guanglie Zhang (gl.zhang@cityu.edu.hk) or Lianqing Liu (lqliu@sia.cn) or Dapeng Wu (dapengwu@cityu.edu.hk) or Wen Jung Li (wenjli@cityu.edu.hk)
[1]Department of Mechanical Engineering, City University of Hong Kong, Hong Kong SAR, China
[2]City University of Hong Kong Shenzhen Research Institute (CityUSRI), Shenzhen 518000, China
Full list of author information is available at the end of the article
These authors contributed equally: Hui Sun, Hao Luo, Feifei Wang.

stochastic optical reconstruction microscopy (STORM)[8], and super-resolution optical fluctuation imaging (SOFI)[9]. STED achieves sub-diffraction resolution by suppressing peripheral fluorescence through stimulated emission, while SIM uses patterned illumination and computational reconstruction to enhance resolution. PALM and STORM utilize photoswitchable fluorophores for nanometer-precision localization, and SOFI improves resolution and signal-to-noise ratio by analyzing temporal fluorescence fluctuations, offering a less phototoxic solution for live-cell imaging.

Despite their success, these methods face inherent limitations due to their dependence on fluorescent molecules, which can restrict their broader application. The high cost of equipment and difficulties in fluorescent tagging further hinder widespread adoption. In 2011, Wang et al.[10] introduced a microsphere-based super-resolution imaging technique that uses transparent dielectric microspheres placed on sample surfaces to achieve resolutions of up to 50 nm under a conventional optical microscope. This method offers a straightforward and efficient approach to real-time, in-situ super-resolution imaging. Subsequently, in 2015, our team (Wang et al.[11]) enhanced this technique by integrating dielectric microspheres with atomic force microscopy (AFM) probes, enabling precise nanoscale control and large-field-of-view imaging. In that work, we employed a Z-stack scanning method and quantitatively demonstrated that the SSUM system could achieve an effective depth-of-field of approximately 2 μm, allowing reliable visualization of nanoscale features across multiple axial planes. However, while effective, the contrast and depth-of-field of the images obtained remain inferior to those produced by scanning electron microscopy (SEM). SEM, despite its ability to surpass the optical diffraction limit, presents challenges such as the requirement for vacuum environments and sample coating with conductive films, rendering it unsuitable for certain applications. Moreover, achieving high-resolution SEM images requires complex and expensive hardware, making it less practical for many scenarios.

In contrast, "super-resolution" in computer vision refers to improving image resolution through software, often by reconstructing high-frequency details using deep learning. This process, known as "neural network-based super-resolution," typically involves training deep convolutional networks on paired low- and high-resolution images to learn the mapping between them. Deep learning methods have been successfully applied to achieve super-resolution imaging with optical microscope data, bridging the gap between low-resolution input and high-resolution output[12–16]. In this work, we apply deep learning to reconstruct SEM images from optical super-resolution images, bypassing traditional SEM constraints while enhancing resolution.

Our proposed approach integrates a custom scanning superlens microscopy (SSUM) system with a cycle-consistent generative adversarial network (CycleGAN) deep learning model.

The SSUM system, which utilizes microsphere-assisted imaging (as shown in Fig. 1 and Fig. S1), is combined with deep learning to bridge the gap between optical and SEM imaging. Specifically, a probe-lens assembly was designed using barium titanate glass microspheres in combination with a scanning platform capable of large-field-of-view imaging. By training the CycleGAN model on paired OSR and SEM images, our method achieves high-resolution reconstructions that closely mimic SEM images, particularly in chip-level applications, as illustrated in the SSUM-CycleGAN super-resolution workflow (Fig. 2). In summary, this work demonstrates a novel method for transforming optical super-resolution images into SEM images through deep learning, addressing key challenges such as vacuum requirements and sample preparation while pushing the limits of resolution beyond conventional optical systems.

Results

Analysis and comparison of SSUM-CycleGAN super-resolution results in the spatial domain

The imaging quality of the super-resolution images generated using the deep learning approach progressively improves as the number of training iterations increases. Notably, at 500 epochs, both the level of detail and contrast in the images show substantial enhancement. The CycleGAN model, incorporating multi-residual blocks, is particularly effective in producing high-resolution images, as demonstrated in Fig. 3. The use of residual blocks allows the neural network to efficiently learn and retain high-resolution features, leading to faster convergence during training.

To assess the quality and fidelity of the reconstructed images, the grayscale distribution within the 3D spatial domain, reflecting the intensity magnitude of each pixel, was analyzed. This distribution serves as an important metric for evaluating global image similarity. By examining the grayscale distribution, we can visualize and quantify changes in the overall image characteristics, offering a more objective basis for comparison. Therefore, the SSUM image (Fig. 3a), the SEM reference image (Fig. 3b), and the selected region used for testing the super-resolution model (Fig. 3c) are depicted. The optical super-resolution image obtained from the SSUM system serves as the input to the model, with Fig. 3d showing the reconstructed super-resolution image generated by the deep learning-based super-resolution model, and Fig. 3e presenting the corresponding local reference image.

In Fig. 3f–h, the global grayscale distributions of the optical super-resolution images obtained using the SSUM system, the super-resolution images reconstructed

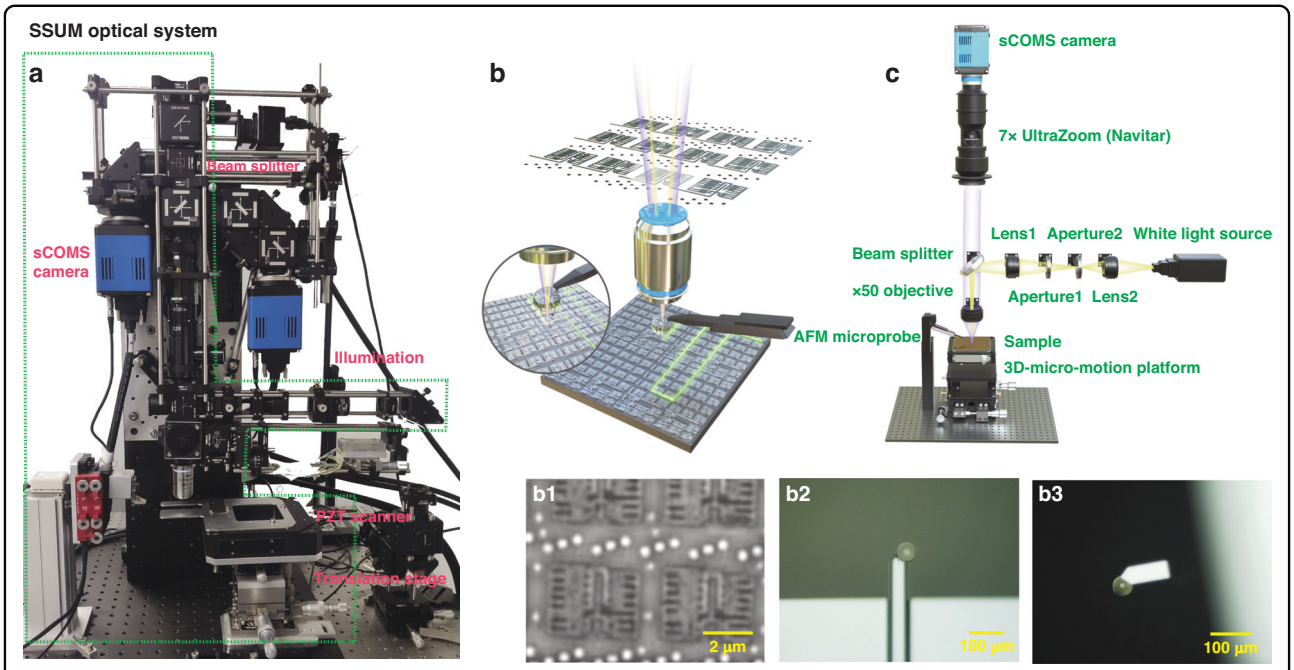

Fig. 1 Scanning superlens microscopy (SSUM). a Photograph of the SSUM setup. **b** Schematic of a microsphere-based SSUM that integrates a microsphere superlens into an AFM scanning system by connecting the microsphere to an AFM cantilever. The objective collects virtual images containing sub-diffraction-limited object information while also focusing and collecting the laser beam utilized in the cantilever deflection detection system. **b1** An original virtual image observed using the microsphere superlens. The inset shows an SSUM image; **b2** and **b3** are front-side and rear-side views, respectively, of the AFM cantilever with the attached microsphere superlens. **c** A Schematic diagram of the optical path of the SSUM system. Scale bars: 2 μm (**b1**); 100 μm (**b2**, **b3**)

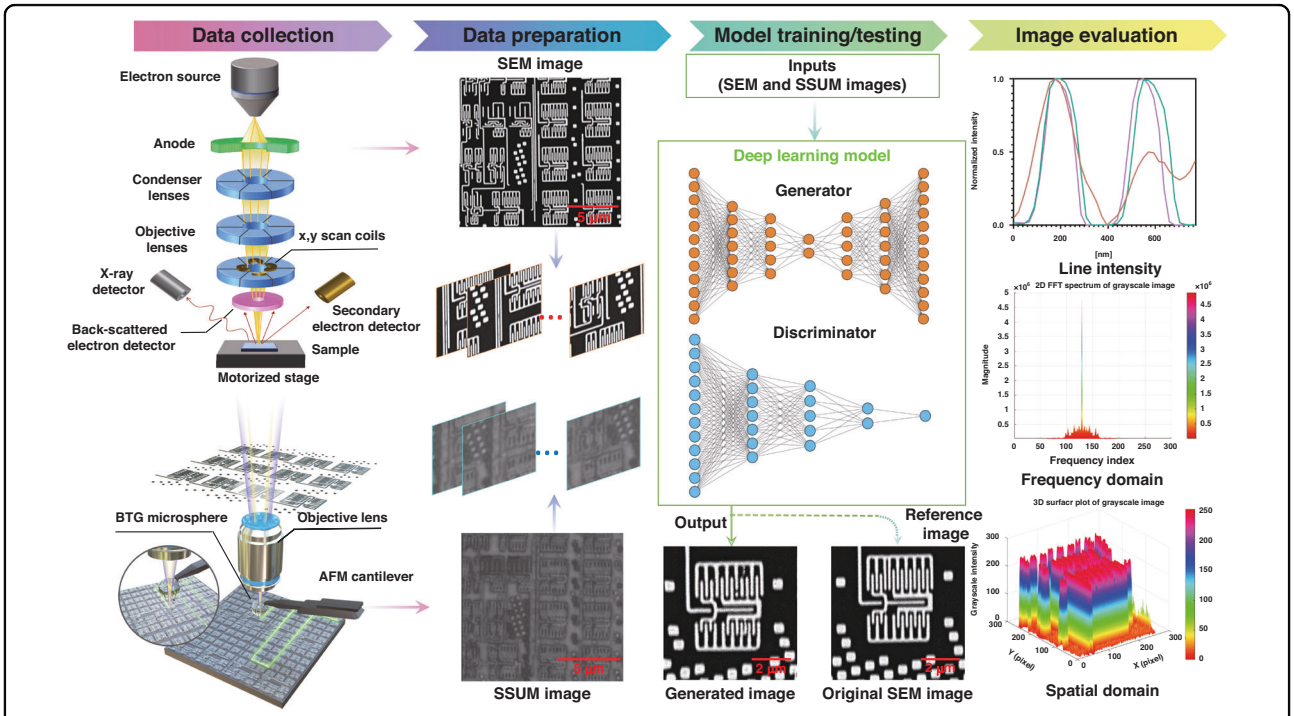

Fig. 2 Schematic diagram of image super-resolution reconstruction for the proposed optical SSUM system using the reference-based super-resolution algorithm

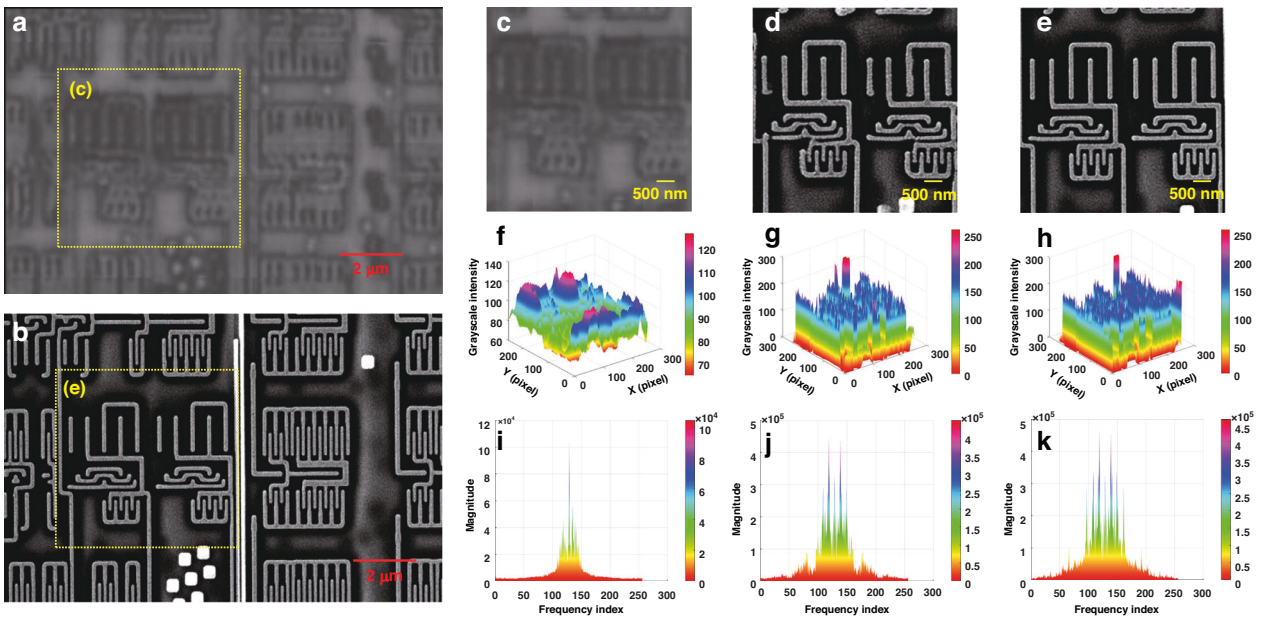

Fig. 3 Comparison of image super-resolution reconstruction results. a SSUM image; **b** SEM reference image; **c** selected region used for testing the super-resolution model. The optical super-resolution image obtained from SSUM serves as the input to the model, and **d** shows the reconstructed super-resolution image generated by the deep learning-based super-resolution model. **e** shows the corresponding local reference image. In the spatial domain, (**f**), (**g**), and (**h**) display the grayscale distribution at different locations for the SSUM image, the CycleGAN-based super-resolution reconstructed image, and the SEM image, respectively. In the frequency domain, (**i**), (**j**), and (**k**) present the results of the Fourier transform and frequency spectrum centering for the SSUM image, the CycleGAN-based super-resolution reconstructed image, and the SEM image, respectively

through CycleGAN, and the SEM images are presented. The results clearly indicate that the deep learning-based super-resolution reconstruction yields images with a grayscale distribution closely resembling that of real SEM images. This alignment is not merely a matter of visual enhancement but provides strong evidence of the method's reliability and accuracy. The proposed super-resolution approach, therefore, produces images that are not only visually comparable to SEM images but also maintain high fidelity, supporting its credibility for practical applications.

Analysis and comparison of SSUM-CycleGAN super-resolution results in the frequency domain

While spatial-domain analysis provides insight into the amplitude variations within an image, it falls short of offering a detailed understanding of the information frequency and the magnitude of various frequency components. In contrast, frequency-domain analysis, derived through the Fourier transform, reveals much richer information about the image's structure. For digital images, which are discrete signals, the frequency magnitude indicates the rate of signal change; higher frequencies correspond to more abrupt variations, typically associated with image edges and noise, while lower frequencies represent smoother transitions, capturing the general outline and background features of the image.

To explore the frequency characteristics of the images generated in this study, we performed a Fourier transform and analyzed the resulting frequency spectra. Figure 3i–k presents the frequency spectra for the SSUM optical super-resolution images, the deep learning-based super-resolution reconstructed images, and the SEM images, respectively. As expected, the low-frequency components, which represent the overall structure and background of the images, exhibit high energy and are concentrated near the center of the spectrum. Conversely, the high-frequency components, which correspond to finer details and edges, are distributed towards the periphery.

The analysis reveals that SSUM images, due to their inherent limitations, often lack significant high-frequency content, resulting in a relatively blurred appearance when compared to the images reconstructed using the multi-residual-block CycleGAN and the SEM images. This deficiency is particularly evident in the reduced sharpness and clarity of the SSUM images.

In contrast, the super-resolution images generated through the deep learning approach exhibit a frequency distribution closely aligned with that of the SEM images. The presence of well-distributed high-frequency components in these images highlights the effectiveness of the CycleGAN model in capturing and reconstructing fine details, thereby producing images that not only visually resemble SEM images but also retain critical structural

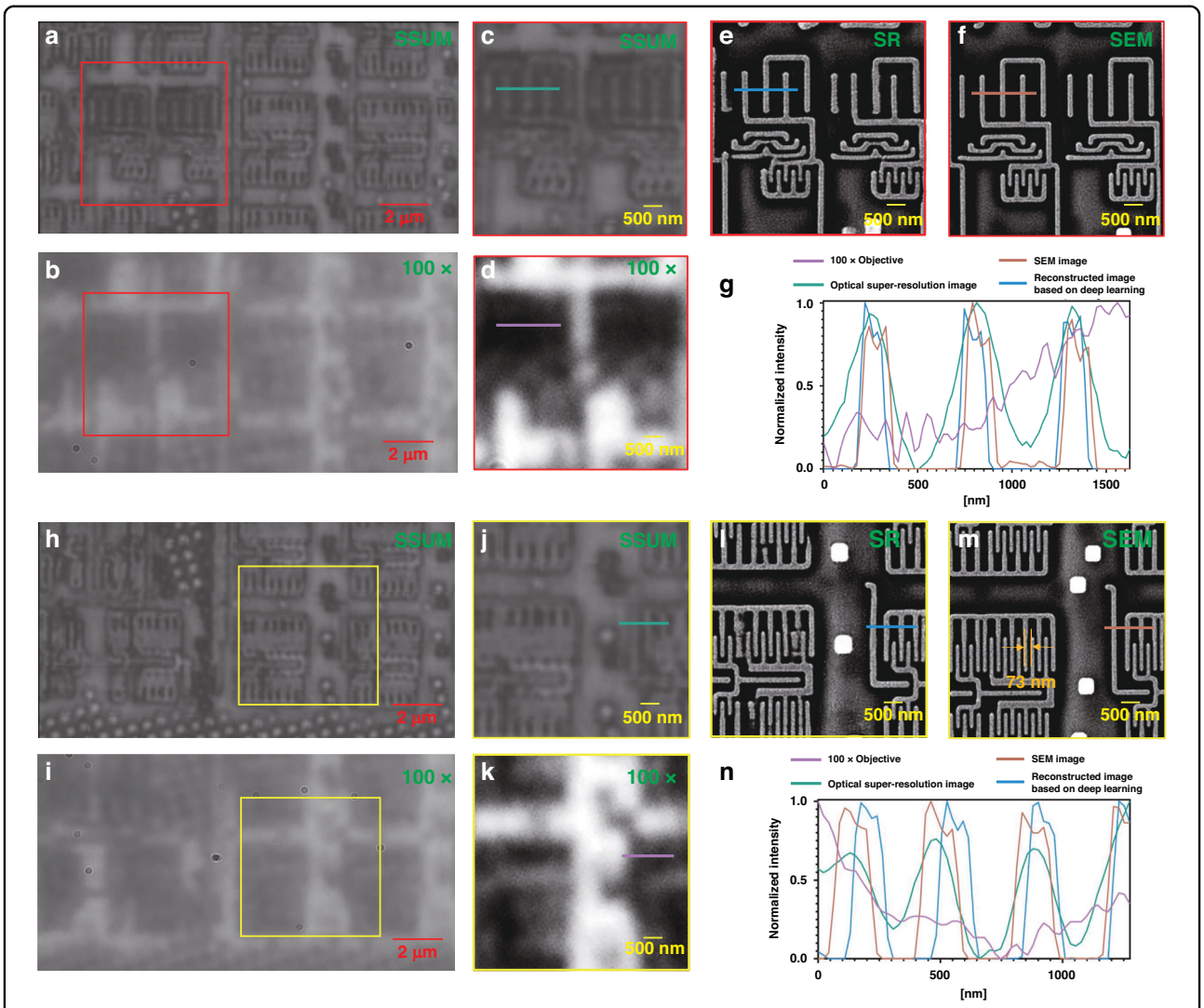

Fig. 4 Line profile analysis was used to detect the smallest features. a and **h** SSUM images of a silicon wafer. **b** and **i** Corresponding areas observed with a 100× objective lens, aligned with (**a**) and (**h**), respectively. Zoomed-in views of the red rectangular area (Location A) are shown in (**c**)–(**f**), and the yellow rectangular area (Location B) in (**j**)–(**m**). **d** and **k** 100× in-air imaging; **c** and **j** Optical super-resolution imaging with a microsphere. **e** and **l** Super-resolution reconstruction based on deep learning; **f** and **m** SEM images. **g** and **n** Line profiles of the blue dashed lines in (**c**)–(**f**) and (**j**)–(**m**), with normalized intensity

information across various frequency bands. This alignment underscores the capability of the deep learning method to enhance the resolution and fidelity of super-resolution images, making them suitable for applications that require precise detail preservation.

Line profile analysis

A wide range of areas on the silicon wafer sample were inspected using the SSUM, as shown in Fig. 4a, h, corresponding to the same areas observed using a 100× objective, as shown in Fig. 4b, i. Two regions were then randomly selected for comparison among images acquired using SSUM, a 100×/0.90 objective lens, deep learning-based super-resolution imaging, and SEM, as shown in Fig. 4c–f and Fig. 4j–m, respectively. A line profile analysis was conducted (Fig. 4g) and (Fig. 4n) along the horizontal line passing through the "gap" between the neighboring lines. It is obvious from the results that the imaging results are least clear when using an optical microscope with a 100× objective, and the details of the object under the optical diffraction limit cannot be seen. SSUM imaging technique can break the optical diffraction limit but cannot further improve the resolution of the image. For the super-resolution reconstructions based on deep learning, the intensity profiles of the images are close to those observed from SEM images. The smallest structural unit in this silicon wafer we used is 73 nm, which can be well observed. These results show that the resolution

using microspheres exceeds that of conventional optical lenses and that combining super-resolution image reconstruction using deep learning with optical super-resolution images can achieve results that are very close to those of SEM imaging. Our proposed method can make the generated image features clearer, with less noise interference and better visual effects.

Quantification of super-resolution artifacts using NanoJ-SQUIRREL

The resolution is typically estimated by measuring the minimum resolvable distance between two adjacent structures in the image. The Fourier ring correlation (FRC)[17] method of measuring effective image resolution is a straightforward and objective approach. FRC evaluates the correlation between two images across different spatial frequencies and is a prominent approach for assessing image resolution in super-resolution and electron microscopy. To determine the resolution threshold (the spatial frequency) at which both reconstructions are consistent, FRC compares the similarity of two independent reconstructions of the same object in frequency space. Block FRC resolution mapping was used to obtain local resolution measurements within the NanoJ-SQUIRREL package. The images were then spatially segmented into equal-sized blocks, as shown in Fig. 5d, h, l, and p. An FRC analysis was performed on each block, with Fig. 5d and l subdividing the image into equal blocks of size 10 and Fig. 5h and p subdividing the image into equal blocks of size 5. Figure 5 shows the resolution and quality of the images for different types and areas, indicating the variation and distribution of image resolution across blocks with different FRC values. There may be some non-square blocks on the map; this happens when the correlation between the two images at this point is insufficient to compute the FRC value. The FRC values in these blocks are tessellated from neighboring blocks to adjust for this. In addition, other quantitative image performance indicators, i.e., PSNR[18,19] and edge preservation index (EPI)[20–22], were selected to assess the quality of the generated super-resolution images. Thirty images were randomly selected for testing, with the results shown in Supplementary Fig. S2.

Comparison of different super-resolution methods

Figure 6 presents a comparison of super-resolution images generated by three different algorithms: Super-Resolution Generative Adversarial Network (SRGAN)[23], Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN)[24], and Pix2Pix[25], in addition to the CycleGAN method. Figure 6a shows the SEM image, while Fig. 6b presents the corresponding SSUM image. Selected regions from Fig. 6a and b are depicted in Fig. 6h and c, respectively, where various super-resolution reconstruction methods were applied. The resulting images generated by ESRGAN, SRGAN, Pix2Pix, and CycleGAN are displayed in Fig. 6d–g, respectively. The low-resolution input images in the original algorithm were generated by downsampling the ground truth images, and the high-resolution images and low-resolution images were replaced by SEM images and SSUM images, respectively, as the training model input. Because SRGAN and ESRGAN are both SISR algorithms, learning the image-to-image mapping relationship between low-resolution and high-resolution images is difficult, resulting in unsatisfactory image generation. Although Pix2Pix is a reference-based super-resolution algorithm, it overemphasizes the image-to-image mapping relationship between image pairs, and because the SEM images and optical super-resolution images are matched manually during the dataset production, some image-matching errors are inevitably generated and gradually accumulate over a series of image cropping and other operations, thus limiting the effect of the algorithm. Therefore, the CycleGAN algorithm, which has relatively lenient image-matching requirements, was adopted to facilitate the conversion of different types of images, enabling the generation of SEM super-resolution images. From the results in Fig. 6, the super-resolution images obtained using the CycleGAN algorithm are closest to the real images.

Super-resolution reconstruction of images with a larger field of view

To obtain super-resolution images with a larger field of view, deep learning-based super-resolution image reconstruction was performed on optical images of silicon wafers with an actual physical size of 18.5 μm × 18.5 μm. The results shown in Fig. 6i–k indicate that using the image super-resolution method in this paper, the SEM super-resolution images can be transformed very well, not only for optical images with a small field of view but also for those with a large field of view. The method combining SSUM-CycleGAN expands our horizon in the super-resolution imaging domain, enabling us not only to observe details within a confined area with precision but also to acquire high-resolution image information across broader physical scales.

Discussion and conclusion

In this paper, critical challenges in nanoscale imaging are addressed by integrating microsphere-assisted optical microscopy with deep learning to achieve super-resolution imaging resembling SEM. Our approach effectively overcomes the limitations of conventional SEM, such as high costs, vacuum requirements, and restrictive sample preparation, while enabling imaging of sub-diffraction-limit features. Building upon our custom-

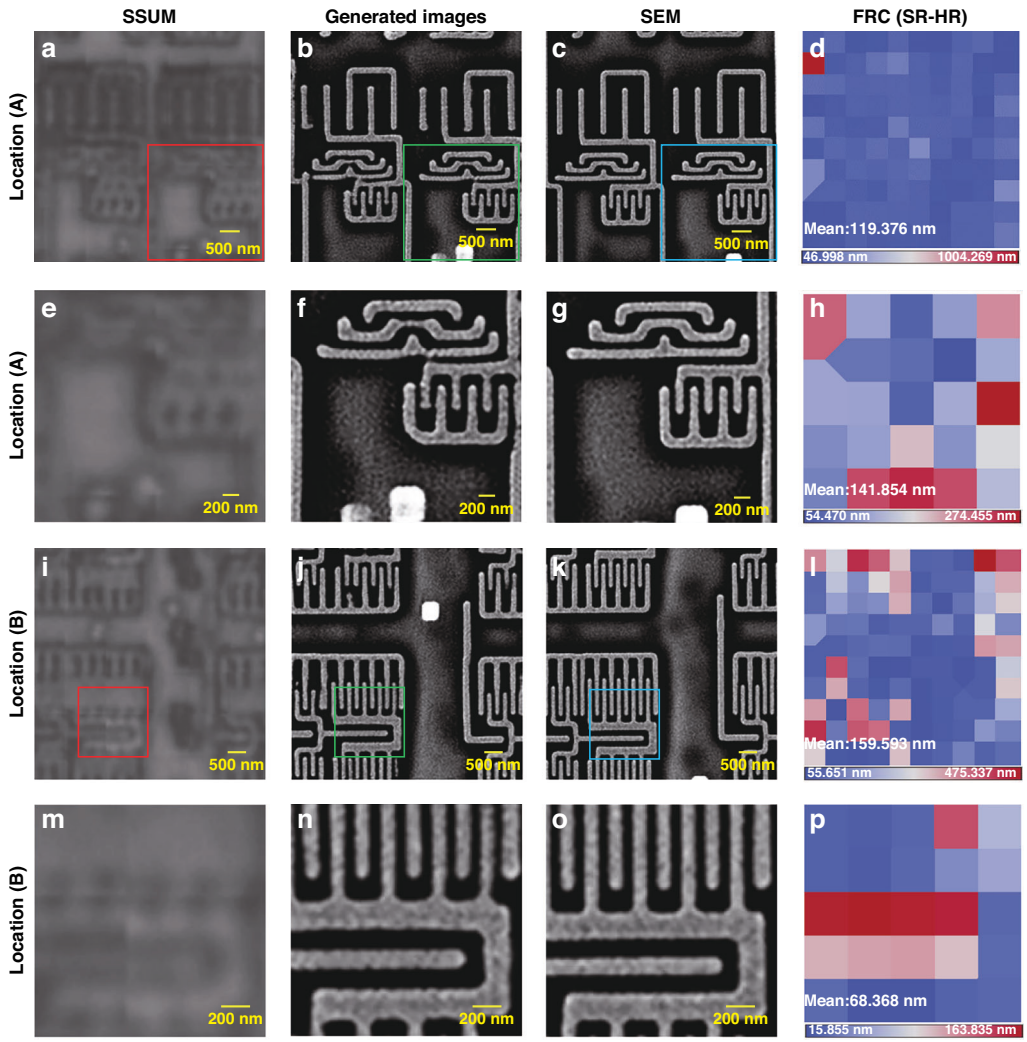

Fig. 5 Image super-resolution imaging results based on CycleGAN with multi-residual blocks and resolution mapping with NanoJ-SQUIRREL. a and **i** SSUM imaging; **b** and **j** Super-resolution imaging results generated by CycleGAN with multi-residual blocks. **c** and **k** SEM imaging (reference); **e**–**g** Zoomed-in views corresponding to regions labeled in (**a**)–(**c**). **m**–**o** Zoomed-in views corresponding to regions labeled in (**i**)–(**k**). **d**, **h**, **l**, and **p** SQUIRREL resolution maps comparing the super-resolution images with the SEM reference images. Scale bars: 500 nm for (**a**)–(**c**) and (**i**)–(**k**); 200 nm for (**e**)–(**g**) and (**m**)–(**o**)

designed SSUM optical super-resolution (OSR) platform, our algorithm further refines the nanoscale microstructures captured by SSUM. Specifically, it enhances structural clarity in regions where the SSUM platform alone struggles to resolve fine details, enabling a more complete and detailed representation of the observed samples.

By leveraging a CycleGAN model with multi-residual blocks, combined with robust data augmentation and regularization techniques, our method achieves a PSNR improvement of approximately 1.64 dB over input optical images. Quantitative and qualitative analyses demonstrate that our super-resolution images closely match SEM images in both grayscale distribution and frequency components, capturing nanoscale details with high fidelity. This achievement not only enhances image quality but also reduces dependency on specialized SEM equipment.

Beyond chip-level applications, the potential of this method extends to a broader nanotechnology context, such as nanomaterials characterization and inspection. Our deep learning framework paves the way for cost-effective, high-throughput imaging solutions that maintain SEM-level resolution without requiring a vacuum environment or complex sample preparation. Moreover, the ability to process large-field-of-view images with consistent super-resolution reconstruction demonstrates the versatility and scalability of this technique for diverse nanoscale applications.

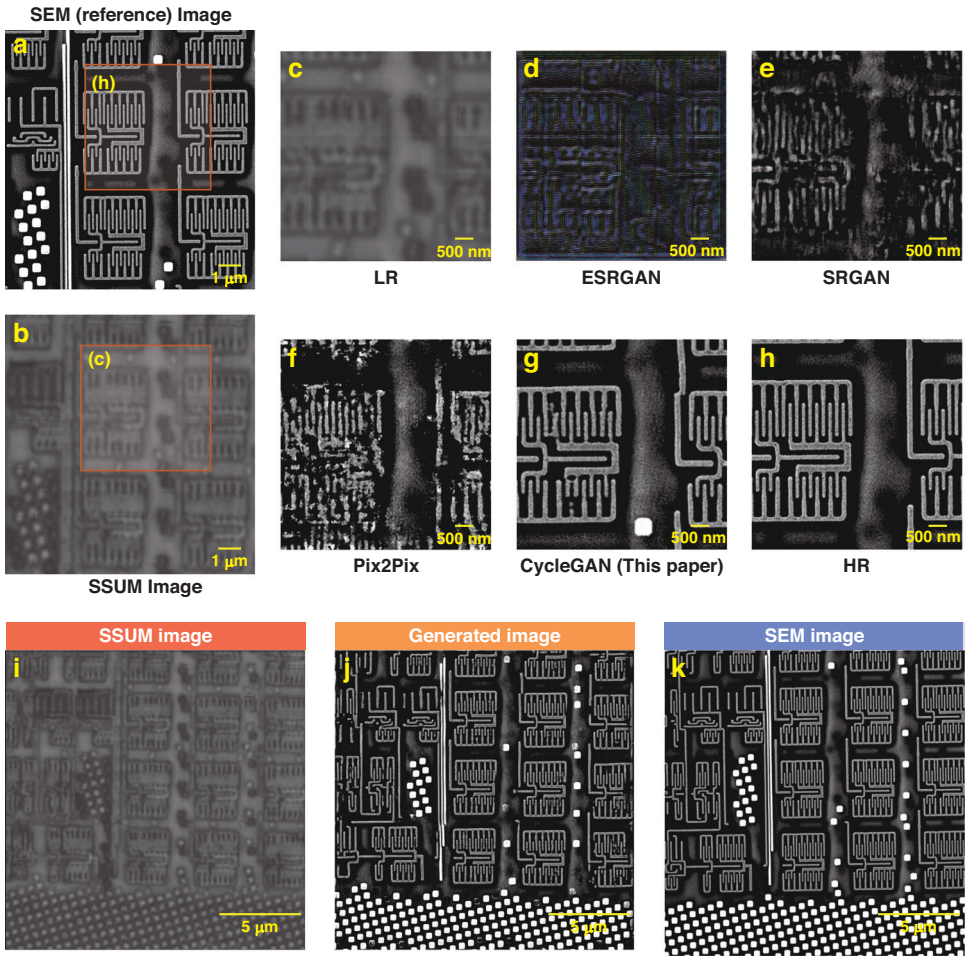

Fig. 6 Comparison of super-resolution methods and the reconstruction of a super-resolution image with a large field of view. a SEM image. **b** SSUM image. **h** and **c** Selected regions from (**a**) and (**b**), respectively, where various super-resolution reconstruction methods were applied. The super-resolution techniques based on ESRGAN, SRGAN, Pix2Pix, and CycleGAN are shown in (**d**)–(**g**), respectively (Scale bar: 500 nm). **i** Image to be reconstructed. **j** Large-field-of-view super-resolution reconstruction based on CycleGAN. **k** Reference image (Scale bar: 5 μm)

Recognizing the current limitations, such as dataset diversity and generalization across sample types, we propose future directions including expanding the dataset scope, improving the alignment between image pairs, and exploring applications in biomedical nanotechnology and advanced manufacturing. Such efforts will further solidify the method's relevance and transformative potential in the nanotechnology domain.

Future studies will specifically focus on evaluating the model's applicability to heterogeneous, multi-material chip structures that integrate both metallic and dielectric components. These investigations will be conducted under a broader range of imaging conditions to assess the model's robustness, generalization capabilities, and adaptability across varying material contrasts and structural complexities. Such efforts are expected to extend the model's practical relevance and promote its broader adoption in diverse nanofabrication and characterization applications.

In summary, this work exemplifies the synergistic application of nanotechnology and deep learning, offering a practical, scalable, and high-resolution imaging methodology. The presented approach represents a significant advancement in nanoscale imaging, bridging the gap between optical microscopy and SEM, and opening new avenues for innovation in nanoscience and nanotechnology.

Methods

Description of microsphere-based SSUM utilizing deep learning

In microsphere-based SSUM, the diameter of microspheres determines their focal length, directly influencing the system's magnification. Smaller microspheres yield higher magnification and more detailed imaging, though this enhancement is fundamentally constrained by the diffraction limit. Differences in refractive indices between microspheres and the surrounding medium can cause

image aberrations, significantly affecting resolution, especially when these differences are substantial. Therefore, when optimizing the design of super-resolution imaging systems, it is crucial to consider these factors comprehensively. By understanding and carefully manipulating these parameters, researchers can effectively enhance the performance of microsphere-based SSUM.

The well-known Rayleigh criterion unequivocally indicates that resolution is fundamentally constrained by both wavelength and numerical aperture[26], and these limits are insurmountable. Conventional methods to enhance microscope resolution typically involve employing light sources with shorter wavelengths, increasing numerical aperture, utilizing media with refractive indices matched to the system, and correcting optical system aberrations. Hence, microspheres with higher refractive indices have been used to improve image quality beyond the diffraction limit. For example, when a microsphere comes in contact with a specimen, the solid-immersion concept is employed to approximate the resolution of diffraction-limited imaging based on microspheres, yielding approximately $\lambda/(2\,n_s)$, where n_s represents the refractive index of the microsphere[27,28]. For instance, when n_s equals 1.9 (as observed in barium titanate glass microspheres), $\lambda/(2\,n_s)$ equals $\lambda/3.8$. Consequently, optical super-resolution based on microspheres is defined as a resolution surpassing $\lambda/3.8$. Nevertheless, the physical factors and conditions present formidable challenges, hampering significant strides in imaging resolution. Additionally, factors such as microsphere diameter, non-uniform incident light distribution, and discrepancies in focusing markedly influence the eventual image quality. For example, the diameter of microspheres affects their focal length, which in turn influences the overall magnification of the system. Decreasing the microsphere diameter increases the magnification, allowing for more detailed imaging. Moreover, refractive index differences between the microsphere and the surrounding medium can introduce aberrations in the images formed by the spherical lens. These aberrations can significantly impact resolution, particularly when there is a substantial difference in refractive indices. Therefore, when optimizing the design of super-resolution imaging systems, it is crucial to consider several factors comprehensively. By understanding and carefully manipulating these parameters, researchers can effectively enhance the performance of microsphere-based SSUM.

In this work, deep learning methods are used to transform microsphere-based super-resolution images into SEM large depth-of-field images. Deep learning leverages the inherently powerful fitting capability of neural networks and advanced feature learning techniques to generate high-resolution images based on lower-resolution inputs. By training Generative Adversarial Networks (GANs), this approach comprehensively interprets complex image features, thus recovering finer details from blurry and low-resolution images, as shown in Eq. 1.

$$\min_G \max_D V(D,G) = \mathbb{E}_{I^{HR}\sim P_{data}(I^{HR})}\left[\log D\left(I^{HR}\right)\right] + \mathbb{E}_{I^{LR}\sim P_{data}(I^{LR})}\left[\log\left(1-D\left(G\left(I^{LR}\right)\right)\right)\right] \tag{1}$$

The entire formulation consists of two components. I^{HR} represents the SEM super-resolution image, and I^{LR} represents the optical super-resolution image input to the G network. Also, $G(I^{LR})$ represents the SEM super-resolution image generated by the G network, and $D(I^{HR})$ represents the probability assigned by the D network to ascertain the authenticity of the SEM image (since I^{HR} is a genuine SEM image, for D, the closer this value is to 1, the better). On the other hand, $D(G(I^{LR}))$ is the probability assigned by the D network to determine whether the image generated by G is an authentic SEM image. Therefore, during the training process, to enhance the generation capability of the generator G and the discrimination capability of the discriminator D, the objective is to find the minimum value for G and the maximum value for D.

To transform optical super-resolution images into SEM images, CycleGAN is built upon the GAN model to effectively address the domain adaptation challenges between two distinct image domains, as illustrated in Eq. 2 below.

$$l_{cyc}\left(G_{LR-HR}, G_{HR-LR}, I^{LR}, I^{HR}\right) = \mathbb{E}_{I^{LR}\sim P_{data}(I^{LR})}\left[\left|\left|G_{HR-LR}\left(G_{LR-HR}\left(I^{LR}\right)\right) - I^{LR}\right|\right|_1\right] + \mathbb{E}_{I^{HR}\sim P_{data}(I^{HR})}\left[\left|\left|G_{LR-HR}\left(G_{HR-LR}\left(I^{HR}\right)\right) - I^{HR}\right|\right|_1\right] \tag{2}$$

Here, $G_{LR\text{-}HR}$ and $G_{HR\text{-}LR}$ respectively represent the processes of mapping optical super-resolution images I^{LR} to SEM images I^{HR} and the reverse transformation. In general, optical microscopy images and SEM images typically exhibit distinct appearances, textures, and resolutions, and the relationship between them may be complex and nonlinear. CycleGAN is designed to handle non-paired data, meaning there is no direct correspondence between the two domains. This allows the model to learn to map images from one domain to another, even in the absence of paired training data. Overall, CycleGAN provides an effective framework for learning complex mappings between optical microscopy and SEM images on paired datasets, enabling cross-domain image transformation. Thus, deep learning-based super-resolution imaging techniques using CycleGAN not only address the limitations of traditional optical microscopes in resolving

details smaller than half the wavelength due to optical diffraction but also overcome the challenges encountered by microsphere-based super-resolution imaging systems. These challenges arise from factors such as microsphere diameter and light source wavelength, fundamentally limiting imaging resolution.

Principle and design of the optical super-resolution system

Based on our previous work[11,29], a non-invasive, high-throughput, environmentally compatible optical SSUM system was developed, and the composition and performance characteristics of the systems are shown in the Supplementary. Figure S1 shows the photograph of the physical configuration of the SSUM, which can be used for large-area, super-resolution imaging and data acquisition. The optical super-resolution imaging system consists of an AFM, a commercially available cantilever (TESP probe, Bruker), a 57-μm-diameter BTG microsphere lens (Cospheric), and a 50× objective lens (Nikon LU Plan EPI ELWD). Microspheres are attached to the AFM probe cantilever using a UV-curable adhesive (NOA63, Edmund Optics). Different illumination conditions are achieved by adjusting two stops in the Köhler illumination system (Thorlabs). A high-speed scientific complementary metal oxide semiconductor camera (PCO.Edge 5.5) is used to record the images. Illumination is provided by an intensity-controlled light source (C-HGFI, Nikon, Japan), with the peak illumination wavelength of the system set to ~550 nm for white light imaging by the optical elements. A drop of UV-cured adhesive is positioned close to the microsphere region to act as the sticky material. The selected microspheres are touched by the adhesive-coated tip, which is then subjected to UV light for 90 s until the adhesive is fully cured. The microsphere superlens is mounted to the AFM cantilever and kept in place during scanning to keep the distance between the microsphere and the objective as consistent as possible. The specimen is placed on top of the stage after the standard AFM adjustments are completed. The scanning platform and 3D translation stage are used to carry out horizontal sample scanning and vertical focus adjustment. The IC chip scanning in the horizontal direction and the feedback adjustment in the longitudinal direction are realized by a 3D piezoelectric ceramic (PZT) scanner (P-733.3CL, Physik Instrumente, Germany), in the process of which the AFM feedback is acquired to monitor whether the microspheres are touching the sample surface and to adjust the distance between the microspheres and the sample. When the distance or force reaches a certain value, the optical microscope is driven by a translation stage with nanoscale resolution (IMS100V, Newport) to capture a virtual image generated by the microsphere. The interval of the IC chip scan and the interval of the signal used to trigger the camera image recording are both adjusted according to the region of the field of view of the microsphere superlens during horizontal scanning. Before scanning, the camera's captured area can be modified to satisfy the overlapping requirement for image stitching, and there is no considerable aberration in the field of view of the microsphere superlens, which effectively reduces data processing time and allows for fast image processing before or after scanning.

SEM image acquisition

The HR images are captured using a QuantaTM 450 FEG-SEM (field-emission gun scanning electron microscope; SEI) with a beam voltage of 30 KeV and a spot size of 4.0.

Image pre-processing

To enable the model to better learn the mapping relationship between SSUM images and SEM images, image processing of the captured SSUM images and SEM images is necessary. To gather the SSUM (low-resolution) and SEM (high-resolution) image pairs of the same samples for network training, it is first necessary to stitch together each frame image taken by the SSUM to obtain an optical super-resolution image of the wafer. After that, the SSUM image is registered with the SEM image, matching the fields of view of the LR and HR images so that each training pair shows roughly the same region of the sample, and cropping them to 512 × 512 pixels. If the effective pixel sizes of the LR and HR images differ (e.g., due to being captured with different devices such as the SSUM system and SEM), they are rescaled to the same physical size. The cropped image pairs are then sorted into two groups for use as training and test datasets, with the training dataset being fed into the neural network for model training. In terms of network structure design and loss function, a network model is built by combining convolutional neural networks and multiple residual blocks, and the loss function is designed based on prior knowledge; model training is then performed to determine the optimizer and learning parameters, and the network parameters are updated using a backpropagation algorithm to improve the model's learning ability by minimizing the loss function. The network model is evaluated according to the performance of the trained model on the validation set, and corresponding adjustments are made. Finally, the trained generative model is tested with the test dataset images to evaluate the tested image quality (see Supplementary Fig. S2).

Design of CycleGAN with multi-residual block network architecture

CycleGAN is a method for training deep convolutional neural networks for image-to-image translation

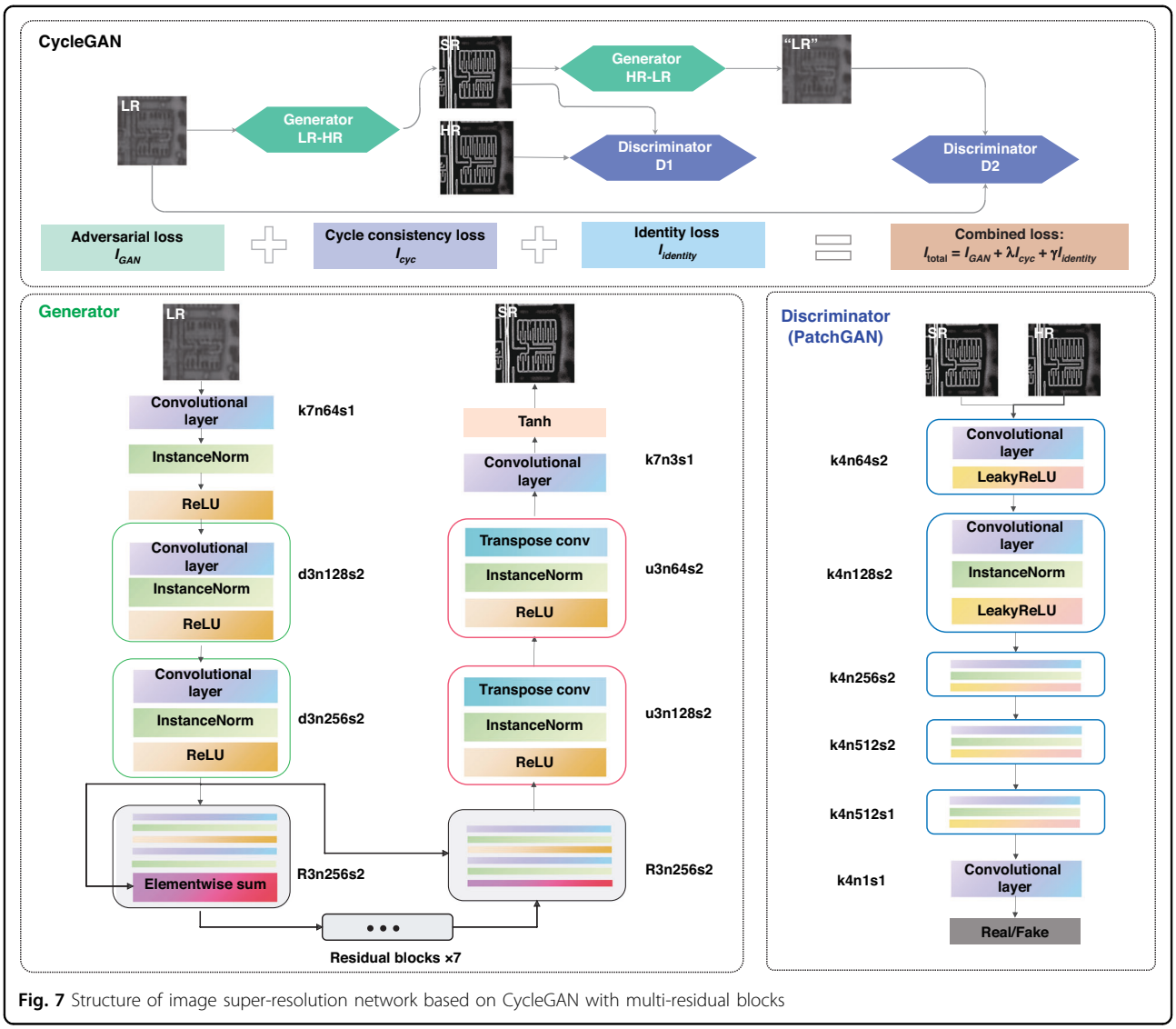

Fig. 7 Structure of image super-resolution network based on CycleGAN with multi-residual blocks

tasks, with the network structure and loss function shown in Supplementary Fig. S3. The basic goal of this translation is to use a network to learn a mapping between input and output images. During the training phase, cycle consistency is needed: that is, one image identical to the original should be produced with the lowest L1 loss value following repeated application of two different generators. Supplementary Fig. S4 shows the variation in loss for the generator and discriminator in the image super-resolution model, after numerous iterations.

Network architecture

The network structure of the CycleGAN with multi-residual blocks is shown in Fig. 7. Its main principle is to train the generator and discriminator models to convert images from low to high resolution with cyclic consistency, such that the generated super-resolution image should subsequently be back-converted to the LR image. The generator is designed to learn how to transform features of SSUM images into features resembling those of real images, while the discriminator is trained to differentiate between real and synthetic images. For the discriminator networks, 70 × 70 PatchGANs are utilized, which focus on classifying each patch in the image as real or fake. A patch-level discriminator architecture has fewer parameters than a full image discriminator and can be applied to arbitrarily sized images in a fully convolutional fashion. For each training batch, the generator and the discriminator compete against each other so that the generator learns to produce features sufficiently similar to the real image to fool the discriminator, bringing the generated super-resolution images closer to the real high-resolution images.

Generator architecture

The generator comprises three primary modules: the downsampling module, the residual module, and the upsampling module. In the training process, the SSUM image is initially downsampled to an LR image. Subsequently, the generator architecture attempts to upsample the LR image to achieve super-resolution, and shallow features are extracted using convolution. The residual blocks are then used to extract the deeper features. The generator architecture then attempts to upsample the image from low resolution to super-resolution using the upsampling module. After this, the image is passed into the discriminator, which tries to distinguish between the SEM image and the generated super-resolution image and generates the adversarial loss for backpropagation into the generator architecture.

In the network structure diagram in Fig. 7, k, d, R, and u represent the convolution kernel size of the convolution layer, downsampling layer, residual block, and upsampling layer, respectively. To extract residual characteristics from both LR and HR images, a deep learning network structure is built with nine residual blocks for 256×256 resolution training images. The generator architecture contains residual networks instead of deep convolutional networks because residual networks are easier to train and can thus be substantially deeper to generate better results. The benefit arises from the residual network using a type of connection called skip connections. There are nine residual blocks, generated by ResNet. Within each residual block, two convolutional layers are used with small 3×3 kernels and 64 feature maps, followed by batch normalization layers and LeakyReLU as the activation function. The resolution of the input image is increased using two trained sub-pixel convolution layers. The above-mentioned network structure can adaptively learn the parameters of the rectifier and improve the accuracy at negligible extra computational cost.

Discriminator architecture

The task of the discriminator is to discriminate between real HR images and generated super-resolution images. The discriminator architecture used in this paper is a 70×70 PatchGAN architecture with LeakyReLU as the activation function. The structure also uses instance normalization. The network contains five convolutional layers. With the deepening of the network layers, the number of features increases, and the feature size decreases. The first four layers have 4×4 filter kernels, increasing by a factor of 2 from 64 to 512 kernels with stride 2, which is the convolutional-InstanceNorm-LeakyReLU structure. The last layers of the discriminator use filter kernels of size 512 with stride 1.

Datasets

The registered SSUM images and SEM images are input into the neural network as LR/HR image pairs and cropped to 512×512 pixels. To improve training efficiency, the input image size is resized to 256×256 pixels. The training set has 704 image pairs. All the network output images shown in this paper were blindly generated by the deep network; that is, the input images had not previously been seen by the network. However, if such training image pairs are not available when using our super-resolution image transformation framework, an existing trained model could be used, although this might not produce ideal results in all cases.

Strategy to avoid overfitting

To address the challenge of a model performing well on the training set but struggling to generalize effectively to cross-validation or unseen data, several strategies were implemented to enhance the model's adaptability while avoiding overfitting. The primary approach involves data augmentation, which increases the diversity of the training dataset, thereby improving the model's ability to generalize to unknown samples. Specifically, the original images are augmented by flipping them horizontally and vertically, followed by rotations of 30°, 45°, 60°, and 90°. Additionally, the images are scaled to various sizes, further enriching the dataset and helping the model learn more robust features across different spatial scales.

To complement data augmentation, regularization techniques such as dropout are integrated into the network design. Dropout is applied during the training process to mitigate the risk of overfitting by reducing the co-adaptation of neurons. By randomly deactivating a subset of neural nodes during training, dropout forces the network to become less reliant on specific features, thereby improving its robustness and preventing the model from becoming overly dependent on local patterns. This is particularly important in scenarios where the dataset is limited in size and diversity, as it helps the model maintain generalization across different samples.

Furthermore, the design of the model itself plays a critical role in enhancing generalizability. Our approach utilizes a CycleGAN architecture with multiple residual blocks, which not only streamlines feature extraction but also supports the model in capturing complex patterns within small datasets. The residual connections help mitigate the vanishing gradient problem, enabling the model to converge more effectively while retaining important details. This design choice is especially beneficial in small-data scenarios, as it allows the model to extract and preserve meaningful features that generalize well to diverse sample types. Super-resolution model robustness validation tests, as shown in Fig. S5, further demonstrate the effectiveness of this approach.

Additionally, different generator network configurations, such as the number of residual blocks and dropout settings, have been explored, which influence the results, as shown in Fig. S6.

Strategies to prevent artifact generation

The data augmentation techniques discussed above also play a crucial role in reducing artifacts, as they help the model learn more robust mapping relationships between image pairs. By training the model on a diverse set of augmented images, the likelihood of generating artifacts during inference is minimized. However, noise in the optical super-resolution images can still pose challenges for accurate image-to-image mapping. To address this, noise reduction is performed, and image edges and contrast are enhanced during the pre-processing phase to ensure that the input data is of high quality. These steps are vital for producing sharper and clearer SEM images.

In addition to pre-processing, the design of the GAN model, which is based on a deep residual network, contributes significantly to artifact suppression. The model leverages multiple residual blocks, skip connections, and block regularization to produce stable and high-quality outputs. The generator network focuses on creating noise-free images, while the discriminator network, trained on these enhanced images, becomes adept at distinguishing real from generated images. This dual approach not only improves image fidelity but also effectively suppresses the formation of artifacts.

To further validate the spatial consistency of the generated images and assess residual discrepancies, we conducted a spatial error analysis of super-resolution reconstruction via pixel-wise difference mapping, as shown in Fig. S7. This analysis reveals that the most significant reconstruction errors occur near sharp edges and high-frequency regions, providing valuable insights into potential sources of visual artifacts.

Given that the image pairings in this study are manually aligned, alignment errors could introduce inaccuracies. As such, our loss function does not include traditional image evaluation metrics like PSNR or EPI, which are sensitive to global alignment issues. Instead, a specialized loss function from ref. [30] is adopted, which is better suited for handling misalignments in image-to-image translation. For future work, if precise alignment can be achieved, incorporating PSNR or EPI as part of the loss function could further enhance the quality of the super-resolution results by guiding the network toward more accurate reconstructions.

Acknowledgements

This work was supported by the Hong Kong Research Grants Council (Project No. 11216120) and the Science, Technology, and Innovation Commission of Shenzhen Municipality (Grant Nos. SGDX2019081623121725 and JCYJ20190808181803703).

Author details

[1]Department of Mechanical Engineering, City University of Hong Kong, Hong Kong SAR, China. [2]City University of Hong Kong Shenzhen Research Institute (CityUSRI), Shenzhen 518000, China. [3]State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China. [4]Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169, China. [5]Department of Electrical and Electronics Engineering, The University of Hong Kong, Hong Kong SAR, China. [6]Department of Computer Science, City University of Hong Kong, Hong Kong SAR, China

Author contributions

W.J.L., G.Z., and L.L. supervised and guided the project. W.J.L. and G.Z. conceived the initial concept of this work. M.C. and X.W. carried out the initial background review. W.J.L., H.S., H.L., H.Y., and G.Z. conceived the experimental design. F.W., W.J.L., and L.L. conceived the design of the SSUM system. H.L., X.W., and H.Y. used the SSUM system to acquire the SSUM images used in this work. H.L., Q.C., M.C., and X.W. helped with data collection. H.S. and G.Z. performed the experiments and analyzed the data with help from M.C., Q.C., X.W., and F.W. H.S. and G.Z. developed the deep learning algorithms with input from J.W. and D.W. H.S., H.L., G.Z., and W.J.L. co-drafted the manuscript, with input from F.W., J.W., and D.W. H.Y., J.W., D.W., L.L., G.Z., and W.J.L. provided critical technical advice and comments for the experimental studies and manuscript revision.

Conflict of interest

The authors declare no competing interests.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41378-025-01060-9.

Received: 2 April 2025 Revised: 26 July 2025 Accepted: 11 August 2025
Published online: 19 December 2025

References

1. Zhao, W. et al. Sparse deconvolution improves the resolution of live-cell super-resolution fluorescence microscopy. *Nat. Biotechnol.* **40**, 606–617 (2022).
2. Okamoto, H., Kamada, S., Yamaguchi, K., Haraguchi, M. & Okamoto, T. Experimental confirmation of self-imaging effect between guided light and surface plasmon polaritons in hybrid plasmonic waveguides. *Sci. Rep.* **12**, 17943 (2022).
3. Zhanghao, K. et al. Super-resolution imaging of fluorescent dipoles via polarized structured illumination microscopy. *Nat. Commun.* **10**, 4694 (2019).
4. Kwon, S., Park, J., Kim, K., Cho, Y. & Lee, M. Microsphere-assisted, nanospot, non-destructive metrology for semiconductor devices. *Light Sci. Appl.* **11**, 32 (2022).
5. Hell, S. W. & Wichmann, J. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Opt. Lett.* **19**, 780–782 (1994).
6. Gustafsson, M. G. L. Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy. *J. Microsc.* **198**, 82–87 (2000).
7. Betzig, E. et al. Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313**, 1642–1645 (2006).
8. Rust, M. J., Bates, M. & Zhuang, X. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3**, 793–796 (2006).
9. Dertinger, T., Colyer, R., Iyer, G., Weiss, S. & Enderlein, J. Fast, background-free, 3D super-resolution optical fluctuation imaging (SOFI). *Proc. Natl. Acad. Sci. USA* **106**, 22287–22292 (2009).
10. Wang, Z. et al. Optical virtual imaging at 50 nm lateral resolution with a white-light nanoscope. *Nat. Commun.* **2**, 218 (2011).
11. Wang, F. et al. Scanning superlens microscopy for non-invasive large field-of-view visible light nanoscale imaging. *Nat. Commun.* **7**, 13748–13758 (2016).
12. Wang, H. et al. Deep learning enables cross-modality super-resolution in fluorescence microscopy. *Nat. Methods* **16**, 103–110 (2019).
13. Qiao, C. et al. Evaluation and development of deep neural networks for image super-resolution in optical microscopy. *Nat. Methods* **18**, 194–202 (2021).

14. Nehme, E., Weiss, L. E., Michaeli, T. & Shechtman, Y. Deep-STORM: super-resolution single-molecule microscopy by deep learning. *Optica* **5**, 458–464 (2018).
15. Qiao, C. et al. Rationalized deep learning super-resolution microscopy for sustained live imaging of rapid subcellular processes. *Nat. Biotechnol.* **41**, 367–377 (2023).
16. Zhang, H. et al. High-throughput, high-resolution deep learning microscopy based on registration-free generative adversarial network. *Biomed. Opt. Express* **10**, 1044–1063 (2019).
17. Nieuwenhuizen, R. P. J. et al. Measuring image resolution in optical nanoscopy. *Nat. Methods* **10**, 557–562 (2013).
18. Hore, A. & Ziou, D. Image quality metrics: PSNR vs. SSIM. In *2010 20th International Conference on Pattern Recognition* 2366–2369 (IEEE, 2010).
19. Sara, U., Akter, M. & Uddin, M. S. Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. *J. Comput. Commun.* **7**, 8–18 (2019).
20. Joseph, J., Jayaraman, S., Periyasamy, R. & Renuka, S. An edge preservation index for evaluating nonlinear spatial restoration in MR images. *Curr. Med. Imaging* **13**, 58–65 (2017).
21. Martini, M. G., Hewage, C. T. & Villarini, B. Image quality assessment based on edge preservation. *Signal Process.: Image Commun.* **27**, 875–882 (2012).
22. Chen, L. et al. Edge preservation ratio for image sharpness assessment. In *2016 12th World Congress on Intelligent Control and Automation (WCICA)* 1377–1381 (IEEE, 2016).
23. Ledig, C. et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition,* pp. 4681–4690 (IEEE, 2017).
24. Wang, X. et al. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proc. European Conference on Computer Vision (ECCV) Workshops* (Springer Science + Business Media, 2018).
25. Isola, P., Efros, A. A., Ai, B. & Berkeley, U. C. Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition,* pp. 1125–1134 (IEEE, 2017).
26. Born, M. & Wolf, E. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light* (Elsevier, 2013).
27. Allen, K. W. et al. Super-resolution microscopy by movable thin-films with embedded microspheres: resolution analysis. *Ann. Phys.* **527**, 513–522 (2015).
28. Allen, K. W. et al. Overcoming the diffraction limit of imaging nanoplasmonic arrays by microspheres and microfibers. *Opt. express* **23**, 24484–24496 (2015).
29. Wang, F. et al. Three-dimensional super-resolution morphology by near-field assisted white-light interferometry. *Sci. Rep.* **6**, 24703–24713 (2016).
30. Zhu, J., Park, T., Efros, A. A., Ai, B. & Berkeley, U. C. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE International Conference on Computer Vision,* pp. 2223–2232 (IEEE Computer Society, 2017).