PhotoniX

RESEARCH Open Access

Transformer-based neural network enabled subpixel-resolution in wide-field meta-microscope

Shanshan Hu[1], Zihan Wang[1], Jian Li[1], Junyi Wang[1], Wenjing Shen[1], Jiacheng Sun[1], Jitao Ji[1], Xian Long[1], Xu Liu[2], Chen Chen[1*], Shining Zhu[1] and Tao Li[1*]

*Correspondence: chenchen2021@nju.edu.cn; taoli@nju.edu.cn

[1] National Laboratory of Solid State Microstructures, Key Laboratory of Intelligent Optical Sensing and Manipulation, Jiangsu Key Laboratory of Artificial Functional Materials, Collaborative Innovation Center of Advanced Microstructures, School of Physics, College of Engineering and Applied Sciences, Nanjing University, Nanjing 210093, China
[2] ZJU-Hangzhou Global Scientific and Technological Innovation Center, College of Optical Science and Engineering, State Key Laboratory of Extreme Photonics and Instrumentation, Zhejiang University, Hangzhou 310027, China

Abstract

The pursuit of compact microscopy systems faces dual constraints from cascaded optical elements and sensor pixel limits. While the integration of metalens and sensor eliminates the bulky elements, the resolution remains confined by pixel-induced under-sampling. Here, we propose a computational imaging framework that synergizes a compact metalens microscope with a transformer-based neural network to achieve subpixel-resolution. To bridge the simulation-to-reality gap, we construct the first experimental dataset of metalens-acquired thyroid pathological sections images. The training strategy enables rapid (~0.2s for 110 μm × 110 μm FOV), high-fidelity (structural similarity up to 91%) reconstruction from single-frame inputs, achieving 3 × spatial sampling density with a high resolution (close to the ground truth resolution of 0.87 μm). We further demonstrate its scalability by implementing the trained network in a metalens array-based system, achieving wide-field (4 mm × 6 mm) and high-resolution (close to the Olympus 10 × /0.25NA objective) imaging, with a field of view approximately 14.5 times that of the Olympus objective. The proposed framework highlights the synergy between simplified optical hardware and computational reconstruction, paving the way for compact and intelligent microscopy.

Keywords: Metalens, Deep learning, Microscope, Subpixel-resolution, Wide-field imaging

Introduction

Optical microscopy is a cornerstone of modern medical diagnostics, particularly in pathological analysis, and supports a broad range of scientific and industrial applications. Nevertheless, conventional microscopy systems confront two inherent physical constraints that constrain their performance and utility. First, the spatial bandwidth product constraints of the uniaxial imaging architecture impose a trade-off between resolution and field of view (FOV), making it challenging to account for both simultaneously. Second, the sensor-imposed resolution barrier necessitates bulky and complicated cascaded magnification systems. To satisfy Nyquist-Shannon sampling criteria [1] for the pixel size of image sensors, conventional designs generally employ multi-stage lens

configurations that dramatically extend the system size. The ongoing miniaturization of optical systems necessitates revolutionary integration strategies for microscopic imaging technologies.

In contrast to conventional refractive lenses, metalenses [2–4] enable transformative ultracompact imaging systems through subwavelength-engineered structures that precisely manipulate light's amplitude, phase, and polarization [5–9]. While notable metalens-based microscopy systems have been demonstrated [10–17], simply replacing refractive components in conventional optics fails to exploit their full compact potential. A significant strategy has demonstrated a chip-scale metalens-integrated microscope (CSM) with large FOV and depth of field by mounting a polarization multiplexed metalens array directly on a complementary metal-oxide semiconductor (CMOS) image sensor [18, 19]. This configuration (unit-magnification with object-to-image distance of 4f) enlarges the FOV by stitching images of multiple sub-lenses, while the resolution remains constrained by the pixel size of the image sensor rather than the optical limit. Current resolution enhancement approaches face critical trade-offs: increasing numerical aperture (NA) or imaging magnification expands system dimensions, degrades signal-to-noise ratio, narrows FOV, and escalates aberration correction complexity. Additionally, specialized image sensors with smaller pixel sizes require costly hardware modifications. These constraints highlight the urgent need for computational strategies that enhance CSM resolution without altering its compact hardware architecture.

Recently, deep learning has revolutionized computational imaging, demonstrating excellent capabilities in image processing tasks, including enhancement, deblurring, denoising, and super-resolution [20–22]. This progress has spurred the innovative integration of neural networks with meta-optics, establishing a high-performance computational framework that spans from intelligent design to back-end processing, bringing new ideas to improve the imaging performance of the system [23–32]. These developments open up promising avenues for achieving high-fidelity, wide-field imaging in compact metalens systems.

Here, we propose and experimentally demonstrate a synergistic computational imaging framework that integrates a compact unit-magnification metalens microscope with a transformer-based neural network, achieving significant performance improvement without hardware modifications. First, two metalenses with different NAs are integrated onto the CMOS image sensor as a dataset acquisition system for paired equal-size imaging and three-fold magnification imaging under the same effective FOV, respectively. Following imaging characterization, we establish a preprocessed experimental dataset of metalens-captured thyroid pathological sections images for transformer-based neural network training. The trained transformer model demonstrates computational sub-pixel-resolution reconstruction, producing high-fidelity, high-resolution images from a single-frame input. To further extend the FOV, we incorporate the metalens array-based integrated microscope with the trained model, adopting a scanning-based strategy by only nine object shifts. Compared to conventional whole-slide scanners that require extensive mechanical precision and thousands of scanning steps, our approach significantly reduces mechanical complexity while maintaining an ultra-compact and lightweight design. Finally, a large FOV (4 mm × 6 mm) is obtained with high resolution (showing comparable with the Olympus 10 ×/NA = 0.25 objective) and ultra-compact

size, achieving a cost-effective image resolution improvement that does not rely on hardware upgrades.

Methods

Figure 1 presents a schematic of the proposed ultra-compact metalens microscope equipped with the transformer-based neural network. The system employs a metalens array to enable wide-field imaging. Each sub-metalens achieves unit-magnification imaging by setting the object-to-image distance to 4f, thereby preventing overlap between the adjacent FOVs of the densely packed sub-lenses while maximizing the utilization of the effective imaging area. To mitigate the inherent under-sampling issue caused by the sensor's pixel size limitation, the model is developed and trained under supervised learning, using high-NA data as ground truth. This model establishes a mapping between unit-magnification imaging and three-fold magnification, enabling effective $3\times$ pixel sampling from a single image. As a result, the image resolution and quality are significantly improved without modifying the hardware, effectively mitigating the constraints imposed by the sensor's native pixel size.

Given that the metalens array comprises a dense configuration of individual lenses, we initially construct a dataset using a single metalens for the model training, which can subsequently be generalized to the whole array. Figure 2(a) illustrates the schematic of the metalens-integrated imaging system for dataset acquisition. To stabilize the conditions for image collection, we adopted a multi-layer integrated architecture with a fixed image distance. Two metalenses fabricated on the same substrate are utilized to gather high-resolution and low-resolution images, respectively. The metalenses and a fixed circular

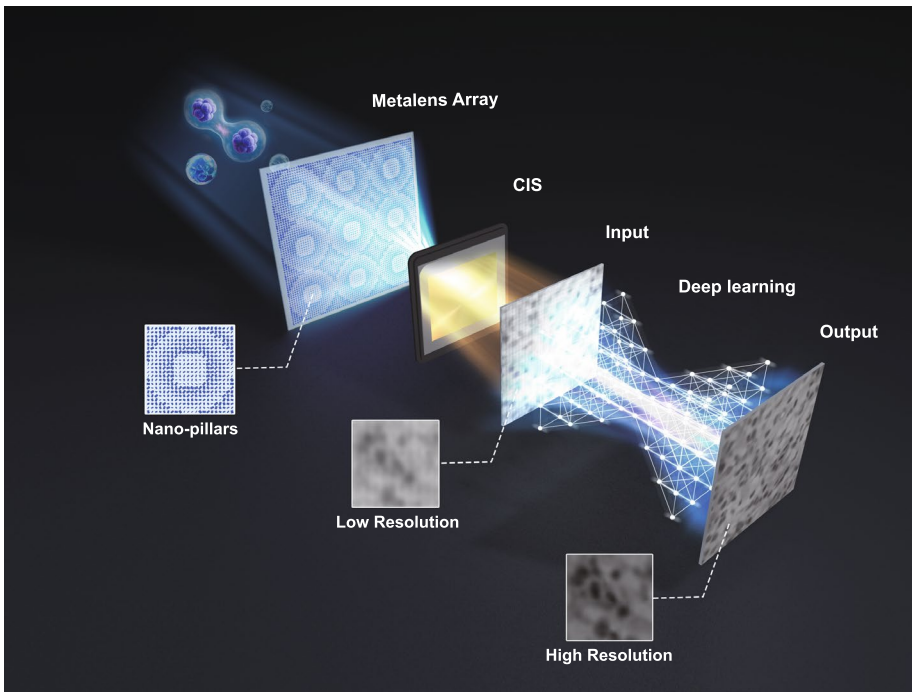

Fig. 1 Schematic diagram of the proposed ultra-compact metalens microscope equipped with the transformer-based neural network. The CIS refers to CMOS Image Sensor

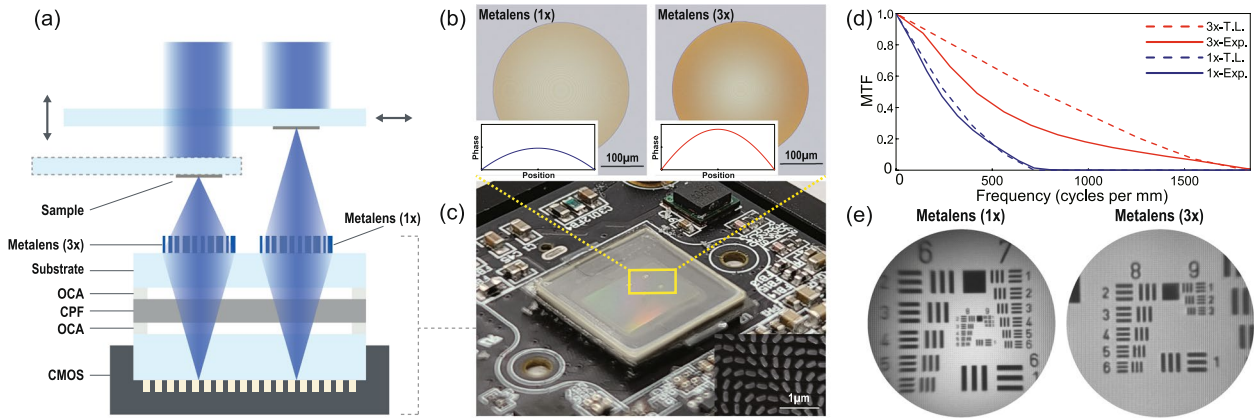

Fig. 2 The dataset acquisition device architecture and metalens characterization. **a** Schematic of the dataset acquisition imaging system. **b** Microscope images of metalenses at magnifications of 1 × and 3 ×, with bottom-left insets showing the corresponding phase profiles of the two lenses, respectively. **c** Photograph of the dataset acquisition device with bottom-right inset for the SEM image of the SiNx metalens. **d** The experimental and theoretically limited MTF of the metalens (1 ×) and the metalens (3 ×). **e** Imaging results of the USAF resolution chart taken from the metalens (1 ×) and the metalens (3 ×)

polarization filter (CPF) with a thickness of 220 μm are mounted onto the CMOS image sensor (Imaging Source: DMM27UJ003-ML, pixel size: 1.67 μm × 1.67 μm) using two layers of optically clear adhesive (OCA) tapes (Tesa 69402; thicknesses: 50 μm). Among them, the CPF can filter the co-polarized background light and significantly improve the imaging quality. The OCA serves both as a mechanical connector and as a spacer to adjust the proper imaging distance (v). The photograph of the acquisition device is shown in Fig. 2(c), where effective v equals 1000 μm when considering the influence of the spacer medium.

The designed metalens is based on the Pancharatnam-Berry (PB) Phase, with phase profile following the hyperbolic distribution

$$\varphi = \frac{2\pi}{\lambda}(f - \sqrt{x^2 + y^2 + f^2}), \tag{1}$$

where λ is the working wavelength, f is the focal length, and x and y are the coordinates of each nano-pillar. The metasurface consists of nano-pillars with a uniform high-aspect-ratio rectangular geometry (height: 1 μm, length: 240 nm, width: 90 nm), arranged with a 300 nm period and exhibiting a 99% polarization conversion ratio at $\lambda = 470$ nm. We fabricated two metalenses with the same diameter of 320 μm while different focal lengths of 500 μm and 250 μm, corresponding to imaging magnifications of 1 × and 3 × at the same fixed image distance. We refer to them as metalens (1 ×) and metalens (3 ×) in the following, respectively, for convenience. All the metalenses were fabricated on a 1-μm-thick SiNx film deposited on a 500-μm-thick fused silica substrate using standard electron-beam lithography and dry etching. Figure 2(b) and the inset of Fig. 2(c) show the optical microscope image and scanning electron microscope (SEM) image of the two metalenses, respectively.

To quantitatively evaluate the imaging performance with these two metalenses, the modulation transfer function (MTF), resolution, and effective FOV were characterized. MTFs were calculated from experimentally measured point spread functions. As shown in Fig. 2(d), both lenses exhibit good agreement with the theoretical diffraction-limited MTFs, with the metalens (3 ×) showing a slight decline at higher spatial frequencies. The reduction mainly arises from the residual spherical aberration when the hyperbolic-phase profile, which ideally focuses light from an object at infinity, is used for finite-distance imaging. In addition, minor defocus during experimental MTF measurement may also contribute to the small deviation from the theoretical diffraction limit. These factors result in a slight background blur that marginally reduces high-frequency modulation without visibly affecting image contrast or feature resolution. The resolution of the two metalenses was evaluated using a positive 1951 United States Air Force (USAF) resolution test chart placed at object distances corresponding to unit and three-fold magnification under incoherent illumination. The metalens (1 ×) demonstrated a half-pitch resolution of 1.74 μm (element 2, group 8), limited by the CMOS pixel size (1.67 μm), while the metalens (3 ×) demonstrated a half-pitch resolution of 0.87 μm (element 2, group 9), as depicted in Fig. 2(e). The resolution and magnification of the two metalenses meet the dataset acquisition conditions. To ensure that the ground truth data within the FOV has sufficient resolution and minimal optical aberration or distortion,

it is necessary to select and crop the imaging region after 3 × magnification. We used a grid distortion target (Grid space: 10 μm, Thorlabs, R1L3S3P) as the imaging target to calibrate the effective FOV of the lenses, and the result of the metalens (3 ×) was used as a reference for selecting the optimal imaging area in the subsequent dataset (refer to Supplementary Note 1). Additional field-dependent resolution analyses, including slanted-edge MTF measurements and imaging validation across the effective FOV, are also presented in Supplementary Note 1.

We obtained the dataset of image pairs from stained thyroid pathological sections. Low-resolution images were captured with the metalens (1 ×) at an object distance of 1000 μm, while high-resolution images were obtained using the metalens (3 ×) at a reduced object distance of 333 μm. By moving the thyroid slices both vertically and horizontally, we can flexibly adjust the object distance and imaging area (Fig. 2(a)), producing unit-magnification and three-fold magnification images corresponding with the low- and high-resolution raw datasets. A total of 3000 sets of raw image pairs were collected, corresponding to various regions from different slices. Thanks to the combined acquisition device, the imaging positions of both metalenses remain fixed on the image sensor, allowing for subsequent batch processing of the images. We extracted a 198 × 198-pixel region from the effective FOV of the metalens (3 ×) as the ground truth data, which corresponds to a physical area of 110 μm × 110 μm. The associated low-resolution input images measured 66 × 66 pixels. For training the end-to-end network, precise alignment of the low-resolution input image and the high-resolution ground truth data is crucial. Through a two-stage image registration process, we created 3000 pairs of low- and high-resolution image patches with subpixel alignment (refer to Supplementary Note 2). From these, we randomly selected 2400 image pairs for the training set, 300 pairs for the validation set, and 300 pairs for the testing set to assess the performance of the final network quantitatively. Furthermore, the training set was enlarged to 21,600 pairs by employing overlapping cropping and data augmentation, resulting in low-resolution inputs of 33 × 33 pixels and high-resolution ground truth images of 99 × 99 pixels. This pixel count was empirically selected to provide enough detail for effective learning while keeping the computation efficient.

The model [33–38] establishes an end-to-end mapping relationship between low-resolution and high-resolution image pairs to achieve computational subpixel-resolution. Trained model enables rapid, high-resolution reconstruction from a single input [39–46]. This methodology enables hardware-independent operation with computational efficiency and strong robustness, paving the way for decoupling spatial resolution from the image sensor pixel size constraints of a low-magnification ultra-compact metalens microscope. Inspired by the successful application of deep learning models in photography and traditional microscopy, we employed a Hybrid Attention Transformer (HAT) model [38] to reconstruct high-resolution images by learning the mapping from unit-magnification and three-fold magnification imaging. Here, the HAT is chosen for its unique hybrid attention mechanism, which achieves an optimal balance among spatial detail fidelity (versus Restormer [47]), cross-window information integration for structural continuity (versus SwinIR [37] and Uformer [48]), and training/inference stability (versus diffusion-based models [49]). These advantages make it particularly suitable for microscopic image reconstruction requiring high fidelity and structural coherence.

As illustrated in Fig. 3, this model first applies a convolutional layer to extract shallow features from the low-resolution input, followed by a series of Residual Hybrid Attention Groups (RHAGs) and a 3×3 convolution layer for deep feature extraction. A global residual connection merges shallow and deep features, which are then up-sampled via a pixel-shuffle-based reconstruction module. Each RHAG contains multiple Hybrid Attention Blocks (HABs), an Overlapping Cross-Attention Block (OCAB), and a 3×3 convolution layer with residual connections. The Channel Attention Block (CAB) is integrated into the standard Swin Transformer block, positioned in parallel with the shifted window-based multi-head self-attention ((S)W-MSA) module after the first LayerNorm layer. The OCAB includes an Overlapping Cross-Attention layer and a multilayer perceptron (MLP) to enable cross-window interaction. This transformer-based design effectively captures long-range dependencies and local structural details through the combination of self-attention, channel attention, and overlapping cross-attention. By activating a broader range of pixels during reconstruction, the model achieves significantly improved performance, which motivated its adoption in our work.

We used the total loss function to optimize the network parameters, which is a combination of mean absolute error (MAE) loss and Gradient loss [39]. The MAE loss captures pixel-level differences, guiding the model to preserve global structural fidelity, while the gradient loss enhances image sharpness by emphasizing local edge details. The total loss function is defined as

$$L_{total} = \| \widehat{Y} - Y^{HR} \|_1 + \alpha(\| \nabla_x * \widehat{Y} - \nabla_x * Y^{HR} \|_1 + \| \nabla_y * \widehat{Y} - \nabla_y * Y^{HR} \|_1), \quad (2)$$

where $\widehat{Y}$ is the model output, Y^{HR} is the responding ground truth, and α is a weighting factor empirically set to 0.05 to balance the relative contributions of the two terms. The gradient loss is computed using the horizontal and vertical image gradient operators, defined as

$$\nabla_x = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}, \ \nabla_y = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix}, \quad (3)$$

where ∇_x and ∇_y denote the horizontal and vertical gradient convolution kernels, respectively. The model was trained until convergence by minimizing the loss function. Additional training details can be found in Supplementary Note 3.

Results

After completing the training process, we test the trained model using the testing dataset containing the input images previously unseen by the network. Figure 4(a) presents a representative low-resolution input captured by the metalens ($1 \times$), with two regions of interest (ROIs) labeled for revealing further details. To benchmark the model output against the widely used interpolation algorithm, the bicubic interpolation algorithm was applied for up-sampling, as shown in Fig. 4(b). The model output is displayed in Fig. 4(c), while the corresponding high-resolution ground truth image, captured by the metalens ($3 \times$), is shown in Fig. 4(d). Detailed visual comparisons of the two ROIs are

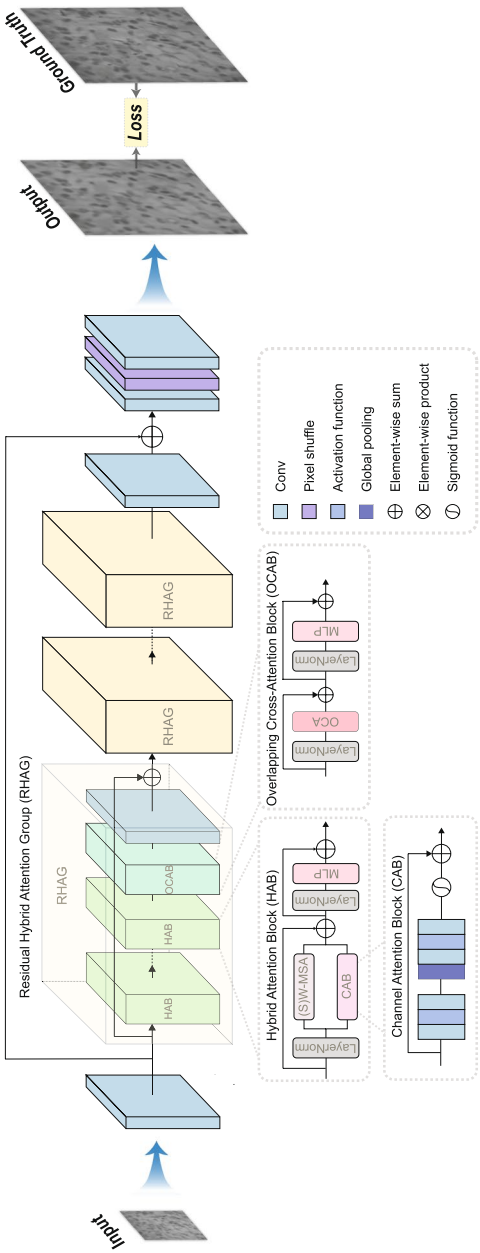

Fig. 3 Architecture of the Hybrid Attention Transformer model. It consists of self-attention, channel attention, and a new overlapping cross-attention to activate more pixels for better reconstruction. The network is trained in an end-to-end manner using experimental image pairs acquired with the metalens (1 ×) and the metalens (3 ×)

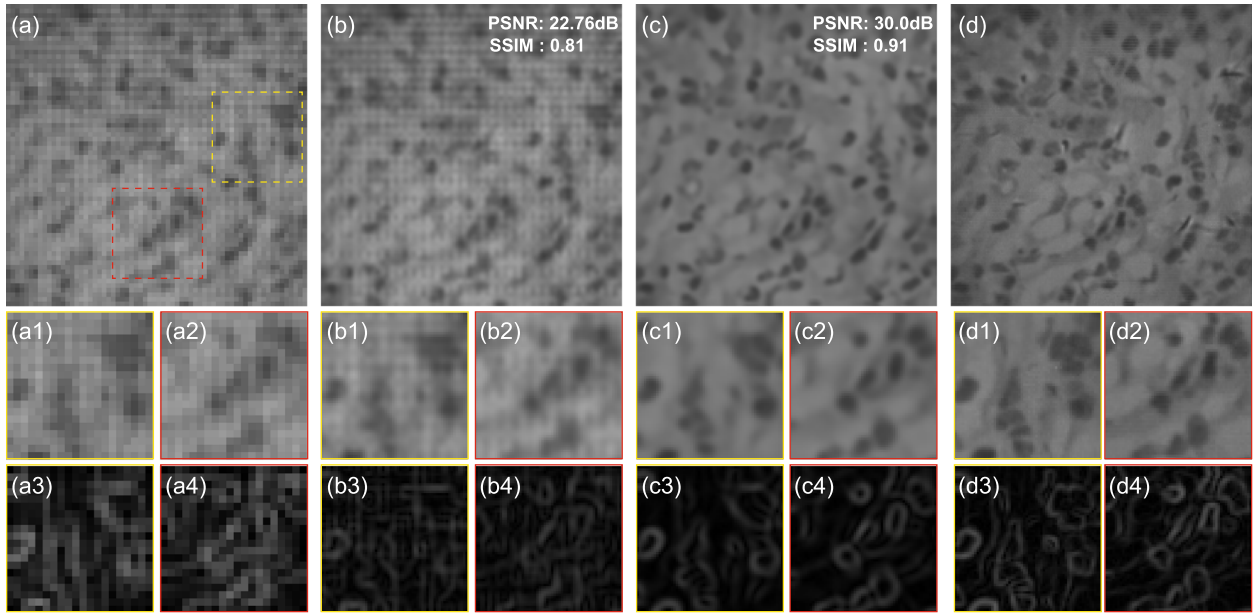

Fig. 4 **a** Network input image captured by the metalens (1 ×). The highlighted regions are two representative regions of interest (ROIs). **b** Bicubic interpolation algorithm-enhanced image. The corresponding PSNR and SSIM values are indicated in the upper-right corner. **c** Network output image. The corresponding PSNR and SSIM values are indicated in the upper-right corner. **d** Ground truth image captured by the metalens (3 ×). (a1, a2) Zoomed-in ROIs of the input image (**a**). (b1, b2) Zoomed-in ROIs of the bicubic interpolation algorithm image (**b**). (c1, c2) Zoomed-in ROIs of the neural network output image (**c**). (d1, d2) Zoomed-in ROIs of the ground truth image (**d**). (a3-d4) present the gradient maps corresponding to the zoomed-in ROIs (a1-d2)

provided in Fig. 4(a1)-(d2). From these zoomed-in images, it is evident that our model significantly improves image resolution, revealing finer cellular details and effectively suppressing grid artifacts present in the input. The reconstruction closely matches the high-resolution ground truth image, whereas the bicubic interpolation yields very limited improvements. To more clearly demonstrate the model's ability to recover high-frequency details, gradient maps of the zoomed-in ROIs are compared in Fig. 4(a3)-(d4). The model output preserves sharp edge features that closely resemble the ground truth, while the interpolated results exhibit noticeable artifacts. To evaluate the performance of the model quantitatively, the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) were calculated. Compared to bicubic interpolation (PSNR: 22.76dB, SSIM: 0.81), the model output (PSNR: 30dB, SSIM: 0.91) exhibits remarkable improvements of 7.24 dB in PSNR and 0.10 in SSIM. In terms of computational efficiency, the model generates each high-resolution output (198×198 pixels, covering a $110\ \mu m \times 110\ \mu m$ FOV) in approximately 0.2 s on a single GPU, demonstrating its potential for rapid, single-frame high-resolution reconstruction.

We further extend the trained model to the metalens array-integrated microscope for wide-field imaging. The array comprises 20×30 sub-metalenses, each with a diameter of 200 μm and a focal length of 250 μm. Figure 5(a) presents a photograph of the metalens array-integrated microscope, showing the overall size of the metalens array to be 4 mm × 6 mm. Figure 5(b) displays the microscope image captured from a portion of the array. Despite minor parameter variations between the array's sub-metalenses and the metalens (1 ×), the imaging conditions remain consistent, enabling direct application of the trained model without modification. To cover the entire FOV, the sample was translated 9 times across a 3×3 grid following a zigzag scanning trajectory to compensate for blind areas between adjacent lenses. By digitally stitching the model-enhanced outputs from each sub-region, a wide-area, high-resolution image was reconstructed, as shown in Fig. 5(c), with the FOV of a standard 10 × Olympus objective (NA = 0.25) indicated by a blue dashed box for scale comparison. The image stitching procedure and the corresponding raw stitched full-field image are detailed in Supplementary Note 4. The final stitched FOV reached 4 mm × 6 mm, nearly covering the entire active area of the CMOS image sensor (4.616 mm × 6.440 mm). To further validate the model's effectiveness, Fig. 5(d)-(f) provide a comparison of part of the raw stitched image, the corresponding model output, and the reference image acquired with the 10 × Olympus objective under the same FOV. The raw image exhibits limited resolution, while the model output reveals clearer cellular structures and substantial improvement in visual quality, demonstrating visual consistency with the Olympus objective (10 ×), emphasizing the high fidelity of the model, as illustrated in detailed ROI comparisons in Fig. 5(g)-(l). In contrast to conventional objectives limited by narrow FOVs, our computational metalens array system maintains high-resolution imaging across a 14.5 × larger area through neural network enhancement. This demonstrates its potential for high-throughput imaging in large-scale biomedical applications. The preliminary clinical diagnostic interpretation relies primarily on cell morphology differences. Normal thyroid cells are relatively sparse and are often accompanied by large vacuoles, while thyroid carcinoma cells exhibit denser and more compact nuclei. As shown in Fig. 5(c), the upper region with sparse cells and large vacuoles corresponds to normal thyroid tissue, while the lower region with densely

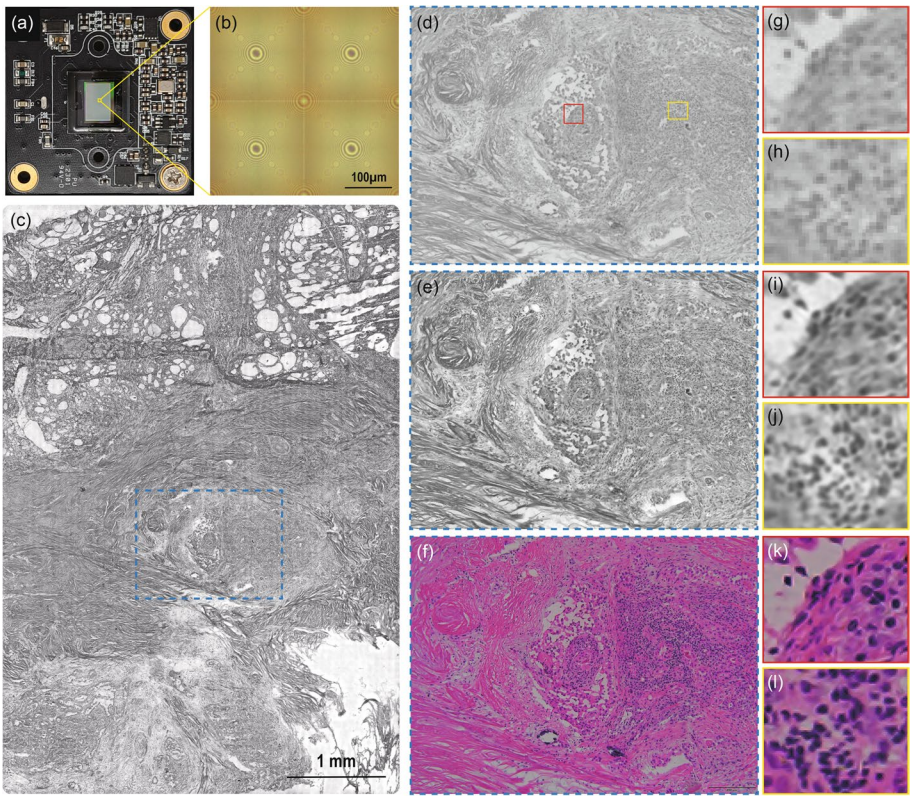

Fig. 5 **a** Photograph of the compact microscope integrated with a 20 × 30 metalens array. **b** Microscope image captured from a portion of the metalens array. **c** Final stitched full-field image of a stained thyroid pathological section reconstructed from the network outputs. The blue highlighted region denotes the FOV of a standard 10 ×/0.25NA Olympus objective. **d** Raw image acquired by the metalens array-based integrated microscope corresponds to the FOV of the Olympus objective. **e** Network output corresponds to the FOV of the Olympus objective. **f** Image acquired by the standard 10 ×/0.25NA Olympus objective. **g**, **h** Zoomed-in ROIs of the raw image (**d**). **i**, **j** Zoomed-in ROIs of the network output image (**e**). **k**, **l** Zoomed-in ROIs of the objective image (**f**)

packed nuclei indicates suspected carcinoma cells. This demonstrates that the system enables clear, simultaneous visualization of normal and suspected cancerous regions within a single field of view, supporting wide-field, portable, and preliminary diagnosis of thyroid cancer for subsequent clinical evaluation.

Discussion and conclusion

In conclusion, we have successfully demonstrated a computational imaging framework addressing the limitations of image sensor resolution inherent in a unit-magnification metalens-based compact microscope. Rather than relying on complex lens designs or modifications to hardware, this methodology employs a transformer-based neural network to enhance image quality from low-resolution inputs. To the best of our knowledge, this constitutes the first effort to establish an experimental dataset comprised of metalens-acquired thyroid pathological sections, thereby providing critical training data for a neural network aimed at enabling accurate and high-quality improvements in image resolution utilizing a single frame input, with a resolution close to the ground truth of 0.87 μm. The trained model was subsequently applied to a metalens array-based

integrated microscope, achieving an extensive FOV measuring 4 mm × 6 mm, with significantly enhanced resolution and high fidelity, approximating the image quality of the Olympus 10 ×/0.25NA objective. The resulting FOV is approximately 14.5 times that of the referenced 10 ×/0.25 NA Olympus objective.

Our research emphasizes the synergistic relationship between simplified optics and neural network-enhanced image processing, illustrating a scalable and cost-effective approach to high-resolution microscopy. It is highlighted that this framework provides wide-field, sub-pixel imaging without added optical complexity via a metalens-transformer design, demonstrates experimentally validated performance on real thyroid pathology sections with reasonable cross-tissue generalizability, and features a compact, low-cost configuration suitable for portable diagnostic applications. While demonstrating promising performance, our reconstruction performance relies on stable illumination and sensor conditions. Factors such as under- or overexposure, nonuniform or oblique lighting, and sensor defects can affect image quality and should be carefully controlled. Additionally, minimizing potential hallucination artifacts and ensuring accurate pixel-level alignment are important for maintaining reconstruction fidelity. Prospective advancements may involve leveraging optimized metalens designs to generate higher-fidelity ground truth data for training, as well as incorporating cross-modality training with high-quality outputs from conventional microscopes to further enhance the network's reconstruction performance. Additionally, a cross-tissue generalization experiment using five other stained tissue types (stomach, lymph node, pancreas, spleen, and rectum) showed that the model maintained good performance on most tissues, while performance would decline a little for morphologically distinct rectum samples. Details are provided in the Supplementary Note 5. Moreover, the generalizability of the model warrants exploration through task-specific training focused on particular pathological conditions or through the employment of transfer learning techniques aimed at enhancing adaptability across diverse sample types, and further extended via training on diverse multi-institutional datasets for broader clinical deployment. Collectively, this framework establishes new avenues for intelligent, compact imaging systems capable of acquiring large-field and high-resolution images, presenting significant potential for applications in biomedical imaging, point-of-care diagnostics, and beyond.

Abbreviations

FOV	Field of view
CSM	Chip-scale metalens-integrated microscope
CMOS	Complementary metal-oxide semiconductor
NA	Numerical aperture
CPF	Circular polarization filter
OCA	Optically clear adhesive
PB	Pancharatnam-Berry
SEM	Scanning electron microscope
MTF	Modulation transfer function
USAF	United States Air Force
HAT	Hybrid Attention Transformer
RHAG	Residual Hybrid Attention Group
HAB	Hybrid Attention Block
OCAB	Overlapping Cross-Attention Block
CAB	Channel Attention Block
(S)W-MSA	Shifted window-based multi-head self-attention
MLP	Multilayer perceptron
MAE	Mean absolute error
ROIs	Regions of interest

PSNR	Peak signal-to-noise ratio
SSIM	Structural similarity index measure

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43074-025-00213-3.

Supplementary Material 1.

Acknowledgements
Not applicable.

Authors' contributions
S.H., C.C., and T.L. conceived the concept. S.H. proposed the metalens design; J.L. and J.S. fabricated the samples; S.H. established the experimental dataset and implemented the deep learning model for the proposed framework; S.H. and Z.W. performed the imaging experiments using the metalens array-based integrated microscope and analyzed the data; S.H. performed the optical measurements with the help of W.S. and J.W.; S.H., Xu L., C.C., and T.L. discussed the results with help from all authors; S.H. and C.C. wrote the manuscript with the help of T.L.; T.L. supervised the project.

Funding
The authors acknowledge financial support from the National Key Research and Development Program of China (2022YFA1404301, 2024YFA1012600), National Natural Science Foundation of China (Nos. 62325504, 62305149, 92250304, 62288101), and Dengfeng Project B of Nanjing University. The authors acknowledge the micro-fabrication center of the National Laboratory of Solid State Microstructures (NLSSM) for technique support.

Data availability
The source data are available from the corresponding author upon reasonable request. All data needed to evaluate the conclusion are present in the manuscript and/or the Supplementary Information.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
All authors agreed to publish this paper.

Competing interests
The authors declare no conflicts of interest.

Received: 19 June 2025 Revised: 22 October 2025 Accepted: 24 November 2025
Published online: 03 December 2025

References

1. Goodman JW. Introduction to Fourier Optics. Roberts and Company publishers, 2005.
2. Yu N, Genevet P, Kats MA, Aieta F, Tetienne J-P, Capasso F, et al. Light propagation with phase discontinuities: generalized laws of reflection and refraction. Science. 2011;334(6054):333–7.
3. Khorasaninejad M, Chen WT, Devlin RC, Oh J, Zhu AY, Capasso F. Metalenses at visible wavelengths: diffraction-limited focusing and subwavelength resolution imaging. Science. 2016;352(6290):1190–4.
4. Li T, Chen C, Xiao X, Chen J, Hu S, Zhu S. Revolutionary meta-imaging: from superlens to metalens. Photonics Insights. 2023;2:R01.
5. Arbabi A, Horie Y, Bagheri M, Faraon A. Dielectric metasurfaces for complete control of phase and polarization with subwavelength spatial resolution and high transmission. Nat Nanotechnol. 2015;10(11):937–43.
6. Arbabi A, Arbabi E, Kamali SM, Horie Y, Han S, Faraon A. Miniature optical planar camera based on a wide-angle metasurface doublet corrected for monochromatic aberrations. Nat Commun. 2016;7:13682.
7. Wang S, Wu PC, Su V-C, Lai Y-C, Chu CH, Chen J-W, et al. Broadband achromatic optical metasurface devices. Nat Commun. 2017;8(1):187.
8. Paniagua-Dominguez R, Yu YF, Khaidarov E, Choi S, Leong V, Bakker RM, et al. A metalens with a near-unity numerical aperture. Nano Lett. 2018;18(3):2124–32.
9. Wang Y, Chen C, Wu S, Ye X, Zhu S, Li T. Metalens with tilted structures for high-efficiency focusing at large-angle incidences. Chin Opt Lett. 2024;22(5):53601.
10. Arbabi E, Li J, Hutchins RJ, Kamali SM, Arbabi A, Horie Y, et al. Two-photon microscopy with a double-wavelength metasurface objective lens. Nano Lett. 2018;18(8):4943–8.
11. Chen C, Song W, Chen J-W, Wang J-H, Chen YH, Xu B, et al. Spectral tomographic imaging with aplanatic metalens. Light Sci Appl. 2019;8(1):99.
12. Kwon H, Arbabi E, Kamali SM, Faraji-Dana M, Faraon A. Single-shot quantitative phase gradient microscopy using a system of multifunctional metasurfaces. Nat Photonics. 2020;14(2):109–14.

13. Liu Y, Yu Q-Y, Chen Z-M, Qiu H-Y, Chen R, Jiang S-J, et al. Meta-objective with sub-micrometer resolution for microendoscopes. Photonics Res. 2021;9(2):106–15.
14. Luo Y, Tseng ML, Vyas S, Hsieh T-Y, Wu J-C, Chen S-Y, et al. Meta-lens light-sheet fluorescence microscopy for in vivo imaging. Nanophotonics. 2022;11(9):1949–59.
15. Long Y, Zhang J, Liu Z, Feng W, Guo S, Sun Q, et al. Metalens-based stereoscopic microscope. Photonics Res. 2022;10(6):1501–8.
16. Li L, Wang S, Zhao F, Zhang Y, Wen S, Chai H, et al. Single-shot deterministic complex amplitude imaging with a single-layer metalens. Sci Adv. 2024;10(1):eadl0501.
17. Ji Z, Chen Q, Sha X, Wang H, Ma X, Liu Z, et al. Multidimensional multiplexing metalens for STED microscopy. Sci Adv. 2025;11(17):eadt2807.
18. Xu B, Li H, Gao S, Hua X, Yang C, Chen C, et al. Metalens-integrated compact imaging devices for wide-field microscopy. Adv Photonics. 2020;2(6):66004.
19. Ye X, Qian X, Chen Y, Yuan R, Xiao X, Chen C, et al. Chip-scale metalens microscope for wide-field and depth-of-field imaging. Adv Photonics. 2022;4(4):46006.
20. Latif J, Xiao C, Imran A, Tu S. Medical imaging using machine learning and deep learning algorithms: a review. in 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (ICoMET) (IEEE). 2019; pp. 1–5.
21. Barbastathis G, Ozcan A, Situ G. On the use of deep learning for computational imaging. Optica. 2019;6(8):921–43.
22. Archana R, Jeevaraj PSE. Deep learning models for digital image processing: a review. Artif Intell Rev. 2024;57(1):11.
23. Wang N, Yan W, Qu Y, Ma S, Li SZ, Qiu M. Intelligent designs in nanophotonics: from optimization towards inverse creation. PhotoniX. 2021;2(1):22.
24. Chen MK, Liu X, Sun Y, Tsai DP. Artificial intelligence in meta-optics. Chem Rev. 2022;122(19):15356–413.
25. Chen J, Hu S, Zhu S, Li T. Metamaterials: from fundamental physics to intelligent design. Interdiscip Mater. 2023;2(1):5–29.
26. Tseng E, Colburn S, Whitehead J, Huang L, Baek S-H, Majumdar A, et al. Neural nano-optics for high-quality thin lens imaging. Nat Commun. 2021;12(1):6493.
27. Fan Q, Xu W, Hu X, Zhu W, Yue T, Zhang C, et al. Trilobite-inspired neural nanophotonic light-field camera with extreme depth-of-field. Nat Commun. 2022;13(1):2130.
28. Chu CH, Chia Y-H, Hsu H-C, Vyas S, Tsai C-M, Yamaguchi T, et al. Intelligent phase contrast meta-microscope system. Nano Lett. 2023;23(24):11630–7.
29. Liu Y, Li W-D, Xin K-Y, Chen Z-M, Chen Z-Y, Chen R, et al. Ultra-wide FOV meta-camera with transformer-neural-network color imaging methodology. Adv Photonics. 2024;6(5):56001.
30. Seo J, Jo J, Kim J, Kang J, Kang C, Moon S-W, et al. Deep-learning-driven end-to-end metalens imaging Adv Photonics. 2024;6(6):66002.
31. Chia Y, Liao W, Vyas S, Chu CH, Yamaguchi T, Liu X, et al. In vivo intelligent fluorescence endo-microscopy by varifocal meta-device and deep learning. Adv Sci. 2024;11(20):2307837.
32. Fröch JE, Chakravarthula P, Sun J, Tseng E, Colburn S, Zhan A, et al. Beating spectral bandwidth limits for large aperture broadband nano-optics. Nat Commun. 2025;16(1):3025.
33. Dong C, Loy CC, He K, Tang X. Learning a deep convolutional network for image super-resolution. in European Conference on Computer Vision (Springer). 2014; pp. 184–199.
34. Tai Y, Yang J, Liu X. Image super-resolution via deep recursive residual network. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017; pp. 3147–55.
35. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, et al. Photo-realistic single image super-resolution using a generative adversarial network. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017; pp. 4681–90.
36. Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, et al. Esrgan: Enhanced super-resolution generative adversarial networks. in Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018; p. 0.
37. Liang J, Cao J, Sun G, Zhang K, Van Gool L, Timofte R. Swinir: Image restoration using swin transformer. in Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021; pp. 1833–44.
38. Chen X, Wang X, Zhou J, Qiao Y, Dong C. Activating more pixels in image super-resolution transformer. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023; pp. 22367–77.
39. Rivenson Y, Göröcs Z, Günaydin H, Zhang Y, Wang H, Ozcan A. Deep learning microscopy. Optica. 2017;4(11):1437–44.
40. Rivenson Y, Ceylan Koydemir H, Wang H, Wei Z, Ren Z, Günaydın H, et al. Deep learning enhanced mobile-phone microscopy. ACS Photonics. 2018;5(6):2354–64.
41. Ouyang W, Aristov A, Lelek M, Hao X, Zimmer C. Deep learning massively accelerates super-resolution localization microscopy. Nat Biotechnol. 2018;36(5):460–8.
42. Wang H, Rivenson Y, Jin Y, Wei Z, Gao R, Günaydın H, et al. Deep learning enables cross-modality super-resolution in fluorescence microscopy. Nat Methods. 2019;16(1):103–10.
43. Qiao C, Li D, Guo Y, Liu C, Jiang T, Dai Q, et al. Evaluation and development of deep neural networks for image super-resolution in optical microscopy. Nat Methods. 2021;18(2):194–202.
44. Huang B, Li J, Yao B, Yang Z, Lam EY, Zhang J, et al. Enhancing image resolution of confocal fluorescence microscopy with deep learning. PhotoniX. 2023;4(1):2.
45. Chen X, Qiao C, Jiang T, Liu J, Meng Q, Zeng Y, et al. Self-supervised denoising for multimodal structured illumination microscopy enables long-term super-resolution live-cell imaging. PhotoniX. 2024;5(1):4.
46. Qian J, Wang C, Wu H, Chen Q, Zuo C. Ensemble deep learning-enabled single-shot composite structured illumination microscopy (eDL-cSIM). PhotoniX. 2025;6(1):13.

47. Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang M-H. Restormer: Efficient transformer for high-resolution image restoration. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022; pp. 5728–39.
48. Wang Z, Cun X, Bao J, Zhou W, Liu J, Li H. Uformer: A general u-shaped transformer for image restoration. in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022; pp. 17683–93.
49. Yue Z, Wang J, Loy CC. ResShift: Efficient diffusion model for image super-resolution by residual shifting. Adv Neural Inf Process Syst. 2023;36:13294–307.

Publisher's Note