MetaScope: Optics-Driven Neural Network for Ultra-Micro Metalens Endoscopy

Supplementary Material

The supplementary is organized as follows.

Supp. A: Details of metalens:
- metalens imaging,
- metalens fabrication,
- chromatic aberration,
- existing physical and algorithmic solutions.

Supp. B: Optical simulation details of:
- spatial optical prior,
- channel optical prior.

Supp. C: Experiments including:
- real scene generalization
- data scaling-up,
- cross-metalens generalization,
- model efficiency,
- dataset analysis,
- extensive qualitative analysis.

Supp. D: Discussions and clarifications about:
- dataset setup,
- physically suitable for endoscopy,
- related works,
- mathematical proofs,
- ethical clarification.

A. Delving into Metalens

Traditional convex lenses focus light by varying the thickness from the center to the edge, altering the optical path length of incoming light. However, this design inherently causes spherical and chromatic aberrations. Chromatic aberration is further divided into axial chromatic aberration (ACA) and lateral chromatic aberration (LCA). ACA involves variations in focal length along the optical (z) axis for different wavelengths. LCA results in the positional displacement of focused colors in the transverse (x-y) plane on the focal plane. Conventional convex lenses employ multiple lens elements arranged in groups to mitigate these aberrations, often leading to bulky optical systems.

Differently, metalenses [15] *(published in Science, 2016)* represent a novel class of lenses that are ultra-lightweight and free from bulky designs. These lenses comprise numerous sub-wavelength structures (scale elements within the visible spectrum) meticulously arranged on a planar surface. Each nanostructure can independently manipulate the phase of the transmitted light wave, allowing for precise deformation of the wavefront and achieving accurate focusing. By programming a target hyperbolic phase profile that inherently satisfies the Abbe sine condition, metalenses can theoretically eliminate spherical aberration at the design stage. Combined with their sub-micron thickness and planar form factor, this paradigm shift allows metalenses to achieve high-performance focusing without the bulk of traditional optics, offering transformative potential for miniaturized imaging systems and photonic integration.

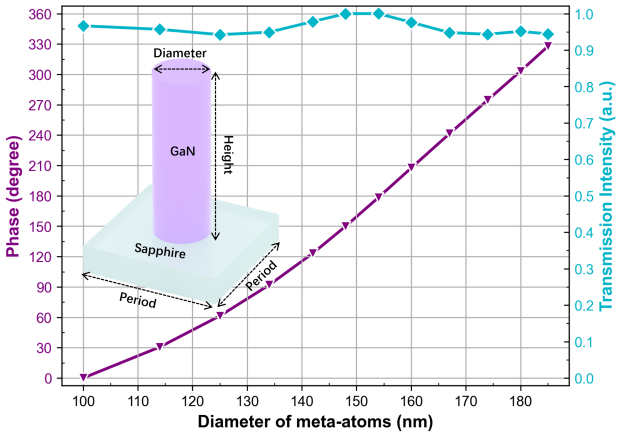

Figure 1. Design of meta-atoms. The phase and transmission intensity of the GaN nanopillars.

A.1. Metalens Design and Fabrication

However, optimizing the achromatic aberration of metalenses remains challenging in meta-optics. The focusing phase of a metalens for a specific working wavelength λ is determined by the following equation:

$$\phi(r,\lambda) = -\frac{2\pi}{\lambda}\left(\sqrt{r^2+f^2}-f\right), \quad \lambda \in [\lambda_{\min}, \lambda_{\max}] \tag{1}$$

where $\phi(r,\lambda)$ represents the required focusing phase at position r for the wavelength λ, and f denotes the desired focal length. In this study, the metalens, with a diameter of 2.6 mm, is designed to achieve a focal length of 10 mm at a wavelength of 532 nm.

Fig. 1 shows the employed polarization-independent meta-atoms, comprising cylindrical gallium nitride (GaN) nanopillars on a sapphire substrate. Each meta-atom features a fixed height of 850 nm and a unit-cell periodicity of 280 nm. The parametric variation of the nanopillar diameters (100–185 nm) enables full 2π-phase coverage for wavefront focusing. The data of phase shift and transmission intensity are derived from numerical simulation with the commercial software COMSOL Multiphysics©. As shown in Eq. 1, different meta-atoms are arranged at corresponding locations according to the focusing phase profile.

In Fig. 2, the metalens is manufactured through a multi-step lithographic patterning and dry etching workflow. The

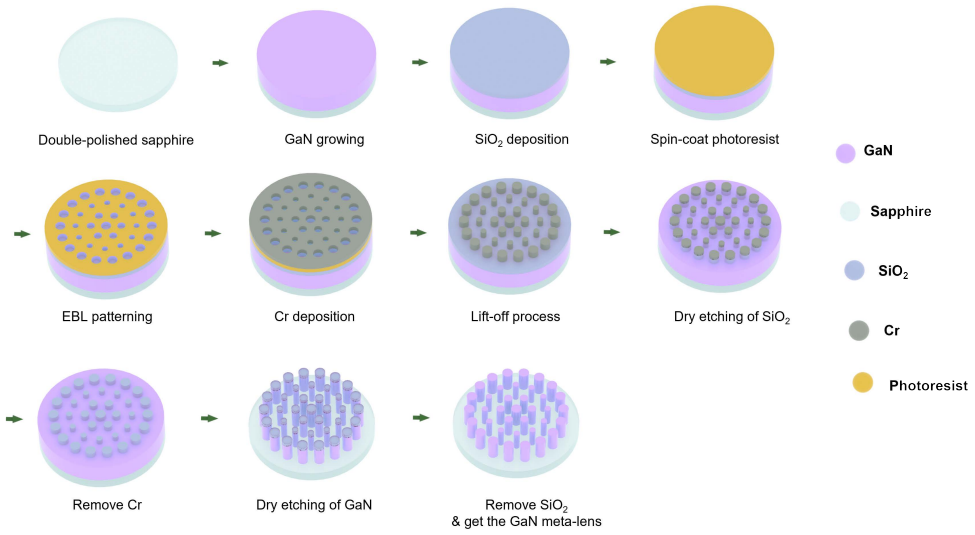

Figure 2. Fabrication flow chart of the metalens.

process begins with an ultrasmooth c-plane sapphire substrate coated with an 850 nm GaN epitaxial layer grown via metalorganic chemical vapor deposition (MOCVD). A 200 nm silicon dioxide (SiO_2) film, serving as an etch-resistant hard mask, is thermally evaporated onto the GaN surface. To pattern the metasurface, a polymethyl methacrylate (PMMA) photoresist layer (200 nm thick) is deposited via spin coating, followed by a 180°C soft bake for 3 minutes. High-resolution electron-beam lithography (ELS-HS50, ELIONIX INC.) directly writes the metalens design into the PMMA layer. Post-exposure development involves immersing the substrate in a methyl isobutyl ketone/isopropyl alcohol (MIBK: IPA = 1:3) solution for 75 seconds, followed by an IPA rinse (20 seconds) to terminate the reaction. A 40 nm chromium (Cr) film is subsequently evaporated onto the patterned resist, and lift-off in acetone selectively removes excess Cr to define the mask geometry. This Cr stencil guides the first inductively coupled plasma reactive ion etching (ICP-RIE) step using CF_4 plasma (Samco RIE-200iPT) to transfer the pattern into the underlying SiO_2 layer. Residual Cr is stripped via wet etching, exposing the SiO_2 hard mask. The final nanostructuring involves a second ICP-RIE cycle with a Cl_2/Ar plasma to anisotropically etch the GaN layer, resulting in high-aspect-ratio nanopillars. Buffered oxide etch (BOE) solution removes the remaining SiO_2 mask, leaving an array of precisely defined GaN nanostructures anchored to the sapphire substrate.

A.2. Chromatic Aberrations

The focal length at a different wavelength can be estimated by scaling the designed focal length of 532 nm proportionally to the wavelength ratio, assuming that the lens material and structure introduce purely dispersive effects without significant aberrations [15, 36]: $f(\lambda) = f_0 \times \frac{\lambda}{\lambda_0}$, where f_0 and λ_0 are the constants of the designed metalens parameters, facilitating the prediction of focal lengths across different wavelengths. As illustrated in Fig. 3, the light of different wavelengths (i.e., different colors) will have different focal lengths (along the z-axis). Red light has a shorter focal length, while blue light has a longer one. Due to these varying focal lengths, specific light colors will appear out of focus when capturing color images, resulting in blurring differently in each color channel. Mismatched color offsets can also create color fringing around objects, especially at high-contrast edges.

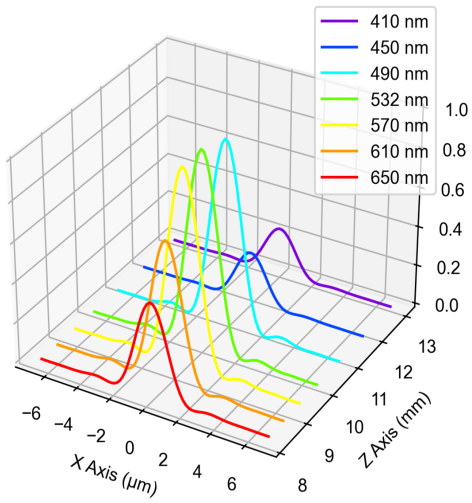

Figure 3. Focal lengths of the metalens for different wavelengths.

Color	Red	Orange	Yellow	Green	Cyan	Blue	Violet
Wavelength	650 nm	610 nm	570 nm	532 nm	490 nm	450 nm	410 nm
Efficiency	0.3524	0.5281	0.7371	0.9920	0.7032	0.1885	0.1738

Table 1. Comparison of the intensity at the focusing points for each color, serving as the transformation efficiency $\mathbf{T}$ in the proposed Optics-informed Intensity Adjustment (OIA) module.

A.3. How to Solve Metalens Chromatic Aberrations

Physical Solution. To physically correct ACA by ensuring identical focal lengths across different wavelengths, the design of an achromatic metalens must incorporate an additional wavelength-dependent phase delay $\Delta\phi(r,\lambda)$ to achieve the optical rectification [11, 26, 38]:

$$\phi_{Achromatic}(r,\lambda) = \phi(r,\lambda_{max}) + \Delta\phi(r,\lambda), \tag{2}$$

$$\Delta\phi(r,\lambda) = -\left[2\pi\left(\sqrt{r^2+f^2}-f\right)\right]\left(\frac{1}{\lambda}-\frac{1}{\lambda_{\max}}\right) + \frac{\delta}{\lambda}\cdot\frac{\lambda_{\min}\lambda_{\max}}{\lambda_{\max}-\lambda_{\min}} - \frac{\delta\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}}. \tag{3}$$

Here, $\phi_{\text{Achromatic}}(r,\lambda)$ is the achromatic focusing phase at position r for wavelength λ, and δ represents the maximum additional phase shift required. However, as the size of the metalens increases, the necessary additional phase delay also grows, posing significant and open challenges. *Due to the limitations of current micro-nano processing technology, finding a solution solely through the geometric design of the nano-antennas proves challenging.*

Computer Vision Solution. To overcome inherent physical constraints, alternative approaches that leverage computer vision and computational optics have garnered significant attention for addressing ACA in metalenses [7, 33, 35, 37, 41]. Recent advancements and the democratization of computing platforms have positioned these methods as promising solutions, effectively bridging both hardware and software limitations. By accurately modeling the degradation patterns caused by ACA, machine learning algorithms can be trained to predict and compensate for aberrations, reducing reliance on intricate nanoantenna designs. Additionally, computer vision techniques enable the learning and adaptation to ACA-induced degradation patterns, offering a versatile and efficient means to mitigate aberrations in metalens systems. This integration of computer vision enhances the performance and reliability of metalenses and facilitates scalable and adaptable solutions for real-world applications, highlighting the transformative potential of combining computational intelligence with optical design.

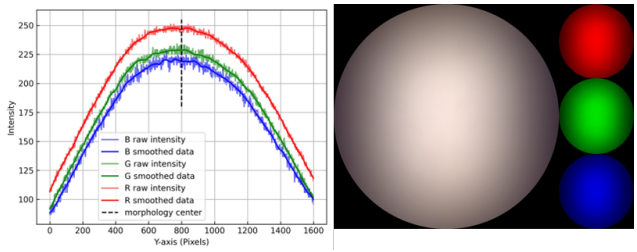

Figure 4. Illustration of the spatial prior of the metalens imaging.

B. Optical Simulation Experiments

B.1. Channel Prior

To model the specific color dispersion and derive the channel prior $\mathbf{T}$, we use Fresnel diffraction [6, 32] to simulate the light propagation, as described by the following equation:

$$I(x,y,z) = |E(x,y,z)|^2 = \left| \iint E_0(u,v)\, P(u,v)\, T(u,v)\, \frac{e^{ikz/\lambda z}}{\lambda z} \times \exp\left(\frac{ik}{2z}\left[(x-u)^2+(y-v)^2\right]\right) du\, dv \right|^2. \tag{4}$$

Here:

- $I(x,y,z)$ represents the light intensity at position (x,y,z), which is the energy of electric field $E(x,y,z)$.
- $E_0(u,v) = A_0(u,v)\exp\left(i\phi_0(u,v)\right)$ is the input electric field on the initial plane (u,v). $A_0(u,v)$ is the input amplitude, and $\phi_0(u,v)$ is the input phase.
- $P(u,v) = \begin{cases} 1 & \text{if } \sqrt{u^2+v^2} \le R \\ 0 & \text{if } \sqrt{u^2+v^2} > R \end{cases}$ is the aperture function. R is the lens radius. An aperture is an optical element that limits the propagation of a light beam, defining the area through which the light passes. The function $P(u,v)$ takes a value of 1 within the aperture area (indicating that light passes) and 0 outside this area (indicating that light does not pass).
- $T(u,v) = A_{\text{meta}}(u,v,\lambda)\exp\left(i\phi_{meta}(u,v,\lambda)\right)$ is the transformation function of metalens, including the amplitude $A_{meta}(u,v,\lambda)$ and phase $\phi_{meta}(u,v,\lambda)$ modulation provided by the metalens at the position (u,v) for the working wavelength λ.
- $k = \frac{2\pi}{\lambda}$ is the wave vector.

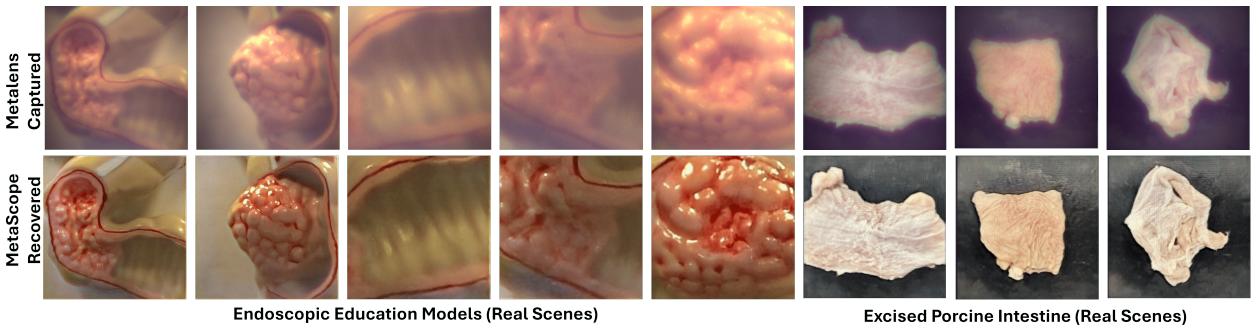

Figure 5. Qualitative restoration on the real scenes, including endoscopic education models (Left) and excised porcine intestine (Right).

	Meta-CVC-Clinic	Meta-CVC-Colon	Meta-Kvasir-Seg	Meta-EndoVis-17	Meta-EndoVis-18
DRMI	34.05	36.50	32.76	30.82	30.90
Ours	**36.47**	**37.53**	**33.40**	**31.40**	**31.55**

Table 2. Performance comparison on a unified model trained with all five datasets.

metalens-1 (10 mm)	metalens-2 (5 mm)	
-	Zero-shot	Fine-tuned
85.55	82.82	**86.37**

Table 3. Cross-metalenses (cross-dataset) generalization of using the MetaScope trained on metalens-1 dataset (mIoU).

- λ is the wavelength.

With this equation, we could derive the Point Spread Function (PSF) details of all wavelengths through the metalens in 3D space. PSF is the imaging response of an optical system to an ideal point light source, directly governing spatial resolution and color fidelity, which can be regarded as the distortion kernel. The chromatically aberrated image $I_{ca}(x, y)$ can be mathematically expressed as the convolution of the true, clean image with the PSF of the optical system, $I_{ca}(x, y) = I_0(x, y) * PSF(x, y)$. For each wavelength, the light intensity $I(x, y, z)$ will reach the maximum at the focal spot position (x_0, y_0, f_λ). Fig. 3 demonstrates the focal sections along the x axis at the focal plane (x, y, f) for seven wavelengths of different colors, $\sum_\lambda I(x, y, z)|_{y=y_0, z=f_\lambda}$. Compared to the designed green light at 532 nm, the focusing efficiency of other colors is slightly lower. Tab. 1 compares the intensity at the focusing points for each color. This efficiency buffer is the transformation efficiency $\mathbf{T}$ in the proposed Optics-informed Intensity Adjustment (OIA) module. Considering the three-channel properties of images, the red, green, and blue (RGB) ones are encoded into optical embeddings to adjust the channel attention.

B.2. Spatial Prior

To analyze the spatial light distribution pattern $\mathbf{Y}$ of the metalens, a picture of a large bright white object is taken by the metalens. As shown in the right panel of Fig. 4, we derive a white image whose color is slightly reddish even applying the sensor built-in automatic white balance function with a 1.67:1:2.34 RGB gain. Wavelength-dependent meta-atom responses fundamentally constrain color channel balancing. The sharp exposure boundaries in the white image are defined by the aperture. Threshold-based morphological analysis can yield the morphological center coordinates of the circular area, which is (x_0, y_0). Cutting a line in the white image along $y = y_0$, we can have the spatial RGB light distribution, as shown in the left panel of Fig. 4. An attenuation exhibits a spatially radial pattern, where the edge regions demonstrate more significant attenuation than the center areas. The radial attenuation pattern stems from two compounding factors: 1) Meta-atom off-axis efficiency decay $\eta_{meta}(\theta)$, where nanostructures exhibit reduced light control capability at oblique angles, and 2) Geometric vignetting governed by the $\cos^4\theta$ law. Hence, we directly encode this optical prior at the spatial level to adjust the feature representation.

C. More Experimental Verifications

C.1. Real Scene Generalization

To verify the real-scene generalization, we capture real intestinal scenes using metalens. Due to the unavailability of in vivo experiments, which require additional biological approvals, an endoscopic education model is employed to simulate intestinal scenes. The model is positioned at varying distances (from 1 cm to 7 cm) from the optical system for imaging. As shown in Fig. 5 (Left), MetaScope demonstrates generalization to real scenes and diverse distances, a finding corroborated by [11]. Note that capturing real scenes precludes obtaining non-degraded ground truth, limiting model training and quantitative evaluation.

We further conduct biomedical verification by photographing the excised porcine intestine (see Fig. 5 (Right)).

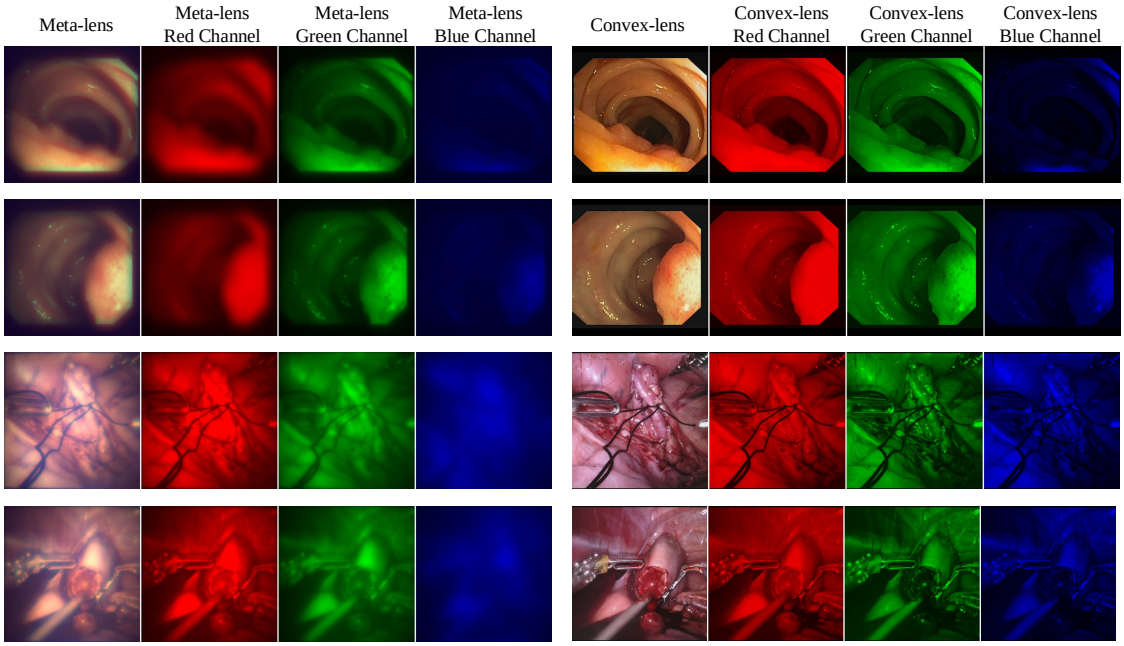

Figure 6. Visualization of the red, green, and blue channels of the samples from our metalens visual datasets.

Due to the inability to preserve biological tissues for extended periods, this experiment is conducted as a one-time trial. Our MetaScope achieves impressive achromatic correction results, revealing its superiority in clinical generalization. As clinical in vivo verification on humans requires strict ethical certifications, we are actively pursuing the necessary clearance.

C.2. Scaling Up Training with All Datasets

To further demonstrate the effectiveness of our algorithmic designs, we train a unified model using all captured data (five datasets) and compare it with the state-of-the-art method in metalens imaging, DRMI [33]. As shown in Tab. 2, our method consistently outperforms DRMI across all datasets, highlighting its superior scalability and generalization capabilities.

C.3. Generalization to Different Metalenses

We specifically focus on designing general metalenses that lack achromatic properties, thereby establishing a uniform degradation model. This model is not only relevant to various metalenses but also extends its applicability to diffractive optical elements (DOE) lenses, thereby broadening its scope and significance in optical research.

To validate our approach and dataset, we recaptured the CVC-Clinic dataset using a new metalens with a focal length of 5 mm. We then evaluated the performance of MetaScope on this dataset (Tab. 3). Remarkably, without any retraining, MetaScope achieved a satisfactory mIoU of 82.82%, which is comparable to the mIoU of 85.55% obtained with the conventional metalens. Furthermore, after fine-tuning, MetaScope demonstrated even superior performance, achieving a mIoU of 86.37%. These results highlight the significant value of the proposed datasets and the generality of our MetaScope.

C.4. Model Efficiency

We further assess the model size and inference speed. MetaScope demonstrates remarkable efficiency, featuring only 12M parameters and achieving an inference speed of 23.4 FPS. This performance significantly outperforms the state-of-the-art Mask2Former, which has 44M parameters and operates at 16.9 FPS. The lightweight architecture of MetaScope enables real-time processing capabilities, making it applicable in endoscopic surgery and diagnostics, where instantaneous imaging and analysis are crucial.

C.5. Dataset Analysis

Fig. 6 showcases metalens images alongside paired convex-lens images from our meticulously constructed datasets. Compared to convex-lens counterparts, metalens images exhibit noticeable variations in color and blurriness, primarily due to axial chromatic aberration inherent in the metalens design. To explore these in detail, we present separate visualizations for each red, green, and blue channel for both metalens and convex-lens configurations. This comparison reveals significant disparities in intensity degradation and dispersion: while the red and green channels show minimal degradation, the blue channel experiences considerable degradation across all samples. This phenomenon is attributed to the optimization of our metalens for a wavelength of 532 nm with a focal length of 10 mm, which aligns

more closely with red and green wavelengths and is less compatible with blue wavelengths (as illustrated in Fig. 3). These observations have driven the design of our OIA and OCC techniques, which enhance metalens image-based in vivo intelligence. This advancement highlights the superior performance and tailored design of our metalens system in mitigating chromatic aberrations, thereby significantly enhancing reliability for real-world clinical applications.

C.6. Metalens Imaging Segmentation

Fig. 7 presents a qualitative comparison of metalens imaging segmentation performance between our proposed MetaScope and state-of-the-art methods, including Rolling Unet [29], U-KAN [18], Mask2Former [10], and EDAFormer [40]. In abnormality segmentation tasks, as illustrated from row 1 to row 6, MetaScope consistently demonstrates superior accuracy in identifying polyps across various complex scenarios, excelling particularly in detecting small polyps (row 4) and accurately segmenting larger polyp areas (row 5). Furthermore, in the more challenging domain of surgical instrument segmentation, MetaScope continues to outperform current state-of-the-art methods by significantly enhancing the completeness of the ultrasound probe (highlighted in pink in row 7) and accurately delineating large needle drivers (highlighted in red in row 9). Remarkably, MetaScope is also capable of precisely identifying suction instruments (highlighted in yellow in row 8) that are only partially visible within the field of view. These results underscore MetaScope's robust capability to handle complex and partially obscured objects, demonstrating its strong potential and effectiveness for real-world practice.

C.7. Metalens Imaging Restoration

Fig. 8 provides a visual comparison of restoration quality between our proposed MetaScope and the latest methods, including SWinIR [25], MambaIR [12], and NeRD-rain [8]. MetaScope significantly outperforms these approaches by accurately restoring the structure and color of polyps (rows 1 to 3), capturing intricate vascular details (row 5), and maintaining the overall scene color fidelity (rows 4 and 6). In more complex surgical scenarios, MetaScope continues to surpass state-of-the-art methods, particularly excelling in rendering accurate colors (rows 7 and 9) and fine texture details (row 8) of surgical instruments. These superior capabilities highlight MetaScope's robust performance in handling diverse and challenging imaging conditions, demonstrating its strong potential and effectiveness for real-world medical applications.

C.8. Data Visualization

We present sample images from our constructed datasets, including Meta-CVC-Clinic, Meta-CVC-Colon, Meta-Kvasir-Seg, Meta-EndoVis-17, and Meta-EndoVis-18. As illustrated in Fig. 9, the images across these datasets exhibit similar meta-distortions. Additionally, our datasets offer extensive diversity, encompassing a wide range of in-vivo clinical scenarios that capture various pathological conditions and anatomical variations inherent to endoscopic procedures. This diversity enables comprehensive benchmarking of metalens imaging analysis in clinical in vivo settings, supporting a variety of research objectives, including generalization, transferability, and robustness.

D. Discussion and Clarifications

D.1. Dataset Information

The dataset details are as follows: **(1) CVC-Clinic** [5] is a publicly available dataset comprising 612 images extracted from 29 colonoscopy videos, each annotated with pixel-wise polyp masks. **(2) CVC-Colon** [4] consists of 380 polyp-annotated images sourced from 15 short colonoscopy video sequences. **(3) Kvasir-Seg** [14] is a gastrointestinal polyp segmentation dataset that includes 1,000 images and corresponding segmentation masks. **(4) EndoVis17** [2] consists of 1,800 frames with annotations for various surgical instrument types, enabling the analysis and recognition of instruments in minimally invasive surgeries. **(5) EndoVis18** [3] consists of 2,384 frames and features more complex porcine tissue and dynamic instrument movements with eight classes. Our metalens photoed datasets are named by adding the **Meta** prefix, such as *Meta-CVC-Clinic*. During dataset construction, we first calibrate the meta-camera to ensure that the photographed and original images are pixel-perfectly aligned [11]. Considering the clinical practice with insufficient illumination, we then generate the dataset under artificial light scenarios without sunlight to simulate endoscopic environments.

D.2. Physically Suitable for Endoscopy

To achieve accurate diagnosis and surgical operations, high-resolution imaging is essential for providing detailed visualization, such as identifying micro-lesions or vascular structures. Specifications for high-resolution imaging typically require a focal length of $f = 10 \sim 50$ mm, a field of view (FOV) of $10^\circ \sim 50^\circ$, and an f-number (F/#) of $3.5 \sim 5.6$. Our optical system, featuring a focal length of $f = 10$ mm, a field of view (FOV) of 31°, and an f-number (F/#) of 3.8, is fully suitable for endoscopy applications that demand high resolution for observing subtle tissues.

D.3. Implementation of the KL Divergence

The proof and implemented version of KL Divergence Loss L_{KL} (Eq. 7 in the main paper) is detailed as follows. Inspired by variational auto-encoders [16], the KL divergence term establishing the proxy offset latent feature $L_{\text{KL}} = \text{KL}\left(q_\phi(\boldsymbol{z} \mid X) | \mathcal{N}(\mathbf{0}, \mathbf{I})\right)$ used in the optics-informed dis-

persion correction module, is derived as follows:

$$
\begin{aligned}
&\mathrm{KL}\left(q_\phi(\boldsymbol{z} \mid X) | \mathcal{N}(\mathbf{0}, \mathbf{I})\right) \\
=&\int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log \frac{\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}} \, \mathrm{d}x \\
=&\int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \log \frac{1}{\sqrt{\sigma^2}} \times e^{\frac{x^2}{2} - \frac{(x-\mu)^2}{2\sigma^2}} \, \mathrm{d}x \\
=&\int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left[-\frac{1}{2}\log\sigma^2 + \frac{1}{2}x^2 - \frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right] \mathrm{d}x \\
=&\frac{1}{2}\int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \left[-\log\sigma^2 + x^2 - \frac{(x-\mu)^2}{\sigma^2}\right] \mathrm{d}x \\
=&\frac{1}{2}\left(-\log\sigma^2 + \mathbb{E}\left[x^2\right] - \frac{1}{\sigma^2}\mathbb{E}\left[(x-u)^2\right]\right) \\
=&\frac{1}{2}\left(-\log\sigma^2 + \sigma^2 + \mu^2 - 1\right),
\end{aligned}
$$

where μ and σ represent the mean and standard deviation of the latent space, respectively. These parameters are learned through two linear layers, $\boldsymbol{\mu}(X)$ and $\log \boldsymbol{\sigma}^2(X)$, as detailed in Section 4.2 of the main paper.

D.4. Related Work

Recent advancements in endoscopic image analysis have significantly enhanced diagnostic capabilities and surgical procedures [1, 13, 17, 19, 27, 28, 30, 31, 34, 39]. Some works focus on improving representation learning [9, 13, 21, 23, 34], removing surgical smoke [39], and integrating complementary modalities [28] to facilitate precise diagnosis and recognition during inspections and surgeries. Other works target accurate depth estimation [31] and pose estimation [30], and domain shifts [20, 22, 23] to enable autonomous navigation [24]. These studies, limited to convex-lens-based systems, hinder the potential for micro-in-vivo intelligence. In contrast, we explore metalens-based perception to advance micro-miniaturized in vivo diagnostics and surgery, fully exploring the inherent optics-driven insights for the methodology design.

D.5. Ethical Clarification

This research complies with all relevant ethical standards and regulations. The data used in this study were sourced from publicly available repositories [2–5, 14], ensuring that no identifiable personal information is included. All datasets used adhere to privacy laws and institutional guidelines governing the use of medical information. Additionally, the study does not involve direct interaction with human subjects, eliminating concerns about consent and participant welfare. We confirm that this work does not present ethical issues and aligns with the ethical principles required for medical AI research.

Figure 7. Qualitative comparison of state-of-the-art segmentation methods on Meta-CVC-Colon, Meta-Kvasir-Seg and Meta-EndoVis-18.

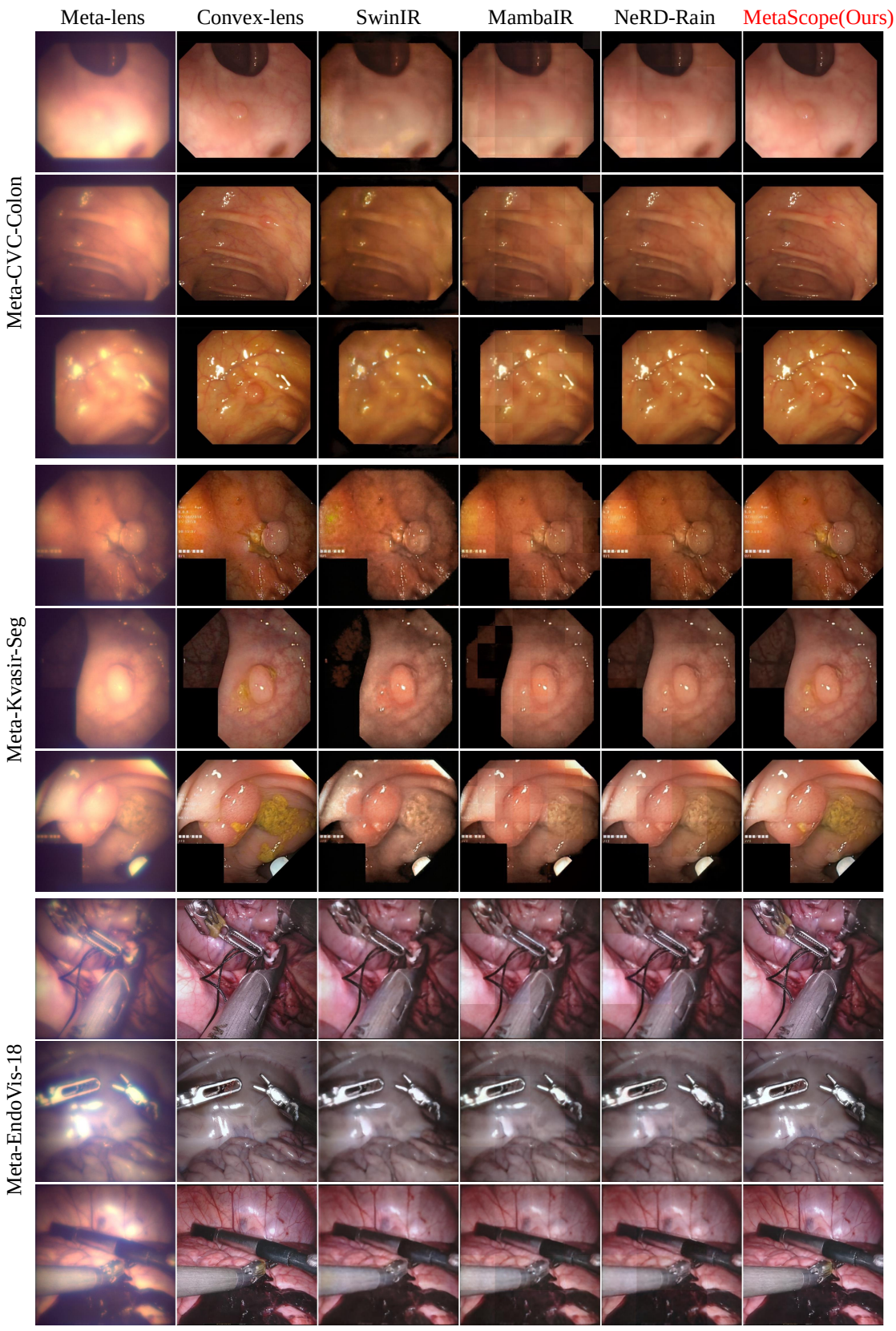

Figure 8. Qualitative comparison with state-of-the-art restoration methods on Meta-CVC-Colon, Meta-Kvasir-Seg and Meta-EndoVis-18.

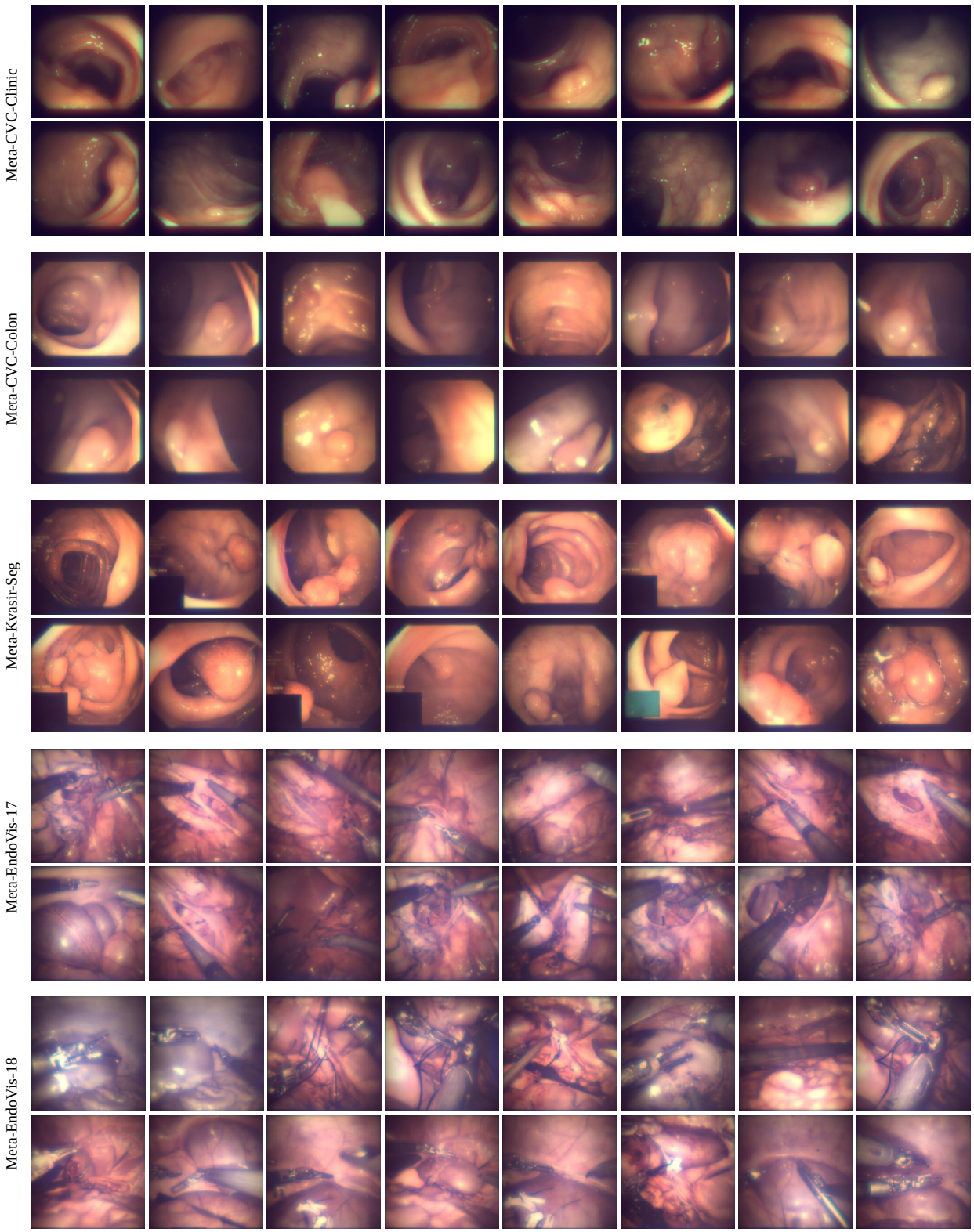

Figure 9. Visualization of samples from five Metalens Imaging datasets.

References

[1] Sharib Ali, Noha Ghatwary, Debesh Jha, Ece Isik-Polat, Gorkem Polat, Chen Yang, Wuyang Li, Adrian Galdran, Miguel-Ángel González Ballester, Vajira Thambawita, et al. Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. *Scientific Reports*, 14(1):2032, 2024. 7

[2] Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2017 robotic instrument segmentation challenge. *arXiv preprint arXiv:1902.06426*, 2019. 6, 7

[3] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020. 6

[4] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012. 6

[5] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015. 6, 7

[6] Max Born and Emil Wolf. *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013. 3

[7] Mu Ku Chen, Xiaoyuan Liu, Yanni Sun, and Din Ping Tsai. Artificial intelligence in meta-optics. *Chemical Reviews*, 122 (19):15356–15413, 2022. 3

[8] Xiang Chen, Jinshan Pan, and Jiangxin Dong. Bidirectional multi-scale implicit neural representations for image deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25627–25636, 2024. 6

[9] Zhen Chen, Wuyang Li, Xiaohan Xing, and Yixuan Yuan. Medical federated learning with joint graph purification for noisy label learning. *Medical Image Analysis*, 90:102976, 2023. 7

[10] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 6

[11] Yunxi Dong, Bowen Zheng, Hang Li, Hong Tang, Huan Zhao, Yi Huang, Sensong An, and Hualiang Zhang. Achromatic single metalens imaging via deep neural network. *ACS Photonics*, 11(4):1645–1656, 2024. 3, 4, 6

[12] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *European Conference on Computer Vision*, pages 222–241. Springer, 2024. 6

[13] Kai Hu, Ye Xiao, Yuan Zhang, and Xieping Gao. Multi-view masked contrastive representation learning for endoscopic video analysis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 7

[14] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*, pages 451–462. Springer, 2020. 6, 7

[15] Mohammadreza Khorasaninejad, Wei Ting Chen, Robert C Devlin, Jaewon Oh, Alexander Y Zhu, and Federico Capasso. Metalenses at visible wavelengths: Diffraction-limited focusing and subwavelength resolution imaging. *Science*, 352(6290):1190–1194, 2016. 1, 2

[16] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 6

[17] Chenxin Li, Hengyu Liu, Yifan Liu, Brandon Y Feng, Wuyang Li, Xinyu Liu, Zhen Chen, Jing Shao, and Yixuan Yuan. Endora: Video generation models as endoscopy simulators. In *International conference on medical image computing and computer-assisted intervention*, pages 230–240. Springer, 2024. 7

[18] Chenxin Li, Xinyu Liu, Wuyang Li, Cheng Wang, Hengyu Liu, and Yixuan Yuan. U-kan makes strong backbone for medical image segmentation and generation. *arXiv preprint arXiv:2406.02918*, 2024. 6

[19] Wuyang Li, Yang Chen, Jie Liu, Xinyu Liu, Xiaoqing Guo, and Yixuan Yuan. Joint polyp detection and segmentation with heterogeneous endoscopic data. In *3rd International Workshop and Challenge on Computer Vision in Endoscopy (EndoCV 2021): co-located with the 17th IEEE International Symposium on Biomedical Imaging (ISBI 2021)*, pages 69–79, 2021. 7

[20] Wuyang Li, Xinyu Liu, Xiwen Yao, and Yixuan Yuan. Scan: Cross domain object detection with semantic conditioned adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1421–1428, 2022. 7

[21] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma: Semantic-complete graph matching for domain adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5291–5300, 2022. 7

[22] Wuyang Li, Xiaoqing Guo, and Yixuan Yuan. Novel scenes & classes: Towards adaptive open-set object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15780–15790, 2023. 7

[23] Wuyang Li, Xinyu Liu, and Yixuan Yuan. Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9022–9040, 2023. 7

[24] Wuyang Li, Zhu Yu, and Alexandre Alahi. Voxdet: Rethinking 3d semantic occupancy prediction as dense object detection. *arXiv preprint arXiv:2506.04623*, 2025. 7

[25] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 6

[26] Ren Jie Lin, Vin-Cent Su, Shuming Wang, Mu Ku Chen, Tsung Lin Chung, Yu Han Chen, Hsin Yu Kuo, Jia-Wern Chen, Ji Chen, Yi-Teng Huang, et al. Achromatic metalens array for full-colour light-field imaging. *Nature nanotechnology*, 14(3):227–231, 2019. 3

[27] Hengyu Liu, Yifan Liu, Chenxin Li, Wuyang Li, and Yixuan Yuan. Lgs: A light-weight 4d gaussian splatting for efficient surgical scene reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 660–670. Springer, 2024. 7

[28] Tianyi Liu, Shuaishuai Zhuang, Jiacheng Nie, Geng Chen, Yusheng Guo, Guangquan Zhou, Jean-Louis Coatrieux, and Yang Chen. A rotation-invariant texture vit for fine-grained recognition of esophageal cancer endoscopic ultrasound images. In *European Conference on Computer Vision*, pages 360–377. Springer, 2025. 7

[29] Yutong Liu, Haijiang Zhu, Mengting Liu, Huaiyuan Yu, Zihan Chen, and Jie Gao. Rolling-unet: Revitalizing mlp's ability to efficiently extract long-distance dependencies for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3819–3827, 2024. 6

[30] Yiyao Ma, Kai Chen, Hon-Sing Tong, Ruofeng Wei, Yui-Lun Ng, Ka-Wai Kwok, and Qi Dou. Shape-guided configuration-aware learning for endoscopic-image-based pose estimation of flexible robotic instruments. In *European Conference on Computer Vision*, pages 259–276. Springer, 2025. 7

[31] Akshay Paruchuri, Samuel Ehrenstein, Shuxian Wang, Inbar Fried, Stephen M Pizer, Marc Niethammer, and Roni Sengupta. Leveraging near-field lighting for monocular depth estimation from endoscopy videos. In *European Conference on Computer Vision*. Springer, 2025. 7

[32] Pierre Pellat-Finet. Fresnel diffraction and the fractional-order fourier transform. *Optics Letters*, 19(18):1388–1390, 1994. 3

[33] Joonhyuk Seo, Jaegang Jo, Joohoon Kim, Joonho Kang, Chanik Kang, Seongwon Moon, Eunji Lee, Jehyeong Hong, Junsuk Rho, and Haejun Chung. Deep-learning-driven end-to-end metalens imaging. *arXiv preprint arXiv:2312.02669*, 2023. 3, 5

[34] Vinkle Srivastav, Nassir Navab, Nicolas Padoy, et al. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 7

[35] Ethan Tseng, Shane Colburn, James Whitehead, Luocheng Huang, Seung-Hwan Baek, Arka Majumdar, and Felix Heide. Neural nano-optics for high-quality thin lens imaging. *Nature communications*, 12(1):6493, 2021. 3

[36] Ming Lun Tseng, Hui-Hsin Hsiao, Cheng Hung Chu, Mu Ku Chen, Greg Sun, Ai-Qun Liu, and Din Ping Tsai. Metalenses: advances and applications. *Advanced Optical Materials*, 6(18):1800554, 2018. 2

[37] Akira Ueno, Juejun Hu, and Sensong An. Ai for optical metasurface. *npj Nanophotonics*, 1(1):36, 2024. 3

[38] Shuming Wang, Pin Chieh Wu, Vin-Cent Su, Yi-Chieh Lai, Mu-Ku Chen, Hsin Yu Kuo, Bo Han Chen, Yu Han Chen, Tzu-Ting Huang, Jung-Hsi Wang, et al. A broadband achromatic metalens in the visible. *Nature nanotechnology*, 13(3): 227–232, 2018. 3

[39] Renlong Wu, Zhilu Zhang, Shuohao Zhang, Longfei Gou, Haobin Chen, Lei Zhang, Hao Chen, and Wangmeng Zuo. Self-supervised video desmoking for laparoscopic surgery. In *European Conference on Computer Vision*, pages 307–324. Springer, 2025. 7

[40] Hyunwoo Yu, Yubin Cho, Beoungwoo Kang, Seunghun Moon, Kyeongbo Kong, and Suk-Ju Kang. Embedding-free transformer with inference spatial reduction for efficient semantic segmentation. In *European Conference on Computer Vision*, pages 92–110, 2025. 6

[41] Maksym V Zhelyeznyakov, Steve Brunton, and Arka Majumdar. Deep learning to accelerate scatterer-to-field mapping for inverse design of dielectric metasurfaces. *ACS Photonics*, 8(2):481–488, 2021. 3