This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Wuyang Li[1,3,*] Wentao Pan[1,*] Xiaoyuan Liu[2,3,*] Zhendong Luo[2] Chenxin Li[1]
Hengyu Liu[1] Din Ping Tsai[2,†] Mu Ku Chen[2,†] Yixuan Yuan[1,†]
[1]The Chinese University of Hong Kong, [2]City University of Hong Kong, [3]EPFL
Project Page: https://cuhk-aim-group.github.io/MetaScope/

Abstract

Miniaturized endoscopy has advanced accurate visual perception within the human body. Prevailing research remains limited to conventional cameras employing convex lenses, where the physical constraints with millimetre-scale thickness impose serious impediments on the micro-level clinical. Recently, with the emergence of meta-optics, ultra-micro imaging based on metalenses (micron-scale) has garnered great attention, serving as a promising solution. However, due to the physical difference of metalens, there is a large gap in data acquisition and algorithm research. In light of this, we aim to bridge this unexplored gap, advancing the novel metalens endoscopy. First, we establish datasets for metalens endoscopy and conduct preliminary optical simulation, identifying two derived optical issues that physically adhere to strong optical priors. Second, we propose MetaScope, a novel optics-driven neural network tailored for metalens endoscopy driven by physical optics. MetaScope comprises two novel designs: Optics-informed Intensity Adjustment (OIA), rectifying intensity decay by learning optical embeddings, and Optics-informed Chromatic Correction (OCC), mitigating chromatic aberration by learning spatial deformations informed by learned Point Spread Function (PSF) distributions. To enhance joint learning, we further deploy a gradient-guided distillation to transfer knowledge from the foundational model adaptively. Extensive experiments demonstrate that MetaScope not only outperforms state-of-the-art methods in both metalens segmentation and restoration but also achieves impressive generalized ability in real biomedical scenes.

1. Introduction

Miniaturized endoscopy provides reliable internal visual perception within the human body (Fig. 1a). Recent advancements in computer vision have significantly enhanced real-time clinical support, such as segmenting the gastrointestinal tract. This is vital for the early detection of serious diseases, aiding medical professionals in surgical decisions.

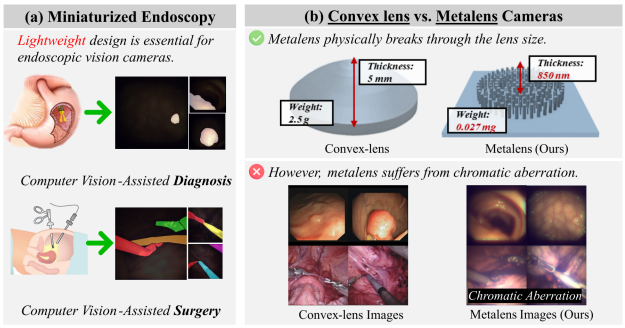


Figure 1. (a) Miniaturized endoscopy; (b) Comparing convex-lens and our meta-lens about physical nature and imaging quality.

To facilitate endoscopic diagnosis and surgery [57, 67, 72, 73], lots of efforts have been made to develop more robust, accurate, and efficient solutions for disease recognition and segmentation. Some works [1, 9, 32, 41, 50, 56] focus on data engineering by enhancing diversity, quality, and scale. Some other works [2, 23, 39, 55, 92, 95] develop more advanced deep-learning models by addressing numerous algorithmic challenges, including domain generalization [2, 46, 47], privacy preservation [17, 35, 58], and label-efficient learning [48, 60, 88], among others.

Despite significant algorithmic advancements, existing studies are limited to conventional cameras that utilize convex lenses [44, 52], whose physical constraints pose substantial barriers to micro-level clinical applications. The inherent curvature limitations of convex lenses (Fig. 1b left) hinder the development of ultra-lightweight profiles, restricting them to millimetre-scale thickness. This impairs navigation through narrow or curved anatomical structures and negatively impacts device size, power consumption, and operational efficiency, hindering progress toward micro-miniaturized clinical applications.

Recently, with the emergence of meta-optics [13, 36, 69, 70, 82], ultra-micro imaging based on metalenses has garnered significant attention in micro-robotics, showing a

*Equal Contribution. †Corresponding Authors.

promising solution [19, 66]. As shown in Fig. 1b, these ultra-thin lenses utilize nanostructured arrays to manipulate light at subwavelength scales. In contrast to conventional in-vivo lenses, which typically weigh around **2.5 grams** with **5-millimetre** thickness, meta-lenses weigh less than **0.05 milligrams** and measure under **850-nanometer** in thickness. These superior physical properties significantly reduce the device size [59, 61, 63]. This enables access to narrow anatomical structures and facilitates minimally invasive technologies within the human body.

While promising, these advancements pose challenges in maintaining imaging quality, as severe chromatic aberration [14, 52, 86] fundamentally limits their diagnostic reliability by distorting spectral signatures critical for tissue characterisation. The key challenge lies in achieving achromatic design, where the main difficulty stems from the wavelength-dependent phase modulation of nanostructures. Their resonant properties and material dispersion make it physically challenging to achieve consistent wavefront control across broad spectra, causing different wavelengths of light to focus at different points. Current physical solutions typically stack multiple layers [43] or frequency-dependent phase engineering combining geometric-phase and propagation-phase elements [15], but these face limitations in fabrication complexity, limited bandwidth, and trade-offs between efficiency and device size. Hence, *we propose the computer vision solution for computationally achromatic meta-lens endoscopy (refer to Sec. 3.1),* underscoring its substantial industrial and academic value [82].

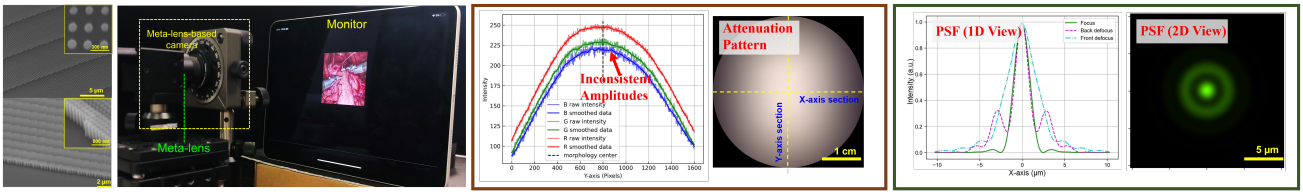
To address this challenge, we delve into the nature of chromatic aberration with optical simulation, encapsulating it into two sub-issues (Sec. 3.2). **(1) Color and Position-Dependent Intensity Loss**: Different colors are attenuated unevenly with inherent optical prior. **(2) Color-Dependent Blur and Misalignment**: Different color has a varied "blur kernel" (Point Spread Function, PSF). Mismatched color offsets create color fringing around objects, especially at high-contrast edges. Based on these observations, we further propose **MetaScope**, a novel optics-driven network to eliminate chromatic aberration and deliver precise segmentation for clinical applications. MetaScope employs a metalens encoder to correct distorted representations via Optics-informed Intensity Adjustment (OIA) and Optics-informed Chromatic Correction (OCC). Additionally, it incorporates a dual-branch decoder designed for accurate segmentation, enhanced by an auxiliary restoration branch (see Fig. 3). Specifically, with optical simulations, OIA encodes optical priors into optical embeddings and then adjusts the attention map. OCC estimates PSF distributions by learning implicit distributions adaptively, eliminating chromatic aberrations through spatial aggregation. Furthermore, to enhance the robustness against chromatic aberration, we introduce gradient-informed distillation, which transfers foundational knowledge to enhance restored representation. *By addressing both hardware and algorithmic barriers, we make clinical micro-miniaturization possible with enhanced physical efficiency.* The contributions are fivefold:

- We are the first to break through the endoscopy barrier via metalenses from both **data and algorithmic** aspects, illuminating ultra-micro applications of healthcare robotics.
- We propose five datasets for metalens-based diagnostics and surgery, starting and advancing the clinical research.
- Delving into two imaging challenges with optical simulations, we propose a novel MetaScope to achieve joint restoration and segmentation guided by optical priors.
- The proposed MetaScope consists of an Optics-informed Intensity Adjustment to correct intensity attenuation with encoded optical embedding and an Optics-informed Chromatic Correction to rectify chromatic aberration. Besides, a gradient-informed distillation is proposed to enhance joint restoration-aware learning.
- Extensive experiments verify that MetaScope achieves SoTA segmentation and restoration, which also works impressively in real biomedical scenarios (Supp. C).

2. Related Works

Metalens Imaging uses optical metasurfaces to enhance the imaging capabilities of cameras [36, 69, 70, 82], which consist of arrays of nano-scale structures for precise focusing control, as shown in Fig. 2a. Their lightweight design and wide fields of view, *make them highly suitable for applications in augmented reality, micro-robotics, and in-vivo healthcare* [19, 66]. However, manufacturing imperfections and inherent limitations lead to image distortion with severe chromatic aberration [83]. To break through this, artificial intelligence has been increasingly integrated into meta-lens imaging to facilitate meta-lens design [5, 33, 78, 87, 96] and enhance image quality through deep models [22, 29, 80, 81]. Unlike existing studies focusing on restoration, we are the first to explore micro-miniaturized endoscopic applications, proposing a unified perception model driven by optical priors.

Medical Image Segmentation [6, 40, 62] autonomously delineates intricate anatomical structures, streamlining clinical procedures. With the U-Net architecture [79], numerous variants [10, 20, 27, 30, 45, 54, 76, 77, 85, 91] have been developed to enhance multi-scale fusion, skip connections, and bottleneck features [30]. Recently, U-shaped models such as UNeXt [84], DC-UNet [65], Rolling-Unet [64], U-Mamba [68], and U-KAN [42] have improved model efficiency and interoperability. Additionally, generic segmentation methods have shown remarkable efficacy in medical applications, such as Mask2Former [18], EDAformer [93], and SegNext [25]. Unlike traditional methods that rely on high-fidelity images, our approach targets the more challenging metalens images via optical observations.



(a) Meta-lens (b) Meta-lens dataset establishment (c) Color and position-dependent intensity loss (d) Color-dependent kernel - PSF

Figure 2. (a) Our metalens nanostructure via scanning electron microscope; (b) Optical setup for meta-lens data collection; (c) Intensity loss in metalens imaging with color and spatial dependence; (d) Color-dependent blur and misalignment due to inconsistent PSF patterns.

Image Restoration [11, 21, 37, 51, 53] seeks to reconstruct high-quality images from degraded inputs, such as rainy, foggy, etc [16, 93]. Chen *et al.* [11] proposes a comprehensive study of restoration-oriented CNN design and introduces NAFNet, a lightweight model for efficient image restoration. With the scalability of the Transformers architecture, SwinIR [51] employs shifted-window attention to enhance global context while preserving local details. Additionally, some methods, such as Restormer [94], SFHformer [34], and MambaIR [24] have gained attention due to their novel network architectures. Instead of focusing on the natural restoration with convex lenses, we investigate advanced meta-lens imaging driven by distinct optical simulation and explore its potential to assist perception tasks.

3. Preliminaries

Metalens Imaging. Conventional lenses achieve light focusing by varying thickness in a convex geometry and using multiple lens elements, which poses challenges for micron-miniaturized practice due to the bulky optical systems. In contrast, **metalenses** employ nanoscale structures, meta-atoms patterned on a flat surface (Fig. 2a) to manipulate the light wavefront in a much more compact way. Each meta-atom provides specific amplitude and phase modulation. The focusing phase profile can be expressed as:

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

where $\phi(r,\lambda)$ is the required phase for the meta-atom at position r for wavelength λ, and f is the desired focal length. Without achromatic design, metalenses exhibit wavelength-dependent focal lengths, expressed as $f(\lambda)$. *This variation prevents different wavelengths from converging on a single imaging plane.* We designed and fabricated polarization-independent gallium nitride (GaN) metalens with a diameter of 2.6 mm and a focal length of 10 mm.

3.1. Metalens Dataset Establishment

Due to the limited imaging quality of metalens cameras, a significant barrier exists to clinical practice, resulting in insufficient endoscopic data collection. This scarcity of data impedes the design of algorithms aimed at correcting chromatic aberration. To overcome this, we employ a proxy photographic solution to establish datasets, as shown in Fig. 2b. *As the chromatic aberration patterns adhere to consistent physical principles, we can create paired metalens mapping data with and without distortion to train the model that approximates the restoration function.* Based on this insight, we construct the necessary pairs by projecting images from publicly available endoscopic datasets onto a screen and capturing images with a metalens camera. Thus, we can generate the metalens dataset with a triple data format, including the metalens image, associated high-quality ground truth, and the corresponding semantic labels, such as segmentation. Note that we empirically find that the model trained on our datasets exhibits **generalization capacity to real scenes with varied depth and different metalenses**, demonstrating significant value to the community. More details can be found in the Supp C.

3.2. Identifying Metalens Imaging Challenges

The challenge of metalens images arises from the wavelength dependence of the focal length, which fails to focus all light onto a single focal plane. To explore this unique characteristic, we conduct optical simulations and identify two derived issues, motivating our method designs.

(1) Color and Position-Dependent Intensity Loss. We illuminate the metalens with the directional white light of uniform intensity while the CMOS sensor applies gains of 1.67, 1, and 2.34 to RGB channels (built-in white balance function). We analyze the resulting intensity variations in the captured images, as shown in Fig. 2c, summarizing two observations: (a) Imbalanced RGB Channel Intensities: Even under the automatic white balance, color channels exhibit inconsistent brightness due to wavelength-dependent responses of meta-atoms. Specifically, non-optimized wavelengths (e.g., red/blue) suffer from simultaneous amplitude and phase mismatch in meta-atoms, significantly reducing modulation efficiency and causing light intensity loss. (b) Vignetting and Structured Noise: The surrounding regions experience significantly greater attenuation compared to the central area, which is analogous to the *dark channel* observed in de-hazing [26]. This is because of the superposition of off-axis efficiency drop of nanostructures $\eta_{\text{meta}}(\theta)$ and traditional geometric vignetting ($\cos^4\theta$ law), which could be approximated as $I_{\text{total}}(\theta) =$

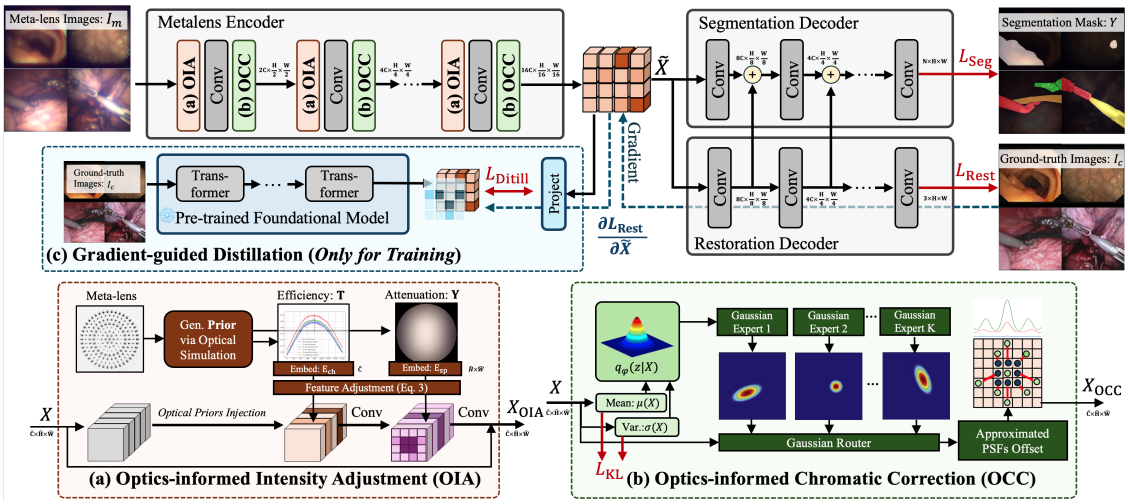


Figure 3. Illustration of the proposed MetaScope, which consists of two optics-informed modules, named (a) OIA and (b) OCC injected in the encoder to correct the meta-distortion, and a (c) gradient-guided distillation strategy to facilitate the restoration-aware representation.

$\eta_{\text{meta}}(\theta) \cdot \cos^4\theta \cdot I_0$. Inhomogeneous amplitude responses of different meta-atoms and nanofabrication defects induce local random fluctuations in transmitted light intensity, manifesting as spatially correlated brightness noise. ***Motivation:*** To fully leverage observed optical priors, we propose to encode the efficiency and radial priors to adjust the feature learning with restoration awareness (Sec. 4.1).

(2) Color-Dependent Blur and Misalignment. PSF is the imaging response of an optical system to an ideal point light source, directly governing spatial resolution and color fidelity, which could be approximated as $I_{\text{output}} = \sum_\lambda \left[T(\lambda) \cdot \left(\eta_{\text{meta}}(\theta) \cdot \cos^4\theta \cdot I_0(\lambda) * \text{PSF}(\lambda, x, y)\right)\right] + \epsilon$, where ϵ is noise. Fig. 2d shows the simulated PSFs of a coherent plane wave (ideal light source) under the wavelength of 532 nm (green). In metalenses, chromatic aberrations cause wavelength-dependent PSF variations. Each wavelength (color) has a different focal length for a sensor fixed at the design wavelength (e.g., green), shorter (red) and longer (blue) wavelengths defocus, broadening their PSFs. Blue/red edges appear more blurred than green ones, akin to applying color-specific Gaussian kernels. Imperfect phase gradient compensation of the metalens results in PSF centroid displacement across wavelengths. Hence, high-contrast edges exhibit color fringing, as misaligned RGB PSFs create artificial chromatic gradients. Our observations indicate that such blur and fringing cause the signal to spread from the source pixel into adjacent pixels. ***Motivation:*** To tackle this issue, we propose a novel solution by estimating the full-color PSF **distribution** via learning inherent Gaussian mixtures adaptively, thereby spatially guiding correction of chromatic aberrations (Sec. 4.2).

Problem Formulation. We have triplet data $\{I_m, I_c, Y\}$ for training, including degraded images captured by metalens I_m, the associated high-quality ones captured from the convex lens I_c, and the segmentation mask Y. Our goal is to train a unified model $\mathcal{F}$ that can achieve reliable perception $\tilde{Y}$ and restoration I_c using metalens images I_m. In the test stage, the model can predict restored images and masks directly from metalens images: $\{\tilde{Y}, \tilde{I}_c\} = \mathcal{F}(I_m)$.

4. Methodology

As shown in Fig. 3, given metalens images I_m, we employ a metalens encoder $\mathcal{F}_{\text{enc}}$, which consists of two key modules, (a) OIA and (b) OCC, to generate the restoration-aware latent $\tilde{X}$, addressing the two issues in chromatic aberration (Sec. 3.2) with optical priors. Then, $\tilde{X}$ is sent to a dual-branch decoder $\mathcal{F}_{\text{dec}}$ to generate both the restored image $\tilde{I}_m$ and the segmentation mask $\tilde{Y}$. In this process, the intermediate feature of meta-restoration (low-level) guides the segmentation branch with a simple residual connection. During the training, we implement (c) gradient-guided distillation that encodes the convex-lens image I_c with the foundation model to facilitate the learning of restoration.

4.1. Optics-informed Intensity Adjustment

Metalens images experience severe attenuation in adhering to specific priors. However, the limited data scale makes it challenging to learn these priors solely through data fitting. To overcome this, we aim to adjust the feature using the optical priors derived from simulations, as shown in Fig. 3a.

Optical Prior Encoding. We begin by sending the metalens image I_m through a metalens encoding block to obtain the feature X. Due to inconsistent wavelengths, the captured images I_m exhibit varying intensity attenuation with optical evidence. To tackle this, we use the optical simulation to generate two types of optical priors (see Sec. 3.2): the color-depended vector $\mathbf{T} \in \mathbb{R}^L$ encoding the efficiency of L visible light in sequential wavelengths, and spatial at-

tenuation prior $\mathbf{Y} \in \mathbb{R}^{H\times W}$. Note that those priors can be generalized to different metalens due to the similar attenuation patterns. Then, we encode them into the optical embedding at the channel $\mathbf{E}_{\text{ch}} \in \mathbb{R}^{\hat{C}}$ and spatial $\mathbf{E}_{\text{sp}} \in \mathbb{R}^{\hat{H}\times\hat{W}}$ levels to calibrate the input feature $X \in \mathbb{R}^{\hat{C}\times\hat{H}\times\hat{W}}$:

$$\mathbf{E}_{\text{ch}} = \mathcal{F}_{\text{Fc}}^{\text{Enc}}\left(\mathbf{T}\right); \mathbf{E}_{\text{sp}} = \mathcal{F}_{\text{Conv}}^{\text{Enc}}\left(\mathbf{Y} \oplus \mathbf{C}_x \oplus \mathbf{C}_y\right), \quad (2)$$

where $\mathbf{C}_{x/y} \in \mathbb{R}^{H\times W}$ indicate the normalized coordinates, $\oplus$ is the concatenation, and $\mathcal{F}_{\text{Fc}}^{\text{Enc}}/\mathcal{F}_{\text{Conv}}^{\text{Enc}}$ are simple fully connected and convolutional based layers, abbreviated as `Conv/Fc` in following sections. These optical embeddings provide significant optical guidance to correct the intensity attenuation at the spatial and channel levels.

Feature Adjustment with Optical Prior. To integrate optical prior, a natural idea is to inject embedding into the network like diffusion models [28]. However, instead of sharing across all features, metalens images introduce attenuation patterns exhibiting variability across wavelengths and positions. Hence, we consider the distinct physical properties of efficiency (color-based) and spatial priors (position-based) and propose a simple yet effective method that uses optical embedding to rectify channel and spatial attention. Specifically, the adjusted attention is integrated as follows:

$$\begin{aligned}\text{Attn}_{\text{ch}}(X) &= \text{Sigmoid}\left(\mathcal{F}_{\text{Fc}}^{\text{Proj}}\left(\text{Mean}_{\text{sp}}(X) + \mathbf{E}_{\text{ch}}\right)\right), \\ \text{Attn}_{\text{sp}}(X) &= \text{Sigmoid}\left(\mathcal{F}_{\text{Conv}}^{\text{Proj}}\left(\text{Mean}_{\text{ch}}(X) + \mathbf{E}_{\text{sp}}\right)\right),\end{aligned} \quad (3)$$

where $\text{Mean}_{\text{sp/ch}}$ indicates the averaging over the spatial and channel dimension, and $\mathcal{F}_{\text{Fc/Conv}}^{\text{Proj}}$ is a projection layer. Then, given the feature X, the adjusted one can be represented as: $X_{\text{OIA}} = \texttt{Conv}(\text{Attn}_{\text{sp}}(X)\odot(\texttt{Conv}(\text{Attn}_{\text{ch}}(X)\odot X)))+X$, where $\odot$ is Hadamard product. Thus, by embedding the optical prior, the model can dynamically adjust representations based on the color-specific and spatially varying optical prior, improving the restoration-aware encoding.

4.2. Optics-informed Chromatic Correction

Metalens images suffer from chromatic artifacts due to wavelength-specific PSFs. Considering the difficulty in estimating explicit PSFs. we model the implicit PFS distribution via Gaussian Mixture Models (GMM) in an adaptive manner for chromatic correction (Fig. 3b).

Formulating Implicit PSF Distribution. Given an input light with wavelength λ, metalens projects it to the 2-D image at a specific position $\mathbf{p} = [x, y]$ coupled with dispersed components to the neighbouring pixels, which can be formulated as the spatial dispersion $\Delta\mathbf{p} = [\Delta x, \Delta y]$, adhering the implicit PSF distribution $P(\Delta\mathbf{p})$. As shown in Fig. 2 (d), such dispersion can be naturally modeled as a mixture of Gaussians with a main lobe accompanied by symmetric side lobes. Based on this observation, we formulate the implicit PSF distribution with GMM as follows,

$$P(\Delta\mathbf{p}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\Delta\mathbf{p} \middle| \mu_{\mathbf{p},k}, \sigma_{\mathbf{p},k}^2\right), \quad (4)$$

where $P(\Delta\mathbf{p})$ represents the dispersion probability, $\mathbf{p} = [x; y]$ is the 2-D position, π_k is the coefficient for the k-th Gaussian parameterized with mean $\mu_{\mathbf{p},k}$ and variance $\sigma_{\mathbf{p},k}^2$.

Learning PSF Distribution via Multi-Expert Gaussian. Based on the GMM formulation, we aim to learn a spatial deformation $\Delta\mathbf{p}$ for each position $\mathbf{p}$ to model PSFs, adhering $P(\Delta\mathbf{p})$. This can thereby facilitate the adaptive aggregation of dispersed information. However, directly estimating the GMM in Eq. 4 with conventional methods like the EM algorithm is time-consuming and challenging since each wavelength λ is associated with a specific GMM with diverse patterns. To break through this barrier, inspired by the Mixture of Experts (MoE) [71], we propose a Multi-expert Gaussian (MeG) paradigm with K Gaussian experts, where each expert MeG_k serves for a specific Gaussian pattern. Notably, instead of directly estimating the Gaussian parameters, the expert is designed to estimate the linear transformation from a normal Gaussian, i.e., *residual* part, based on the scaling invariance of the Gaussian.

Specifically, given the input feature X, we first assume a normally distributed latent $P(z|X)$ with variational autoencoding for the PSF distributed offset space. To this end, we deploy two linear layers for the mean $\boldsymbol{\mu}(X)$ and log-variance $\log\boldsymbol{\sigma}^2(X)$ as Gaussian parameters and implement Kullback-Leibler divergence to ensure that offset space to be Gaussian: $\text{KL}\left(q_\phi(\boldsymbol{z} \mid X)\|\mathcal{N}(0, \mathbf{I})\right)$ as [38], where q_ϕ is simply formulated with the former *Conv* block [49]. Then, we establish K experts by leveraging the linear invariance property of Gaussian distributions, where each of the experts parameterized with $\{\mathbf{a}_k, \mathbf{b}_k\}$ aims to learn a linear projection based on a sampled $z \sim P(z|X)$ using the reparameterization trick [38]. Thus, this can generate a transformed Gaussian for each expert: $\text{MoG}_k(X) \sim \mathcal{N}(\mathbf{a}_k\boldsymbol{\mu}(X) + \mathbf{b}_k, \mathbf{a}_k^2\boldsymbol{\sigma}^2(X))$. Concretely, each MeG is deployed with a convolution layer outputting $2\times M\times M$ channels where $M = 3$, which generates M^2 offset candidates $[\Delta x; \Delta y]$ modeling the implicit PSFs[1]. Finally, a small router layer $\text{Router}(X)$ is deployed with `Conv−Pooling` to learn the aggregation weight with K-channel output. Thus, the whole process can be formally represented as,

$$\Delta\mathbf{p} = \sum_{k=1}^{K} \pi_k \cdot \text{MeG}_k(X), \text{where } \pi = \text{Router}(X), \quad (5)$$

and $\Delta\mathbf{p}$ is the estimated spatial offsets in modeling dispersion patterns. These offsets are utilized to aggregate dispersed points with the well-estimated guidance of PSFs.

[1]To avoid the collapse to the same Gaussian and facilitate learning orthogonally, we deploy the MeG layers with different dilation rates.

Metalens Segmentation												
	In-Vivo Diagnosis						In-Vivo Surgery					
	Meta-CVC-Clinic		Meta-CVC-Colon		Meta-Kvasir-Seg		Meta-EndoVis-17		Meta-EndoVis-18		**Average**	
Method	mIoU	mDICE	mIoU	mDICE	mIoU	mDICE	mIoU	mDICE	mIoU	mDICE	mIoU	mDICE
UNet [79]	0.6623	0.7710	0.6238	0.7306	0.4814	0.6171	0.7174	0.8274	0.5931	0.7181	0.6156	0.7328
Atten-UNet [74]	0.6705	0.7752	0.6336	0.7306	0.4884	0.6255	0.7040	0.8163	0.5895	0.7200	0.6172	0.7335
UNet++ [97]	0.6765	0.7770	0.6417	0.7391	0.4746	0.6084	0.7195	0.8287	0.6003	0.7296	0.6225	0.7366
UNext [84]	0.7006	0.7948	0.6855	0.7707	0.5504	0.6651	0.7290	0.8377	0.5235	0.6496	0.6378	0.7436
Rolling-UNet [64]	0.7499	0.8259	0.6597	0.7460	0.5376	0.6482	0.7978	0.8847	0.6092	0.7359	0.6708	0.7681
U-KAN [42]	0.7454	0.8260	0.7191	0.7924	0.5929	0.6937	0.7753	0.8692	0.6482	0.7766	0.6962	0.7916
DeepLabV3+ [12]	0.6574	0.7506	0.6121	0.6943	0.4629	0.5882	0.7887	0.8784	0.6749	0.7956	0.6392	0.7414
UperNet [89]	0.6312	0.7201	0.7159	0.8041	0.5839	0.6910	0.8038	0.8810	0.6774	0.7999	0.6824	0.7792
Segformer [90]	0.7612	0.8378	0.6863	0.7729	0.4739	0.5970	0.8025	0.8878	0.6745	0.7965	0.6797	0.7784
EDAformer [93]	**0.8169**	**0.8859**	**0.7507**	**0.8312**	0.5705	0.6813	0.8161	0.8965	**0.7310**	**0.8380**	0.7370	0.8266
SegNext [25]	0.7639	0.8399	0.6528	0.7352	0.5790	0.6803	0.8207	0.8994	0.6266	0.7555	0.6886	0.7821
Mask2former [18]	0.7912	0.8612	0.7353	0.8157	**0.6664**	**0.7558**	**0.8367**	**0.9010**	0.7300	0.8371	**0.7519**	**0.8342**
MetaScope (Ours)	**0.8555**	**0.9137**	**0.7868**	**0.8604**	**0.6972**	**0.7845**	**0.8802**	**0.9355**	**0.8057**	**0.8896**	**0.8051**	**0.8767**

Metalens Restoration												
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SwinIR [51]	25.3100	0.8442	24.1285	0.8338	23.7643	0.8328	26.1958	0.9143	25.2748	0.8786	24.9347	0.8607
Restormer [94]	33.5640	0.9581	33.9541	**0.9719**	31.3738	0.9778	30.9982	0.9561	30.6164	0.9356	32.1013	0.9599
NAFNet [11]	**33.8970**	0.9668	34.6834	**0.9856**	31.5939	**0.9793**	30.9660	0.9553	30.9087	0.9397	32.4098	**0.9653**
DRMI [80]	33.5517	**0.9784**	**34.7755**	0.9672	**31.6620**	0.9705	**31.6709**	**0.9601**	**31.4767**	**0.9451**	**32.6274**	0.9643
MambaIR [24]	25.5517	0.8689	26.9477	0.9128	24.7844	0.9165	24.1641	0.8852	24.3246	0.8572	25.1545	0.8881
NeRD-Rain [16]	30.3351	0.9370	31.1423	0.9558	28.3791	0.9564	25.7874	0.9072	25.4942	0.8712	28.2276	0.9255
MetaScope (Ours)	**34.8483**	**0.9854**	**35.6603**	**0.9856**	**32.4454**	**0.9818**	**32.0365**	**0.9619**	**31.8664**	**0.9474**	**33.3714**	**0.9724**

Table 1. Comparison with state-of-the-art methods on five benchmarks in terms of metalens segmentation (top) and restoration (bottom). The best and second-best results are highlighted in red and blue, respectively.

Chromatic Correction. With the learned offsets, we further aggregate the spatially dispersed points with a simple deformable convolution layer, facilitating local aggregation of dispersed information. This can ensure that the pixel colors are accurately reconstructed by aligning with the underlying physical distribution modeled by the GMM, denoted as $X(\mathbf{p})_{\text{OCC}} = \sum_{i=1}^{M+N} w_i \cdot X(\mathbf{p}+\Delta\mathbf{p}_i)$, where $\Delta\mathbf{p}_i$ represents the learned offset for the i-th element ($i \in [1, M \times M]$) at location $\mathbf{p}$, and w_i is the learnable weight. By learning the positions of the sampled points, we can effectively mitigate color dispersion caused by PSFs, leading to more accurate and visually consistent color correction.

4.3. Optimization

Gradient-guided Distillation. To better optimize the proposed optics-informed modules in chromatic aberration learning, we propose a distillation method that leverages the pre-trained DINO v2 [75] to refine under-optimized regions of $\tilde{X}$, as illustrated in Fig. 3c. Given the feature $\tilde{X}$, we back-propagate the restoration Loss L_{Rest} and collect the gradient amplitude map $\mathbf{M} = |\frac{\partial L_{\text{Rest}}}{\partial \tilde{X}}|$ to guide the feature distillation, which can be denoted as follows,

$$L_{\text{Distill}} = \frac{1}{\text{sum}(\tilde{\mathbf{M}})} \sum \tilde{\mathbf{M}} \odot \left| \mathcal{F}_{\text{DINO}}(I_c) - \mathcal{F}_{\text{Conv}}^{\text{Proj}}(\tilde{X}) \right|_1, \tag{6}$$

where $\tilde{\mathbf{M}} = \frac{\mathbf{M}-\min(\mathbf{M})}{\max(\mathbf{M})-\min(\mathbf{M})}$ is the normalized gradient map, $\mathcal{F}_{\text{Conv}}^{\text{Proj}}$ is a projection layer structured as $\texttt{Conv}-\texttt{ReLU}-\texttt{LayerNorm}-\texttt{Conv}$ to match the dimension. Consequently, the mask with higher values corresponds to areas with greater distortion and under-optimization, enabling the distillation to focus preferentially on critical regions.

Overall Objective. During the training stage of MetaScope, the overall optimization objective is denoted as:

$$\mathcal{L}_{\text{MetaScope}} = L_{\text{Seg}} + \lambda L_{\text{Rest}} + \omega_d L_{\text{Distill}} + \omega_k L_{\text{KL}}, \tag{7}$$

where L_{Rest} is for the auxiliary metalens restoration task, deployed with PSNR loss, L_{Seg} is the segmentation loss, implemented with CE loss, L_{KL} is the variational loss to ensure the Gaussian [38] in OCC estimating PSF distributions. ω_d and ω_k balance the two representation learning modules. λ is used to balance the segmentation and restoration.

5. Experiments

5.1. Benchmark Setup

Dataset Setup. We build up five benchmarks focusing on two critical clinical procedures: diagnosis and surgery. After establishing the metalens optical system, as described in Sec. 3.1, we are able to capture the metalens image I_m

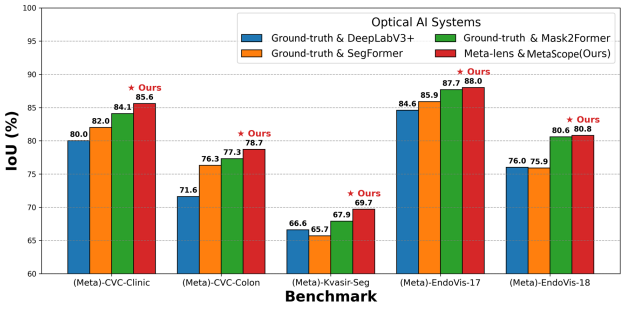


Figure 4. Comparison between conventional systems using high-quality ground-truth images without **Meta**, and our metalens system using the degraded images in **Meta** version dataset.

by photographing the publicly available database $\{I_c, Y\}$ projected on the screen, thereby generating the triplet data: $\{I_m, I_c, Y\}$ in 1024×1024 resolution. Based on our dataset, we carry out two streams of research to advance the metalens vision, including metalens segmentation with $\{I_m, Y\}$ (perception), metalens restoration $\{I_m, I_c\}$ (low-level vision). Hence, we name our datasets by adding the **Meta** prefix on the following five database: (1) CVC-Clinic [8], (2) CVC-Colon [7] (3) Kvasir-Seg [31], (4) EndoVis-17 [3], and (5) EndoVis-18 [4], such as Meta-EndoVis-18. More details can be found in Supp. D.

Evaluation Metrics. Evaluation is conducted in alignment with two clinical needs. (1) Clinical assistance: Metalens segmentation adopts mIoU and mDICE [42] (IoU/DICE for single-class datasets). (2) Clinical visualization: Metalens restoration uses PSNR and SSIM [94].

Implementation Details. The proposed model is based on the NAFNet [11] and is trained using the AdamW optimizer with a learning rate of 1×10^{-3}, a batch size of 2, and a weight decay of 1×10^{-4}. The training schedules of all benchmarked methods are consistent. Training was conducted on two 24G GPUs for 50,000 iterations. The number of MeG is set to $K = 3$. The ω_d and the ω_k are both set to 1, and the λ is set to 0.01/0.1 for diagnostics/surgery datasets. The simulation details and results of the optical priors in Eq. 2 are in Supp. B.

5.2. Comparison with State-of-the-art Methods

Metalens Segmentation. In Tab. 1 (top), we compare with state-of-the-art methods in both the medical field like U-KAN [42] and Rolling-UNet [64] and the natural imaging, including Mask2former [18]. Our MetaScope achieves superior performance, attaining an 80.51% mIoU and 87.67% mDICE across all five benchmarks, significantly outperforming the latest counterpart, EDAformer [93], by 6.81% in mIoU and 5.01% in mDICE. Most importantly, MetaScope exhibits the best performance across all five settings, encompassing varied clinical devices, in-vivo scenes, and diverse classes, demonstrating substantial robustness.

Metalens Restoration. Given promising achromatic metalenses [52], we benchmark metalens restoration and compare our method with state-of-the-arts, as shown in Tab. 1 (bottom). MetaScope achieves the best performance with a PSNR of 33.3714 dB and an SSIM of 0.9724, surpassing the latest metalens restoration method DRMI [80] and the natural imaging method MambaIR [24], revealing that our method can thoroughly correct distorted patterns. We also observe that methods tailored for specific restoration, such as NeRD-Rain for de-raining, perform unsatisfactorily, highlighting the unique characteristics of metalens (Sec. 3).

Comparing Metalens and Conventional Systems. Fig. 4 presents a more challenging comparison between metalens and conventional systems, where the latter utilizes high-quality ground-truth images and state-of-the-art segmentation algorithms. Surprisingly, despite metalens introducing chromatic aberration, our system (red bar) achieves the best performance across all benchmarks. This demonstrates that our method is robust to chromatic aberration and capable of unleashing the potential of metalens. *Notably, our system transcends the physical weight limitations and algorithmic precision constraints of conventional lens systems.*

Setting		Segmentation		Restoration	
-	w/o	mIoU	mDICE	PSNR	SSIM
Seg-Base	-	0.7376	0.8272	-	-
Meta-Base	-	0.7749	0.8573	31.1748	0.9538
+ Grad-Distill	Grad	0.7975	0.8750	32.4870	0.9632
	-	0.8312	0.8927	32.8153	0.9610
+ Grad-Distill;	MeG	0.8382	0.9012	32.8078	0.9482
+ OCC	-	0.8512	0.9088	34.0011	0.9603
+ Grad-Distill;	Prior	0.8523	0.9109	34.1341	0.9806
+ OCC; + OIA	-	**0.8555**	**0.9137**	**34.8483**	**0.9854**

Table 2. Detailed ablation study of the proposed modules on the Meta-CVC-Clinic benchmark, where 'w/o' indicates removing the sub-components, and 'Grad' indicates the gradient guidance.

Module	Setting	mIoU	mDICE	PSNR	SSIM
Optical Prior	Learned	0.8700	0.9293	31.6896	0.9600
	Only ch.	0.8715	0.9303	31.8751	0.9615
	Only sp.	0.8746	0.9320	32.0141	0.9612
	Full	**0.8802**	**0.9355**	**32.0365**	**0.9619**
Num. MeG K in Eq. 5	1	0.8682	0.9279	31.8780	0.9616
	3	**0.8802**	**0.9355**	32.0365	0.9619
	5	0.8793	0.9349	**32.0637**	**0.9625**
Loss Weight λ in Eq. 7	0.01	0.8676	0.9280	29.8646	0.9452
	0.1	**0.8802**	**0.9355**	32.0365	0.9619
	1.0	0.8497	0.9168	**32.0491**	**0.9622**

Table 3. Sensitivity analysis of the simulated optical prior and hyper-parameters on the Meta-EndoVis-17 benchmark, where 'Learned' indicates directly learning optical embedding.

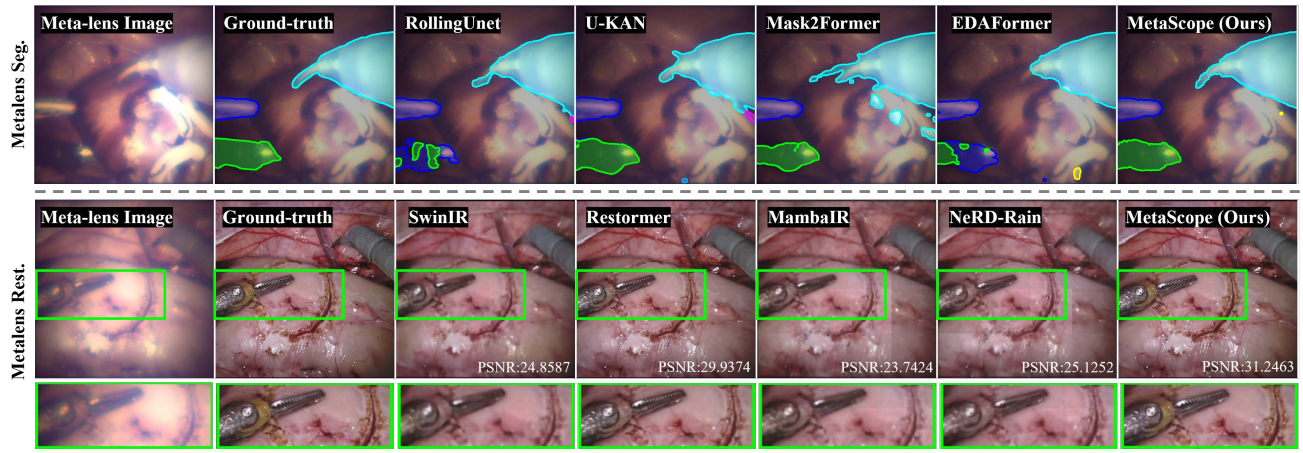


Figure 5. Qualitative comparison in terms of state-of-the-art segmentation methods and image restoration methods.

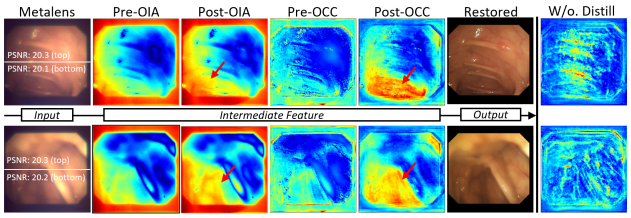


Figure 6. Heatmaps of the feature before and after each module.

5.3. Ablation Study

Effect of Each Module. We conduct a comprehensive ablation study presented in Tab. 2 and summarize several valuable observations. (1) Compared with the pure baseline (Seg-Base), introducing the restoration branch (Meta-Base) to guide segmentation yields a noticeable 3.01% mDICE gain. (2) The gradient-guided distillation contributes an additional 3.54% mDICE enhancement, highlighting the critical role of the restoration gradient (Grad). (3) Introducing OCC enhances the mDICE by 1.61% and achieves a 1.18 dB gain in PSNR, with the restoration gains primarily arising from the MeG design, emphasizing the importance of modeling dispersion patterns. (4) Incorporating OIA further refines the restoration representation, achieving the best performance in both segmentation (91.37% mDICE) and restoration (34.8483 PSNR), demonstrating the effectiveness of the optical prior in meta-restoration.

Sensitivity Analysis. We further conduct in-depth exploration in Tab. 3. (1) Removing the spatial and channel priors results in a noticeable decrease in PSNR, highlighting the essential role of correcting intensity attenuation. (2) Regarding the number of MeG, the performance remains robust when $K \geq 3$ but experiences a noticeable decrease when $K = 1$, revealing the effectiveness of modeling PSFs by the Gaussian mixture. (3) There is a trade-off in joint learning: Increasing the restoration loss weight yields further gains in PSNR while sacrificing some mDICE.

5.4. Qualitative Analysis

Qualitative Comparison. Fig. 5 shows a comparison between segmentation (top) and restoration methods (bottom). Compared to the latest method, EDAFormer [93], MetaScope shows visually significant improvements in the spatial completeness of challenging local regions. This is particularly evident in the prograsp forceps (green) and monopolar curved scissors (cyan) in row 1. Regarding the metalens restoration, Metascope achieves significant gains in both color accuracy and structural fidelity of local regions.

Analyze each Module with Heatmaps. In Fig. 6, we progressively visualize features before and after each module, highlighting ***insightful interoperability and well-aligned optical motivation***. **(1)** By modulating features with spatial priors, OIA enhances the near-border regions and relieves the position-dependent intensity loss, showing a similar pattern to Fig. 2c. **(2)** OCC encourages each pixel to cluster with its surrounding pixels, addressing PSF-caused dispersion with a spatial grouping (consistent with Fig. 2d). It also focuses on more distorted regions (bottom part) instead of segmenting targets, perfectly aligning with chromatic correction designs. **(3)** Without distillation, only low-level details are focused, yet lacking semantic guidance.

6. Conclusion

For the first attempt, we study endoscopic intelligence on the microscale with advanced metalenses. We establish benchmarks with five datasets across diagnostic and surgical applications. To address the chromatic aberration, we propose a novel MetaScope with two key components: *Optical-informed Intensity Adjustment* corrects intensity attenuation by learning optical embeddings; *Optical-informed Chromatic Correction* models and eliminates PSF distortion with multi-expert Gaussian. A gradient-guided distillation is integrated to enhance the meta-representation. Extensive experiments demonstrate that MetaScope significantly surpasses state-of-the-art methods.

Acknowledgement

This research was partly supported by Hong Kong PhD Fellowship from Research Grants Council of Hong Kong. This work was supported by Innovation and Technology Fund ITS/229/22, PRP/082/24FX and Hong Kong Research Grants Council (RGC) General Research Fund 14204321. We acknowledge the support from the University Grants Committee / Research Grants Council of the Hong Kong Special Administrative Region, China [Project No. AoE/P-502/20, CRF Project C5031-22G, GRF Project: CityU11305223; CityU11300224, CityU11310522, CityU11300123], National Natural Science Foundation of China [Grant No. 62375232], Guangdong Basic and Applied Basic Research Foundation [2025A1515011846], City University of Hong Kong [Project No. 9380131, 9610628] and the Shenzhen Research Institute, City University of Hong Kong [R-IND27901].

References

[1] Sharib Ali, Felix Zhou, Christian Daul, Barbara Braden, Adam Bailey, Stefano Realdon, James East, Georges Wagnieres, Victor Loschenov, Enrico Grisan, et al. Endoscopy artifact detection (ead 2019) challenge dataset. *arXiv preprint arXiv:1905.03209*, 2019. 1

[2] 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. 1

[3] 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. 7

[4] 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. 7

[5] Sensong An, Bowen Zheng, Mikhail Y Shalaginov, Hong Tang, Hang Li, Li Zhou, Yunxi Dong, Mohammad Haerinia, Anuradha Murthy Agarwal, Clara Rivero-Baleine, et al. Deep convolutional neural networks to predict mutual coupling effects in metasurfaces. *Advanced Optical Materials*, 10(3):2102113, 2022. 2

[6] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[7] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012. 7

[8] 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. 7

[9] Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283, 2020. 1

[10] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 2

[11] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022. 3, 6, 7

[12] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 6

[13] Mu Ku Chen, Xiaoyuan Liu, Yanni Sun, and Din Ping Tsai. Artificial intelligence in meta-optics. *Chemical Reviews*, 122 (19):15356–15413, 2022. 1

[14] Mu Ku Chen, Xiaoyuan Liu, Yongfeng Wu, Jingcheng Zhang, Jiaqi Yuan, Zhengnan Zhang, and Din Ping Tsai. A meta-device for intelligent depth perception. *Advanced Materials*, 35(34):2107465, 2023. 2

[15] Wei Ting Chen, Alexander Y Zhu, Vyshakh Sanjeev, Mohammadreza Khorasaninejad, Zhujun Shi, Eric Lee, and Federico Capasso. A broadband achromatic metalens for focusing and imaging in the visible. *Nature nanotechnology*, 13(3):220–226, 2018. 2

[16] 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. 3, 6

[17] 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. 1

[18] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2, 6, 7

[19] Yu-Hsin Chia, Wei-Hao Liao, Sunil Vyas, Cheng Hung Chu, Takeshi Yamaguchi, Xiaoyuan Liu, Takuo Tanaka, Yi-You Huang, Mu Ku Chen, Wen-Shiang Chen, et al. In vivo intelligent fluorescence endo-microscopy by varifocal meta-device

and deep learning. *Advanced Science*, 11(20):2307837, 2024. 2

[20] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer, 2016. 2

[21] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11065–11074, 2019. 3

[22] 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. 2

[23] Yuqi Fang, Cheng Chen, Yixuan Yuan, and Kai-yu Tong. Selective feature aggregation network with area-boundary constraints for polyp segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 302–310, 2019. 1

[24] 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. 3, 6, 7

[25] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *arXiv preprint arXiv:2209.08575*, 2022. 2, 6

[26] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010. 3

[27] Alessa Hering, Lasse Hansen, Tony CW Mok, Albert CS Chung, Hanna Siebert, Stephanie Häger, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, Wei Shao, et al. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 42(3): 697–712, 2022. 2

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 5

[29] Wei-Lun Hsu, Chen-Fu Huang, Chih-Chun Tan, Noreena Yi-Chin Liu, Cheng Hung Chu, Po-Sheng Huang, Pin Chieh Wu, Shang Jyh Yiin, Takuo Tanaka, Chun-Jen Weng, et al. High-resolution metalens imaging with sequential artificial intelligence models. *Nano letters*, 23(24):11614–11620, 2023. 2

[30] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 2

[31] 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. 7

[32] Debesh Jha, Sharib Ali, Krister Emanuelsen, Steven A Hicks, Vajira Thambawita, Enrique Garcia-Ceja, Michael A Riegler, Thomas de Lange, Peter T Schmidt, Håvard D Johansen, et al. Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27*, pages 218–229. Springer, 2021. 1

[33] Jiaqi Jiang, David Sell, Stephan Hoyer, Jason Hickey, Jianji Yang, and Jonathan A Fan. Free-form diffractive metagrating design based on generative adversarial networks. *ACS nano*, 13(8):8872–8878, 2019. 2

[34] Xingyu Jiang, Xiuhui Zhang, Ning Gao, and Yue Deng. When fast fourier transform meets transformer for image restoration. In *European conference on computer vision*. Springer, 2024. 3

[35] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020. 1

[36] 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

[37] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 3

[38] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5, 6

[39] Baopu Li and Max Q-H Meng. Computer-aided detection of bleeding regions for capsule endoscopy images. *IEEE Transactions on biomedical engineering*, 56(4):1032–1039, 2009. 1

[40] Chenxin Li, Wuyang Li, Hengyu Liu, Xinyu Liu, Qing Xu, Zhen Chen, Yue Huang, and Yixuan Yuan. Flaws can be applause: Unleashing potential of segmenting ambiguous objects in sam. *Advances in Neural Information Processing Systems*, 37:45578–45599, 2024. 2

[41] 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. 1

[42] 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. 2, 6, 7

[43] Mengmeng Li, Shuaishuai Li, Lip Ket Chin, Yefeng Yu, Din Ping Tsai, and Rushan Chen. Dual-layer achromatic

metalens design with an effective abbe number. *Opt. Express*, 28(18):26041–26055, 2020. 2

[44] Tao Li, Chen Chen, Xingjian Xiao, Ji Chen, Shanshan Hu, and Shining Zhu. Revolutionary meta-imaging: from superlens to metalens. *Photonics Insights*, 2(1):R01–R01, 2023. 1

[45] 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. 2

[46] 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. 1

[47] 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. 1

[48] 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. 1

[49] Wuyang Li, Xinyu Liu, Jiayi Ma, and Yixuan Yuan. Cliff: Continual latent diffusion for open-vocabulary object detection. In *European Conference on Computer Vision*, pages 255–273. Springer, 2024. 5

[50] Wuyang Li, Xinyu Liu, Qiushi Yang, and Yixuan Yuan. From static to dynamic diagnostics: Boosting medical image analysis via motion-informed generative videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 195–205. Springer, 2024. 1

[51] 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. 3, 6

[52] 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. 1, 2, 7

[53] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. *arXiv preprint arXiv:2308.15070*, 2023. 3

[54] Yi Lin, Dong Zhang, Xiao Fang, Yufan Chen, Kwang-Ting Cheng, and Hao Chen. Rethinking boundary detection in deep learning-based medical image segmentation. *Medical Image Analysis*, page 103615, 2025. 2

[55] 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. 1

[56] Shengyuan Liu, Zhen Chen, Qiushi Yang, Weihao Yu, Di Dong, Jiancong Hu, and Yixuan Yuan. Polyp-gen: Realistic and diverse polyp image generation for endoscopic dataset expansion. *arXiv preprint arXiv:2501.16679*, 2025. 1

[57] Shengyuan Liu, Boyun Zheng, Wenting Chen, Zhihao Peng, Zhenfei Yin, Jing Shao, Jiancong Hu, and Yixuan Yuan. A comprehensive evaluation of multi-modal large language models for endoscopy analysis. *arXiv preprint arXiv:2505.23601*, 2025. 1

[58] Xinyu Liu, Wuyang Li, and Yixuan Yuan. Intervention & interaction federated abnormality detection with noisy clients. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 309–319. Springer, 2022. 1

[59] Xiaoyuan Liu, Mu Ku Chen, Cheng Hung Chu, Jingcheng Zhang, Borui Leng, Takeshi Yamaguchi, Takuo Tanaka, and Din Ping Tsai. Underwater binocular meta-lens. *ACS Photonics*, 10(7):2382–2389, 2023. 2

[60] Xinyu Liu, Wuyang Li, and Yixuan Yuan. Decoupled unbiased teacher for source-free domain adaptive medical object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):7287–7298, 2023. 1

[61] Xiaoyuan Liu, Wuyang Li, Takeshi Yamaguchi, Zihan Geng, Takuo Tanaka, Din Ping Tsai, and Mu Ku Chen. Stereo vision meta-lens-assisted driving vision. *ACS Photonics*, 11 (7):2546–2555, 2024. 2

[62] Xinyu Liu, Wuyang Li, and Yixuan Yuan. Diffrect: Latent diffusion label rectification for semi-supervised medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 56–66. Springer, 2024. 2

[63] Xiaoyuan Liu, Zhou Zhao, Shengming Xu, Jingcheng Zhang, Yin Zhou, Yulun He, Takeshi Yamaguchi, Hua Ouyang, Takuo Tanaka, Mu Ku Chen, et al. Meta-lens particle image velocimetry. *Advanced Materials*, 36(17): 2310134, 2024. 2

[64] 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. 2, 6, 7

[65] Ange Lou, Shuyue Guan, and Murray Loew. Dc-unet: rethinking the u-net architecture with dual channel efficient cnn for medical image segmentation. In *Medical Imaging 2021: Image Processing*, pages 758–768, 2021. 2

[66] Yuan Luo, Cheng Hung Chu, Sunil Vyas, Hsin Yu Kuo, Yu Hsin Chia, Mu Ku Chen, Xu Shi, Takuo Tanaka, Hiroaki Misawa, Yi-You Huang, et al. Varifocal metalens for optical sectioning fluorescence microscopy. *Nano Letters*, 21(12): 5133–5142, 2021. 2

[67] Yiwen Luo, Wuyang Li, Cheng Chen, Xiang Li, Tianming Liu, Tianye Niu, and Yixuan Yuan. Llm-guided decoupled probabilistic prompt for continual learning in medical image diagnosis. *IEEE Transactions on Medical Imaging*, 2025. 1

[68] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024. 2

[69] Maksim Makarenko, Arturo Burguete-Lopez, Qizhou Wang, Fedor Getman, Silvio Giancola, Bernard Ghanem, and Andrea Fratalocchi. Real-time hyperspectral imaging in hardware via trained metasurface encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12692–12702, 2022. 1, 2

[70] Maksim Makarenko, Arturo Burguete-Lopez, Qizhou Wang, Silvio Giancola, Bernard Ghanem, Luca Passone, and Andrea Fratalocchi. Hardware-accelerated integrated optoelectronic platform towards real-time high-resolution hyperspectral video understanding. *Nature Communications*, 15(1): 7051, 2024. 1, 2

[71] Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42: 275–293, 2014. 5

[72] Yuichi Mori, Shin-ei Kudo, Hussein EN Mohmed, Masashi Misawa, Noriyuki Ogata, Hayato Itoh, Masahiro Oda, and Kensaku Mori. Artificial intelligence and upper gastrointestinal endoscopy: Current status and future perspective. *Digestive endoscopy*, 31(4):378–388, 2019. 1

[73] Yutaka Okagawa, Seiichiro Abe, Masayoshi Yamada, Ichiro Oda, and Yutaka Saito. Artificial intelligence in endoscopy. *Digestive Diseases and Sciences*, 67(5):1553–1572, 2022. 1

[74] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 6

[75] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[76] Wentao Pan, Zhe Xu, Jiangpeng Yan, Zihan Wu, Raymond Kai-yu Tong, Xiu Li, and Jianhua Yao. Semi-supervised semantic segmentation meets masked modeling: Fine-grained locality learning matters in consistency regularization. *arXiv preprint arXiv:2312.08631*, 2023. 2

[77] Wentao Pan, Jiangpeng Yan, Hanbo Chen, Jiawei Yang, Zhe Xu, Xiu Li, and Jianhua Yao. Human-machine interactive tissue prototype learning for label-efficient histopathology image segmentation. In *International Conference on Information Processing in Medical Imaging*, pages 679–691. Springer, 2023. 2

[78] Thaibao Phan, David Sell, Evan W Wang, Sage Doshay, Kofi Edee, Jianji Yang, and Jonathan A Fan. High-efficiency, large-area, topology-optimized metasurfaces. *Light: Science & Applications*, 8(1):48, 2019. 2

[79] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2, 6

[80] 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. 2, 6, 7

[81] 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. 2

[82] 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. 1, 2

[83] Akira Ueno, Juejun Hu, and Sensong An. Ai for optical metasurface. *npj Nanophotonics*, 1(1):36, 2024. 2

[84] Jeya Maria Jose Valanarasu and Vishal M Patel. Unext: Mlp-based rapid medical image segmentation network. In *International conference on medical image computing and computer-assisted intervention*, pages 23–33. Springer, 2022. 2, 6

[85] Chang Wang, Zongya Zhao, Qiongqiong Ren, Yongtao Xu, and Yi Yu. Dense u-net based on patch-based learning for retinal vessel segmentation. *Entropy*, 21(2):168, 2019. 2

[86] 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. 2

[87] Fufang Wen, Jiaqi Jiang, and Jonathan A Fan. Robust freeform metasurface design based on progressively growing generative networks. *Acs Photonics*, 7(8):2098–2104, 2020. 2

[88] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 1

[89] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 6

[90] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems (NeurIPS)*, 2021. 6

[91] Jiangpeng Yan, Hanbo Chen, Kang Wang, Yan Ji, Yuyao Zhu, Jingjing Li, Dong Xie, Zhe Xu, Junzhou Huang, Shuqun Cheng, et al. Hierarchical attention guided framework for multi-resolution collaborative whole slide image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 153–163. Springer, 2021. 2

[92] Qiushi Yang, Wuyang Li, Baopu Li, and Yixuan Yuan. Mrm: Masked relation modeling for medical image pre-training with genetics. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21452–21462, 2023. 1

[93] 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. 2, 3, 6, 7, 8

[94] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 3, 6, 7

[95] Jun Zhang, Zhuoran Zheng, Jingang Zhang, and Wenqi Ren. Ecsnn: Spiking neural networks for efficient exposure correction in endoscopy imaging. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 1

[96] 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. 2

[97] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 6