

Learned Off-aperture Encoding for Wide Field-of-view RGBD Imaging

Haoyu Wei, *Student Member, IEEE*, Xin Liu, Yuhui Liu,
 Qiang Fu, *Member, IEEE*, Wolfgang Heidrich, *Fellow, IEEE*, Edmund Y. Lam, *Fellow, IEEE*,
 and Yifan Peng, *Member, IEEE*



Fig. 1: (Left) Depiction of three potential locations for integrating a DOE for encoding purposes in an imaging system. The on-aperture DOE has many degrees of freedom to affect each ray bundle, but all of them are applied globally to the whole image plane. On the other hand, a DOE near the sensor provides localized control of the PSF, but with much fewer degrees of freedom for each ray bundle. The off-aperture DOE strikes an optimal balance between these two extremes. (Center-left) Cross-sectional view of two custom-fabricated optical imaging systems. The DOEs and apertures are located at separate planes. (Center-right) Resolved wide-FoV results of App.1 (Sec. 5) compared to encoded measurements; (Right) Color and depth results of App.2 (Sec. 6).

Abstract—End-to-end (E2E) designed imaging systems integrate coded optical designs with decoding algorithms to enhance imaging fidelity for diverse visual tasks. However, existing E2E designs encounter significant challenges in maintaining high image fidelity at wide fields of view, due to high computational complexity, as well as difficulties in modeling off-axis wave propagation while accounting for off-axis aberrations. In particular, the common approach of placing the encoding element into the aperture or pupil plane results in only a global control of the wavefront. To overcome these limitations, this work explores an additional design choice by positioning a DOE *off-aperture*, enabling a spatial unmixing of the degrees of freedom and providing local control over the wavefront over the image plane. Our approach further leverages hybrid refractive-diffractive optical systems by linking differentiable ray and wave optics modeling, thereby optimizing depth imaging quality and demonstrating system versatility. Experimental results reveal that the off-aperture DOE enhances the imaging quality by over 5 dB in PSNR at a FoV of approximately 45° when paired with a simple thin lens, outperforming traditional on-aperture systems. Furthermore, we successfully recover color and depth information at nearly 28° FoV using off-aperture DOE configurations with compound optics. Physical prototypes for both applications validate the effectiveness and versatility of the proposed method.

Index Terms—RGBD, wide FoV, off-aperture encoding, computational photography.

1 INTRODUCTION

JOINT-designed optics, dubbed *Deep Optics* in many research studies, have shown significant promise in a wide range of applications, including low-level vision tasks in

- H. Wei, X. Liu, Y. Liu, E. Y. Lam, and Y. Peng are with the Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Rd, Hong Kong SAR.
Corresponding E-mails: evanpeng@hku.hk, elam@eee.hku.hk.
- Q. Fu is with the Department of Computer Science and Electrical and Computer Engineering, King Abdullah University of Science and Technology, Saudi Arabia, 23955.
- W. Heidrich is with the Department of Computer Science and Electrical and Computer Engineering, King Abdullah University of Science and Technology, and the Department of Electrical and Electronic Engineering, The University of Hong Kong.

photography [1], [2], 3D vision and computational imaging [3], [4], as well as holographic displays [5]. Unlike traditional optical systems relying on heuristics [6], deep optics can enhance the encoding capabilities of camera systems, allowing their measurements to be decoded through post-processing algorithms [7], [8], [9]. This encoding-decoding proficiency is jointly engineered to achieve optimal imaging objectives. Specifically, optical elements such as refractive lenses [10], DOE [3], [11], [12], and metasurfaces [5], [13] are incorporated into cameras to modulate the complex incoming wavefront. The parameters governing these optics, such as curvature, height map, or atom diameters, are updated

in conjunction with differentiable algorithms employing backpropagation, typically neural networks [14].

In our work we are particularly interested in hybrid optics that combine classical refractive lenses, which provide high diffraction efficiency for focusing light, with diffractive optical elements (DOEs) for optical encoding [15]. Research on this topic has concentrated on optimizing the parameters of the encoding element, while neglecting the question of where to place the DOE relative to the aperture (or pupil plane). Conventional coded-aperture systems [16], [17], for simplicity, place the optical encoding on or near the aperture plane, creating *global modulation for the entire wave field*. For idealized unaberrated systems, this results in a shift-invariant PSF across the image plane. However, in the presence of off-axis aberrations, the global influence of the DOE parameters on the whole image becomes *an obstacle for localized aberration correction*. To address these issues, we explore separating the encoding plane from the aperture plane, so that different regions of the DOE produce different encoding characteristics based on the incoming light directions.

It is worth noting that in most imaging scenarios where the sensor size significantly surpasses that of the pupil, placing the DOE closer to the sensor increases the total degrees of freedom as well as the locality of the wavefront control. Nonetheless, positioning the DOE directly on the sensor is also not optimal, since the ability to spatially redistribute light is diminished in this position (Fig. 1). In this work, we investigate this inherent trade-off between the locality of control and the ability to redirect light.

The localized control over the wavefront shape is particularly relevant for wide-FoV imaging, where off-axis aberrations require a localized phase modulation for PSF shaping. In this work, we first show that in compact imager settings, pairing a simple thin lens with an off-aperture DOE enables the recovery of aberration-free images at 45°, outperforming the on-aperture DOE system by over 5 dB in peak signal-to-noise ratio (PSNR). Expanding on this achievement, we introduce more complex compound optics to tackle the challenging task of simultaneous color imaging and depth recovery at a wide-FoV of around 28°. This extension leverages the co-optimization of an off-aperture DOE with a Cooke triplet configuration [6]—comprising three refractive lenses positioned near the aperture plane, which partially corrects large-angle aberrations. This setting utilizes the high diffraction efficiency of refractive lenses to focus light, while allowing the DOE to encode high-frequency information essential for depth estimation. This hybrid refractive-diffractive system is modeled in two steps; the differentiable ray-tracing engine *dO* [18] is used for light propagation through the refractive lenses, followed by two-step off-axis wave propagation utilizing the least sampling ASM (LS-ASM) [19]. This hybrid design preserves the precise diffractive attributes of the optics, a feature absent in purely refraction-based systems.

As we showcase applications of multiple optimization objectives, we devise a lightweight multi-head neural network architecture that effectively extracts different information channels (e.g., color and depth) through separate decoding heads. Our design integrates a shared pre-trained feature extractor, with individual heads dedicated to specific

tasks. This framework not only alleviates the training burden compared to conventional single-head networks [20], but also facilitates efficient information encoding. In summary, our contributions are as follows:

- We present a co-optimized off-aperture encoding framework tailored for wide-FoV computational imaging applications. Notably, we explore the critical impact of off-aperture DOE placement.
- Our approach involves a differentiable refractive-diffractive hybrid imaging pipeline that integrates accurate off-axis wave propagation modeling, enabling the co-design of an off-aperture DOE and a multi-head image-processing network.
- Through our simulations, we demonstrate high-fidelity lightweight lens imaging capability at 45° with a PSNR gain over 5 dB compared to on-aperture systems, and RGBD imaging capability at 28°.
- We develop two camera prototypes: one equipped with an off-the-shelf convex lens, while the other features a customized compound lens incorporating bespoke computer numerical control (CNC)-turned aspherical lens geometry. Both prototypes utilize custom nanofabricated DOEs as encoding elements. Our systems are rigorously validated using a diverse range of indoor and outdoor scenarios.

2 RELATED WORK

Coded-Aperture Computational Imaging. Conventional image system design optimizes optics and algorithms separately [16], [21], [22], overlooking their potential synergistic interaction within a unified system. Computational imaging addresses this limitation by jointly optimizing encoding optics and decoding algorithms. Recent advances in AI, particularly deep neural networks, have facilitated the E2E design of computational imaging systems, co-optimizing hardware and software for specific tasks. These systems employ encoding masks to modulate incident light's amplitude, phase, and polarization, capturing rich scene information. Representative applications include hyperspectral imaging [23], [24], superresolution [11], extended depth-of-field [3], [25], and depth estimation [1], [26].

Several off-aperture designs have been studied in other imagers in the past decade. A near-sensor amplitude mask was analytically designed for light-field encoding [27]. Another work placed a DOE at an optimized position in a 4f system, using approximate off-axis wave simulations [28]. Concurrently, a metalens positioned off the aperture stop was optimized to maximize the pupil diameter for telescopic imaging at 20° FoV, simulated with geometric optics [29]. Unlike these approaches, our work aims to systematically explore and characterize this additional design space for wide-FoV compact imagers.

Wide Field-of-View Imaging. Wide-angle imaging settings are particularly challenging for optical systems with reduced complexity due to significant off-axis aberrations. Notably, efforts to recover sharp color and depth information from defocused images [12], [20] often assume spatial shift-invariance, valid only within a limited FoV. Expanding the FoV requires accurate modeling and mitigation of off-axis aberrations inherent in all imaging systems. Peng et al. [30] successfully demonstrated an imaging system that supports

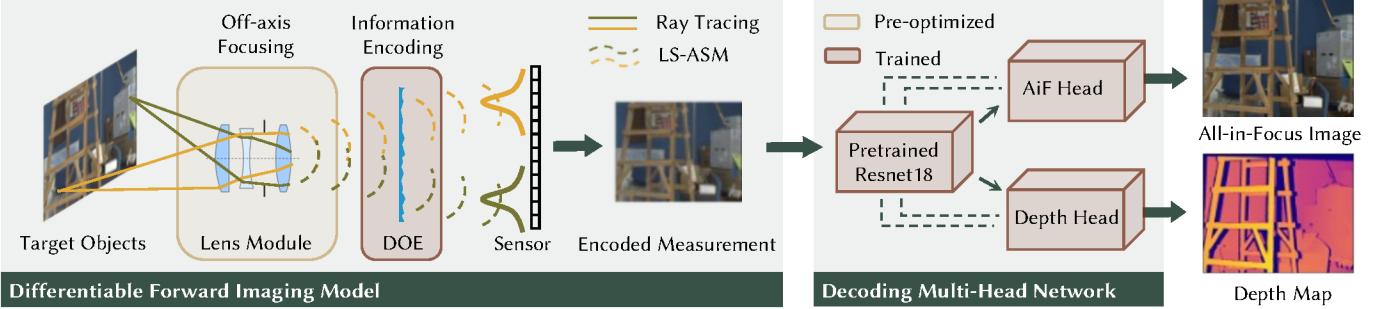


Fig. 2: Imaging pipeline of our proposed system. We model the camera’s light propagation using a combination of ray tracing and wave propagation. The compound lens module is pre-optimized and kept fixed during training, while the entire system is fully differentiable. The DOE placed in-between the lens module and the sensor facilitates information encoding. A multi-head decoding network based on the ResNet architecture is incorporated to support optimization for multiple visual tasks.

up to 53° by first crafting the shift-variant point spread functions (PSF) of a thin-plate lens using ray-tracing. A separate post-processing network was designed to remove residual artifacts. Nevertheless, commercial optical design software is often inefficient in modeling wave optics, particularly at large angles [31]. Moreover, task-driven designs require significant manual effort when using these tools, due to their incompatibility with E2E optimization. Differentiable ray-tracing engines have also been used to design compound optics systems enabling end-to-end all-in-focus (AiF) imaging over a 30° FoV [10], [32], [33] [34]. Despite their effectiveness, compound optics tend to be bulky and have limited bandwidth to encode additional scene information. A comprehensive summary of representative encoded imaging literatures can be found in Supplementary Table. 1.

Learning-based Image and Depth Recovery. Decoding both images and depth from a single encoded image is an ill-posed problem, typically addressed using deep neural networks. Standard U-Net or attention U-Net [35] with RGB-D output channels have been utilized in coded-aperture systems [2], [20]. However, the disparate features and sparsity of color images and depth maps impose significant burdens on the layers preceding the output. Anwar et al. [36] predicted depth and image sequentially, with the deconvolution performed using the intermediate depth predictions. This approach prioritizes depth estimation, as its quality is crucial for image deblurring.

More recently, a two-head network decouples the depth and image estimation tasks using parallel decoders sharing a feature encoder [37]. This allows for task-specific decoder designs but results in a larger model that can be difficult to train and susceptible to gradient vanishing. Therefore, to mitigate memory demands, we adopt a pre-trained ResNet18 [38] as the lightweight feature extractor with skip residual connections linking the encoder and decoders [39].

Refractive-Diffractive Hybrid Optics. Combining simple thin lenses with diffractive lenses offers compact and efficient encoding for coded aperture applications [20], [40], [41]. However, refractive components are still modeled with simple analytical functions, such as the Gaussian lens formula [6], excluding the use of compound optics and limiting the development of more advanced systems. This limitation is particularly critical for wide-FoV imaging, where simple

lenses suffer severe off-axis aberrations.

The most recent studies explore a differentiable hybrid lens design and showcase the potential of aberration correction, depth estimation, and depth-of-field extension. Yang et al. [15] represented a significant advancement in fully E2E optimizing both the optics and image-processing algorithms. However, their learned DOE also acts as the aperture stop of the overall system, whereas each DOE feature *globally* affects the wavefront of the entire image, similar to conventional coded aperture methods. Zhuge et al. [42] similarly optimized a Cooke triplet with a DOE but their DOE is, again, a near-aperture design. Furthermore, these works are limited to RGBD imaging with a FoV of under 20° . In comparison, our work seeks to separate the aperture and encoding planes of the optical system and to comprehensively investigate the impact of DOE positioning for the overall system performance, especially for wide FoV systems. Furthermore, we have experimentally demonstrated not only wider FoV imaging, but also extra depth estimation capabilities, compared to recent counterparts.

3 SHIFT-VARIANT HYBRID IMAGING MODEL

The optical part of our imaging system is composed of refractive lenses and a DOE, as illustrated on the left of Fig. 2. We demonstrate the design paradigm with two configurations tailored for specific imaging tasks. Initially, as the foundational application, we employ a simple main lens and a DOE placed off-aperture to realize wide-FoV imaging. Expanding upon this, the second application aims at not only enabling wide-FoV imaging capability, but also exploiting depth estimation (Fig. 2 right), facilitated by a pre-optimized Cooke triplet and a co-designed DOE.

To co-design these optical elements and image-processing neural networks, we establish an image formation model that simulates light propagation from the object to the image sensor. In this model, we first resort to geometric optics to obtain the response, i.e., the modified complex amplitude, of the refractive lenses (Sec. 3.1). Specifically, we assume the propagation through the simple lens is achromatic and model it by wavelength-dependent quadratic phase, while that of the compound optics is derived by ray tracing. Upon obtaining the response, we employ wave optics theory to accurately calculate the PSFs following DOE

modulation (Sec. 3.2). Finally, wide-FoV images are synthesized through spatial shift-variant convolution between the object and the PSFs.

3.1 Geometric Optics Simulation

Complex lens elements are widely employed in imaging systems to enhance performance, particularly in mitigating off-axis aberrations. To accommodate such design scenarios, we simulate complex lenses with geometric optics.

The lenses can be optimized, or predefined and then simulated using a differentiable ray tracing engine dO [18] to obtain the complex field before wave optics is applied. Rays are traced from the front to the backplane of the lens module's bounding box, which is tangential to the exterior of the lens module and assumed to be fully refractive within its boundaries. The backplane functions as the transition plane, linking refractive and diffractive optics. The complex fields at this plane represent the transfer functions of the compound lens at various angles.

The ray tracing process involves tracing rays between consecutive surfaces of the lenses, the aperture, and the front and back planes. At each point $\mathbf{o}_i \in \mathbb{R}^3$ on the surface i , where $i = 0, \dots, S - 2$ and S is the total number of surfaces including the backplane, a ray $\{\mathbf{o}_i, \mathbf{d}_i\}$ intersects the subsequent surface $i + 1$ at $\mathbf{o}_{i+1} = \mathbf{o}_i + t_i \mathbf{d}_i$. Here, the direction $\mathbf{d}_i \in \mathbb{R}^3$ is a unit vector determined by the previous ray direction \mathbf{d}_{i-1} and the surface material, following Snell's law. Notably, for $i = 0$, \mathbf{d}_{i-1} corresponds to the incident light direction. The ray marching distance $t_i \in \mathbb{R}^+$ is computed at surface intersections on i and $i + 1$ utilizing iterative root-finding techniques such as Newton's method.

The total phase shift of a ray is associated with the optical path length (OPL) through the lenses, given by the sum of OPLs $n_i t_i$ in each propagation medium, where n_i is the refractive index of the medium from surface i to $i + 1$. The rays reach scattered points on the transition plane, but subsequent processes such as wave propagation or sensing require measurements on a uniform grid. Thus, the points are interpolated onto a regular grid (ξ, η) using Clough-Tocher interpolation [43], denoted as g . The phase ϕ_{refrac} at the wavelength λ is then expressed as

$$\phi_{\text{refrac}}(\xi, \eta) = \frac{2\pi}{\lambda} g \left(\sum_{i=0}^{S-2} n_i t_i \right| \xi, \eta \right). \quad (1)$$

Consistent with prior works [18], [33], we ignore energy decay in geometric light propagation and assume uniform energy for each ray. The intensity I_{refrac} at the transition plane is obtained by bilinearly interpolating to adjacent grid points, accumulating ray energies. The transfer function is defined as $E_{\text{refrac}} = \sqrt{I_{\text{refrac}}} \exp(j\phi_{\text{refrac}})$, where $j = \sqrt{-1}$.

3.2 Off-Aperture Diffraction Modeling

Our two-step wave propagation process begins at the transition plane, traverses the off-aperture DOE, and ultimately culminates at the sensor plane. To simulate the highly off-axis wave propagation, we employ the well-established angular spectrum method (ASM) [44]. The propagation of an input field $E_{\text{in}}(\xi, \eta)$ to a finite region on plane (x, y) over a distance z is expressed as:

$$\text{ASM}(E_{\text{in}}|\xi, \eta; x, y) = \mathcal{F}^{-1} \{ \mathcal{F} \{ E_{\text{in}} \} H(f_X, f_Y) \}, \quad (2)$$

where $H(f_X, f_Y) = \exp[jkz\sqrt{1 - (\lambda f_X)^2 - (\lambda f_Y)^2}]$ is the frequency-domain transfer function, $k = 2\pi/\lambda$ is the wave number, and \mathcal{F} denotes the Fourier transform operator.

The off-aperture DOE connects two wave propagation pathways at large angles, which can be computationally demanding. To address this, we employ the LS-ASM [19] that introduces a compensation term to the phase of the input field, resulting in a quasi-on-axis field. This reduces the sampling requirements for simulating off-axis propagation. For an input field at coordinates (x, y) , the compensation term is defined as $\phi_{\text{comp}}(x, y) = -2\pi(f_{X_c}\xi + f_{Y_c}\eta)$, where (f_{X_c}, f_{Y_c}) represents the effective spectrum center. The wave field arriving at the DOE plane is then computed as:

$$E_{\text{DOE}}(\hat{p}, \hat{q}) = \text{ASM} \{ E_{\text{in}}(\xi, \eta) \exp[j\phi_{\text{comp}}(\xi, \eta)] \}. \quad (3)$$

The learned DOE globally modulates the incident field with $\phi_{\text{DOE}}(p, q) = k(n_{\lambda_0} - 1)h(p, q)$, where n_{λ_0} is the refractive index of the substrate material at the nominal wavelength, and h is the height map of the designed DOE. However, only a window at (\hat{p}, \hat{q}) is calculated on the DOE based on the chief ray's origin \mathbf{o}_{S-1} and direction \mathbf{d}_{S-1} at λ_0 from the lens module to encompass the majority of diffracted energy. Note that the DOE is also compensated. The complete wave propagation is thus formulated as:

$$\begin{aligned} \text{Prop}(E_{\text{in}}|\xi, \eta; p, q; x, y) = \text{ASM} & \left\{ E_{\text{DOE}}(\hat{p}, \hat{q}) \times \right. \\ & \left. \exp \{ j[\phi_{\text{comp}}(\hat{p}, \hat{q}) + \phi_{\text{DOE}}(\hat{p}, \hat{q})] \} \right\}. \end{aligned} \quad (4)$$

For our case, a point light source at (x_0, y_0, z_0) results in $E_{\text{in}}(\xi, \eta) = E_{\text{refrac}}(\xi, \eta) \exp(jkr)$, thereby the PSF is computed as the squared amplitude of the complex field arriving at the sensor plane within the specified window (x, y) :

$$K(x, y) = |\text{Prop} \{ E_{\text{refrac}}(\xi, \eta) \exp(jkr) \}|^2, \quad (5)$$

where $r = \sqrt{(x_0 - \xi)^2 + (y_0 - \eta)^2 + z_0^2}$. Refer to Fig. S1 in the supplementary material for details of the propagation modeling.

To simulate camera measurements, we convolve the PSFs with local image patches, assuming local shift-invariance due to the relatively smooth variations of PSFs across the FoV. We consider multiple depths $z_0 = 0, \dots, D - 1$ and utilize the established occlusion-aware model [20] to provide continuous and seamless representation of depth variations:

$$I = \sum_{d=0}^{D-1} \frac{1}{U_d} (K_d * I_d) \prod_{d'=d+1}^{D-1} (1 - \tilde{\alpha}_{d'}) + \mathcal{N}, \quad (6)$$

where $\tilde{\alpha}_{d'} = (K_d * \alpha_d)/U_d$, $U_d = K_d * \sum_{d'=0}^d \alpha_{d'}$, $*$ denotes element-wise multiplication, and K_d is the sensor field from a light source at depth $z_0 = d$. The images are quantized into D depth layers denoted as I_d , accompanied by a binary mask α_d . Each depth layer is normalized by U_d , and Gaussian noise \mathcal{N} is incorporated into the simulation. Sensor spectral response is also applied to image I to address the sensor's sensitivity to different wavelengths.

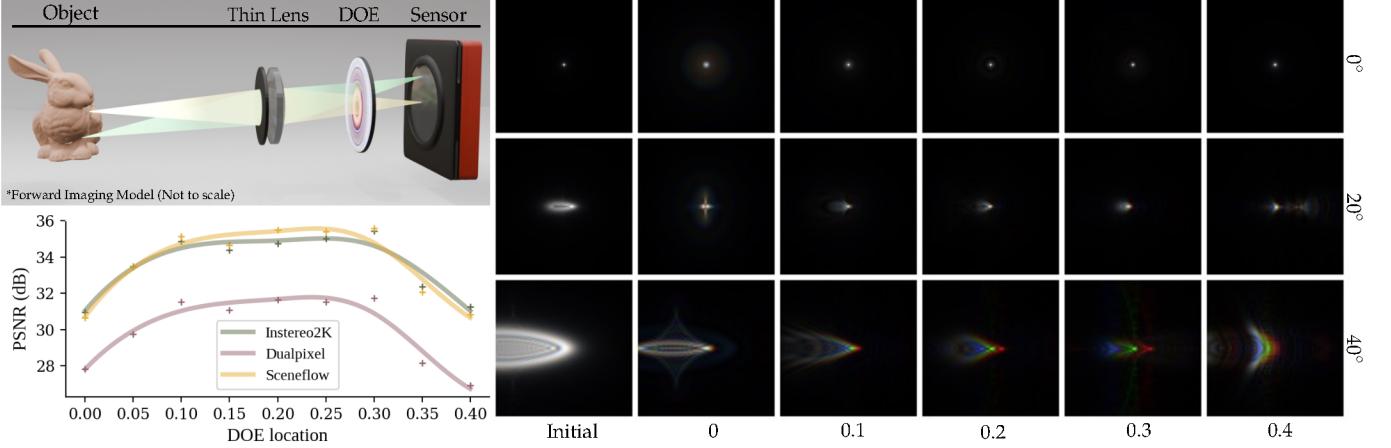


Fig. 3: (Top-left) The conceptual system setup of an off-the-shelf thin lens and an off-aperture DOE. (Bottom-left) The fitted PSNR plots for recovered images across various DOE locations shows that the optimal location is a trade-off point between the aperture and the sensor. The optimal DOE location for each dataset is approximately 0.24. (Right) Comparison of PSF amplitudes at several FoVs when the DOE is placed at different locations. The PSFs are cropped by the central 99 pixels for visualization.

4 MULTI-HEAD IMAGE-PROCESSING NETWORK

4.1 Network Architecture

The reconstruction network is tasked with extracting compressed information from sensor measurements and decoding it into both image and depth data (the right in Fig. 2). However, color images and depth maps possess fundamentally distinct spatial characteristics: images are characterized by dense color features, while depth maps exhibit sparse features. Moreover, sharp image reconstruction promotes the DOE to generate similar PSFs across various depths, whereas depth estimation prefers distinct PSF variations along the depth axis. This inherent difference necessitates extracting different types of information from the encoded features. Consequently, it is essential to employ separate decoding heads for each task, as illustrated in prior work [37].

To streamline training and reduce computational overhead, we adopt a shared feature extractor for both decoding heads. Specifically, we leverage a lightweight pre-trained ResNet18 [38], similar to Monodepth2 [39], and fine-tune it to better align with the encoded features in the sensor measurements. Each decoding head is comprised of four upsampling blocks, with each block integrating bilinear interpolation and residual convolution layers. Additionally, skip connections are applied between the ResNet encoder’s feature maps and the corresponding layers of decoders, facilitating efficient gradient flow and enhanced feature decoding across all levels.

4.2 Loss Functions

During training, the decoder in the network generates multi-scale images used to compute losses tailored to each task. For sharp image reconstruction, we employ the pixel-wise mean squared error (MSE) and the VGG16 perceptual loss $\mathcal{L}_{\text{percep}}$ [45] to ensure the reconstructions closely match the ground truth while maintaining visual fidelity. Given an image patch of size $M \times N$, the MSE loss is calculated as $\mathcal{L}_{\text{mse}} = \frac{1}{MN} \|I - \hat{I}\|_2^2$. For depth estimation, we utilize the \mathcal{L}_1 loss which promotes sparsity, defined as $\mathcal{L}_1 = \frac{1}{MN} \|I - \hat{I}\|_1$. To enhance focus power, we define a region O where the majority of the PSF energy should be concentrated. This

region is a circle with a diameter equal to half the PSF size. The PSF loss is defined as $\mathcal{L}_{\text{focus}} = \sum_{(x,y) \notin O} K(x, y)$.

Notably, multi-task reconstruction requires precise tuning of loss function weights for each task. Instead of relying on heuristic-based tuning methods [12], we adopt the dynamic weight averaging (DWA) [46] to dynamically adjust the weights $c_w (w = 0, 1, \dots)$ based on the average gradients from previous iterations. This strategy avoids complex hyperparameter tuning when all losses hold similar importance. Consequently, the total loss function is a weighted sum of individual losses, as follows:

$$\mathcal{L}_{\text{total}} = c_0 \mathcal{L}_{\text{mse}} + c_1 \mathcal{L}_{\text{percep}} + c_2 \mathcal{L}_1 + c_3 \mathcal{L}_{\text{focus}}. \quad (7)$$

5 WIDE-FoV SIMPLE LENS IMAGING

5.1 Design Overview

Wide-FoV imaging is increasingly challenging when aiming for a compact lens form factor, such as a single refractive lens element. State-of-the-art coded-aperture and/or deep optics imaging solutions mostly assume spatial-invariant PSF behavior and demonstrate in experiments a FoV up to 30°. A detailed summary of specifications in relevant work is presented in the supplementary material Tab. S1. To this end, our wide-FoV imaging setup includes a simple focusing lens with an approximate FoV of 45° and a rotational-symmetric off-aperture DOE, effectively compensating for most aberrations. Due to computational constraints, the camera is configured with an effective focal length (EFL) of 35 mm and an f-number of 12 for proof of concept. Objects are focused at 1.4 m depth.

The system is trained on a total of 26,281 images sourced from three datasets, which include 21,818 images from the simulated Sceneflow FlyingThings3D (SF) [47], alongside 2,506 images from the real-world Dualpixels (DP) [48], and 1,957 images from the Instereo2K (IS) [49]. During testing, we utilize 1,000 images from the SF dataset, 684 images from the DP dataset, and 50 images from the IS dataset, all upsampled and cropped to match the sensor resolution.

For this scenario, the network employs a single image-head with $c_2 = 0$. The learning rates are initialized at

1×10^{-4} and 6×10^{-3} for the network and DOE, respectively. Training is conducted over 50 epochs, incorporating a cosine annealing schedule starting at epoch 35, reducing the learning rate to 1/10 of its initial value. Following this, the system undergoes fine-tuning at the adjusted learning rate for an additional 10 epochs, with early stopping triggered after convergence. We randomly sample two angles for each forward iteration in the off-aperture simulations, aiding in gradient computation within the overlapping regions between each bundle of light. The training process typically takes a day on an Nvidia RTX 4090 GPU.

5.2 Analysis of DOE Position

To assess the trade-off between the angular de-multiplexing and local degree of freedom, we conduct an analysis on the system’s performance with various DOE positions, as in Fig. 3. The DOE is optimized at uniform intervals from 0 to 1 along the optical path, where 0 corresponds to the principal plane where light is refracted and 1 to the sensor plane. Considering assembly difficulties and to mitigate potential performance degradation associated with placing the DOE too close to the sensor, where local degree of freedom is compromised, our analysis focuses on DOE positions of $0 \sim 0.4$, corresponding to distances of $35 \sim 14$ mm from sensor, respectively. A spline curve is fitted to the sampled data points using a smoothing factor of 1.7 [50]. A scaled illustration of the system is presented in the Supplementary Material.

The optimized PSF presented on the right of Fig. 3 imply that the angular compromise diminishes as the DOE is positioned further away from the aperture (top row). For a larger FoV of 40° , the PSF is the most concentrated around positions 0.2 and 0.3. This observation is reinforced by the quantitative results depicted in the plot, indicating the image PSNR of DOEs optimized at various positions when assessed across all datasets. The off-aperture DOE system achieves PSNR values of 35.49 dB, 35.44 dB, and 31.74 dB on the SF, IS, and DP datasets, respectively. This outperforms the on-aperture system by 5.35 dB (SF), 4.41 dB (IS), and 4.13 dB (DP), respectively. These results are detailed in Tab. S1 and S2 in the supplementary material for a comprehensive presentation. The observed trend underscores the trade-off between local and global degree of freedom, implying that the optimal DOE position should lie between the aperture and sensor planes, but not at either extreme.

Notably, while the DOE position could be co-optimized as a parameter, the above analysis suggests that the performance impact is minimal when the DOE is positioned between 0.1 and 0.3 of the EFL. As a result, the DOE positions are not stable over multiple optimization runs, but all found solutions have comparable overall performance levels. For our physical prototypes we therefore opted to abandon the automatic optimization of DOE position and instead chose a position within the high-performance region that is also convenient for mechanical mounting.

5.3 Simulation Results

We partition the entire FoV into smaller local patches where shift invariance can be reasonably assumed. Each patch consists of 256×256 pixels, and we stitch together 16×10 patches to generate the final 4K resolution images. To mitigate stitching artifacts, we incorporate a 32-pixel

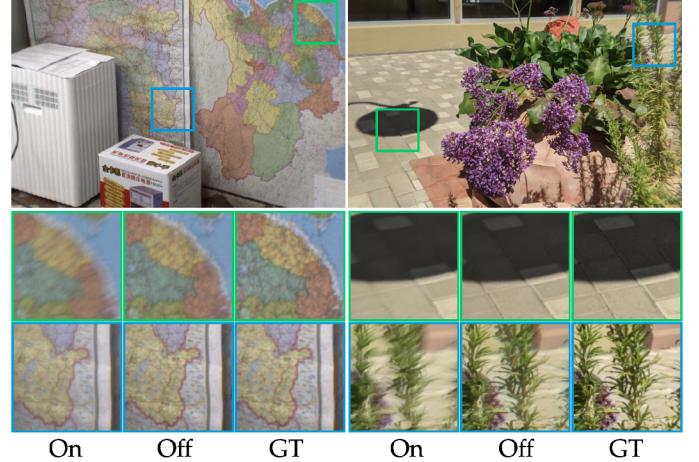


Fig. 4: Recovered 4K-resolution real-world scenes with wide-FoV imaging in simulation. The off-aperture DOE leads to much clearer results than the on-aperture DOE in both central and periphery FoV regions.

overlap at each boundary. Figure 4 compares the recovered images based on on-aperture and off-aperture DOE using the Instereo2K dataset. These results demonstrate that the off-aperture DOE significantly enhances visual quality in both the central and peripheral regions of the FoV, consistent with the corresponding metrics and optimized PSF characteristics.

While DOEs still encounter challenges in correcting off-axis aberrations, our preliminary results demonstrate their potential over an extended field of view. Specifically, at a 60° FoV, the proposed design achieves a 2.5dB improvement compared to on-aperture configurations (see Sec. S6).

6 WIDE-FoV COMPOUND LENS RGBD IMAGING

6.1 Design Overview

While the base application is demonstrated using a simple thin lens, a majority of imaging systems employ compound optics. These systems, though advantageous, introduce several challenges in our study. First, the aperture plane often does not coincide with the principal plane where light refraction occurs. Second, simulating a DOE prior to ray-traced lenses complicates the optimization process. Third, the joint optimization of refractive and diffractive lenses has received limited attention in prior research.

To address these challenges, we introduce a compound-optics prototype designed for wide-FoV depth and color imaging. The system combines an optimized Cooke triplet as the focusing lens module, achieving a practical balance between compactness and aberration correction. The lens profiles are initially optimized in Zemax OpticStudio to minimize aberrations across the entire FoV at the focal distance. For ease of fabrication, the last two lenses are replaced with the closest commercially available options, while the first lens is fine-tuned. Additionally, the first surface of the first lens is designed with an aspherical shape to better control off-axis light rays. Detailed specifications of our designed lenses are presented in the supplementary material Tab. S3. The off-aperture DOE is applied after the lens module.

The camera is configured with a 35 mm EFL and an

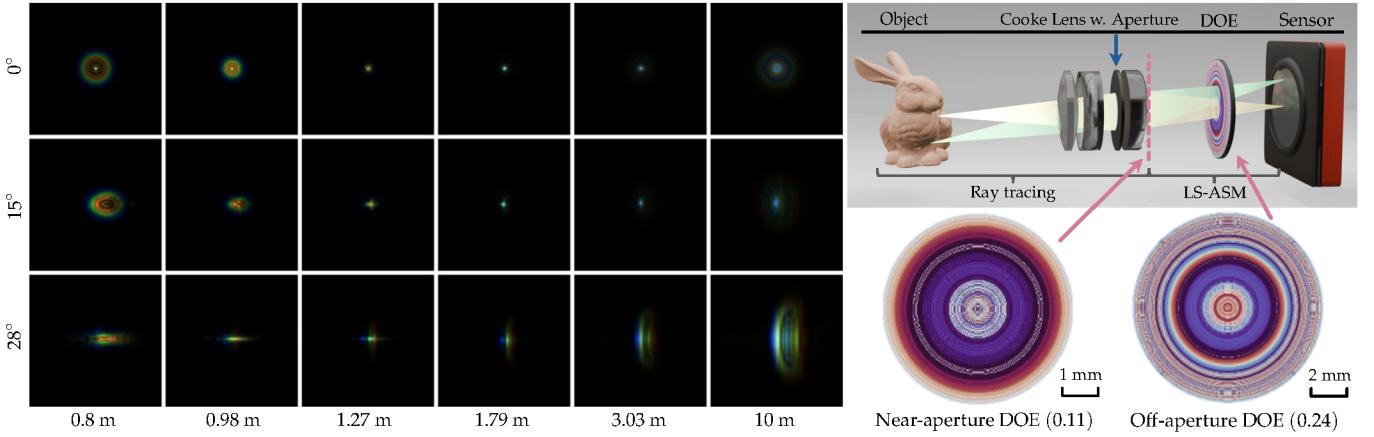


Fig. 5: (Left) Optimized PSFs at depths from 0.8 m to 10 m and FoVs up to 28°, using the off-aperture large-FoV EDoF imaging system. Depth layers are sampled uniformly. (Top-right) The 3D model of the optimized system, where the dotted line denotes the location to place the near-aperture DOE. Figure visualization is not drawn to scale. (Bottom-right) Optimized DOE height maps for near (around 0.11) and off aperture (around 0.24) settings.

f-number of 7. To integrate the lens module effectively, we optimize the DOE’s position within the back focal length, placing it 22 mm from the sensor but before the pre-assembled sensor components for ease of simulation and assembly. The system is designed to capture both depth and color information for objects within a range of 0.8 m to 10 m, corresponding to approximately 1.2 Diopter, with a maximum FoV of around 30 degrees.

We still utilize the three datasets comprising image-depth pairs with diverse features and varying depth map qualities. For this multi-task application, the network architecture employs two decoding heads: one for AiF imaging and another for depth estimation. Based on the optimized compound lens module for AiF imaging and insights on balancing multi-tasks from preliminary studies [12], we opt to yield a higher loss weight for the more challenging depth estimation task. Accordingly, we assign additional weights on the basis of DWA $c_2 = 1$ for depth estimation, while setting $c_0 = c_3 = 0.05$ for other parts. To manage GPU memory constraints associated with larger apertures, we disable perceptual loss by setting $c_1 = 0$. Experimental results demonstrate a relatively balanced performance between the two tasks. The training process retains the setting from previous simulations but with the cosine annealing scheduler starting at epoch 15 to gradually reduce the learning rate to 1/100 of its initial value, ensuring training stability. It typically takes two days to train on the same GPU as before for the increased sampling.

6.2 PSF Generation

While the off-axis wave propagator LS-ASM has proven effective in generating PSFs for simple lenses, it is inadequate for simulating the PSFs for compound optics systems. Similarly, *dO* can simulate geometric propagation for compound systems but fails to model fine details from light diffraction. The fusion of these two parts was recently demonstrated in concurrent work [15], while our additional design space requires considering DOE placement, and the absence of ground truth data for such systems hinders insightful comparisons. In supplementary Fig. S3, we compare the *dO* PSFs with those generated from the fused system, showcasing



Fig. 6: Recovered 4K-resolution real-world image and depth map with large-FoV EDoF imaging. Zoom-in patches of AiF images of near- and off-aperture DOE systems are presented and differences observed among the obtained depth maps are highlighted.

consistent shapes and structures across the entire FoV.

6.3 Simulation Results

To conserve computational memory and reduce runtime, we opt not to train the refractive lens with the DOE E2E, as double wave propagation at large angles requires a very high sampling rate. Although E2E optimization of a fused system is theoretically feasible, as demonstrated in concurrent work [15], it is orthogonal to the scope of this work. Instead, we present a proof-of-concept system that demonstrates the effectiveness of our approach while keeping complexity to a minimum.

We first design a Cooke triplet using Zemax OpticStudio, which minimizes aberrations at discrete angles (0°, 5°, 10°, 14°) for a focal depth of 1.7 m. This optimization yields a lens module with relatively low aberrations across the entire FoV, albeit not throughout all depths. Using this optimized

TABLE 1: Assessment of color images in PSNR (dB) \uparrow and depth maps in MAE \downarrow when optimizing DOEs at two positions closer and farther from the aperture in the application of wide-FoV depth and color imaging, and when using a simple network. These assessments are conducted on three datasets. The percentages of reduction in MAE of depth maps are reported.

Dataset	Near-aperture [42] + Multi-head	Off-aperture + U-Net [12]	Off aperture + Multi-head
Sceneflow	31.00 / 0.037	25.27 / 0.046	32.09 / 0.033 (-10.8%)
Dualpixel	27.95 / 0.025	23.69 / 0.027	28.28 / 0.019 (-24.0%)
Instereo2K	30.51 / 0.032	26.06 / 0.124	31.42 / 0.027 (-15.6%)

lens as a base, we integrate a DOE primarily to encode depth information. Based on the findings discussed in Sec. 5, we position the DOE at approximately 0.24 of the EFL, or 8.4 mm from the aperture plane. Due to the challenges in simulating the placement of the DOE before the refractive optics, we compare this off-aperture configuration with a near-aperture setup, where the DOE is positioned at the tangential surface of the last lens, approximately 4 mm from the aperture plane (0.11 of the EFL). Figure 5 presents the optimized PSFs of the off-aperture system, sampled uniformly at six depths and three angular directions. Notably, the system is trained on seven discrete angles (0°, 2.5°, 5°, 7.5°, 10°, 12°, and 14°), with intermediate angles 1.25°, 3.75°, 6.25°, 8.75°, 11°, and 13° included during testing. The optimized DOEs for near- and off-aperture systems exhibit distinct characteristics in their outer peripheral rings, highlighting the off-aperture DOE’s ability to efficiently encode high-frequency information for large-FoV rays without interference from the inner FoV. This represents a key advantage of the off-aperture design.

Table 1 presents a quantitative comparison of the near- or off-aperture systems. The off-aperture configuration achieves approximately 1 dB higher PSNR for color images and reduces mean absolute error (MAE) for depth maps by up to 24% across three representative datasets. The greater improvement in depth maps can be attributed to two factors: First, the near-aperture system’s DOE placement at 0.11 of the EFL incurs fewer disadvantages compared to the 0.24 location, and second, the Cooke triplet is optimized primarily for image performance, leaving more room for improvement in depth recovery. We further compare the multi-head network with a simple U-Net-based architecture [12], which outputs image and depth from the same layer as multiple channels. Our model outperforms the U-Net architecture in both image and depth recovery.

Figure 6 visually compares the near- and off-aperture systems at 4K resolution. The off-aperture system demonstrates notably superior sharpness in recovered color images compared to its near-aperture counterparts. This enhanced image quality is evident across various depths and throughout FoV, including both inner and outer regions. Moreover, the depth map produced by the off-aperture system exhibits significantly enhanced fidelity, particularly at wider FoVs. We further measure the performance at several sampled points in the Supplementary material. These results align with the prior observation that a distance between 0.2 and 0.3 represents an optimal choice for off-aperture designs.

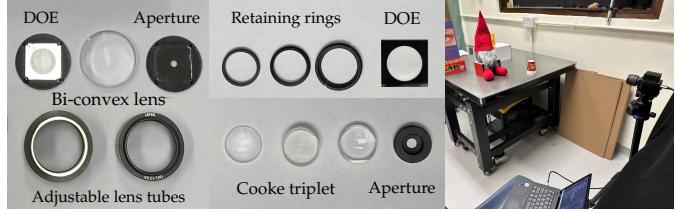


Fig. 7: (Left) The thin lens, DOE, and aperture in App. 1 and (Center) the three refractive lenses, DOE, and other components in App. 2. (Right) The experimental setup.

7 EXPERIMENTS AND RESULTS

7.1 Prototyping Details

In Application 1, we utilized an off-the-shelf anti-reflex (AR) coated N-BK7 bi-convex lens with a focal length of 35 mm (Thorlabs LB1811-A). Given the wide FoV, we employed a Sony A6400 mirrorless camera equipped with an APS-C sensor. The DOE has an effective diameter of approximately 12 mm and a full diameter of 16 mm. It is mounted with a 3D-printed circular mount with a diameter of 1 inch, in conjunction with the bi-convex lens. The distance between the center plane of the lens and the DOE plane is set at 10.5 mm. The total length of the imaging system outside the camera body is only 21.5 mm, resulting in a compact and lightweight wide-FoV imaging system. A mount adapter has been designed and 3D-printed to connect the SM1 lens tubes to the camera body.

In Application 2, we utilized one AR-coated N-SF11 bi-concave lens of -15 mm focal length (LD2060-A) and one CaF2 Positive Meniscus Lens of 50 mm focal length (LE5243), and importantly, fabricated one aspherical lens employing CNC machining capable of 5-axis single point diamond turning, akin to the techniques reported in state-of-the-art works [30], [32]. Given the available turning tool, we opt for polymethyl methacrylate (PMMA) as the substrate material, implying a refractive index of 1.492 at the principal wavelength of 550 nm. We employ the FLIR GS3-U3-123S6C-C sensor with 4K resolution.

Our DOE fabrication employs an iterative photolithography and dry etching approach to create 2^4 discrete phase levels on a fused silica wafer, with a Chromium layer acting as an optical baffle [51], [52]. Repeating the photolithography and reactive-ion etching steps four times preserves the high-frequency spatial features crucial for the DOE design. Detailed fabrication parameters are provided in the supplementary material. We then assembled all optics with customized focal-tuning mounts. The setup diagrams and photographs of the components in both applications are shown in Fig. 1 (center) and Fig. 7.

7.2 PSF Calibration and Network Fine-tuning

We employ a white laser source (LS-WL1 from Edmund Optics) coupled with collimating lenses and a 25 μm pinhole to serve as a point light source. In the first application, the PSFs captured on the sensor at a distance of 1.4 m are measured at 24 distinct locations within a quadrant of the image plane. Each measured PSF is spaced approximately 512 sensor pixels apart. In the second application, we capture PSFs at five FoVs for each depth plane of the nine depths from 0.8 m to 5 m. A detailed comparison between the simulated and

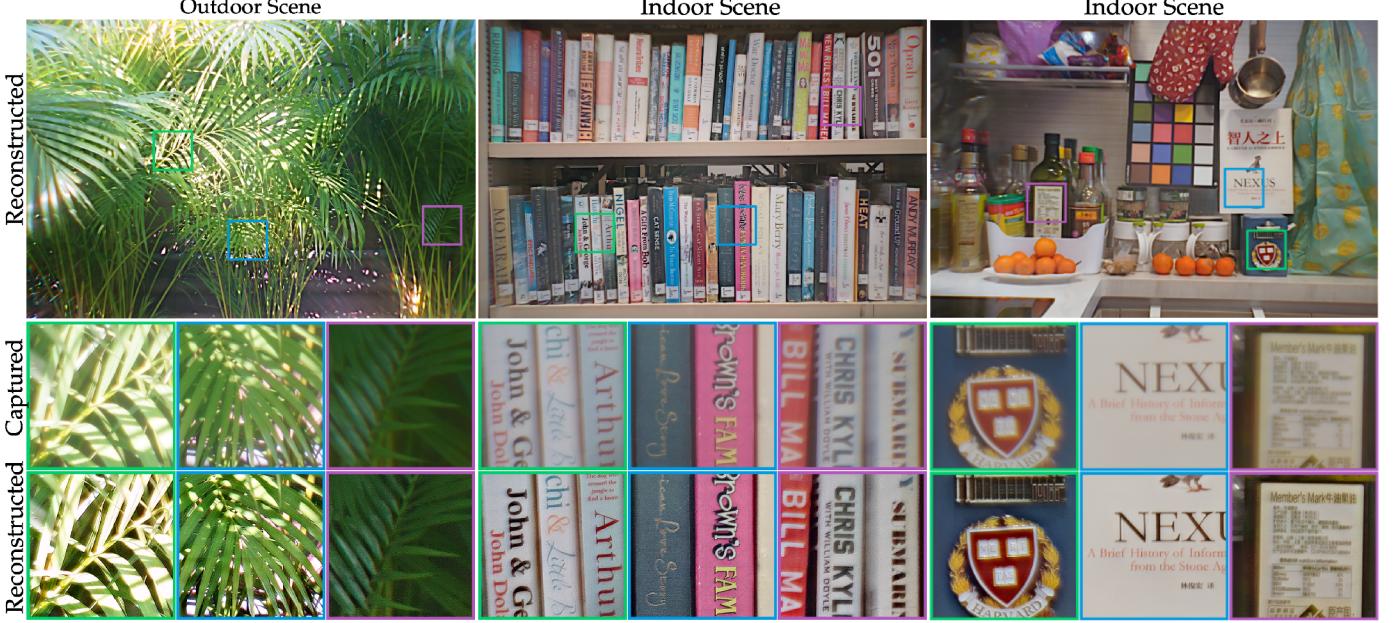


Fig. 8: Indoor and outdoor experimental results for the application of wide-FoV simple lens imaging. The equivalent f-number is 12. ISO is set as 100 for all captures.

captured PSFs is provided in the Supplementary Material.

Upon capturing the PSFs, we simulate measurements and fine-tune the network using the three datasets. However, we observe strong haze effects in the measurements of App. 2, which are likely caused by imperfections in the fabrication and assembling process [53]. To address this issue and enable the network to learn optical encodings from the corrupted measurements, we manually perturb the captured PSFs by introducing a zero-frequency term and applying random scaling to simulate realistic measurement conditions. Furthermore, we enhance the tuning process by incorporating a high-performing semantic-based depth estimation model “Depth Anything” [54]. Despite this, our optical encoded system successfully predicts plane depth maps from fake scenes displayed on a monitor, whereas semantic models fail to do so, highlighting the continued importance of optical encoding in resolving RGBD imaging. Further details and visualization are available in the supplementary material Fig. S6.

7.3 Real-world Results

The experimental results from real-world indoor and outdoor scenes, captured using the customized camera prototypes, are presented in Fig. 8 for App. 1 and Fig. 9 for App. 2, as well as in Fig. 1.

For App. 1, the reconstructed images exhibit significant improvements over the raw measurements across both inner and outer FoVs. High-frequency details, such as text and edges, are notably sharper in the reconstructed images. For instance, even small characters, like those on bottle labels, are accurately recovered, underscoring the method’s capability to reconstruct wide-FoV information effectively.

In App. 2, the AiF images demonstrate sharpness across the entire depth range. Notably, distant objects, such as highly blurred tree leaves that are otherwise unrecognizable to human eye, are faithfully reconstructed. Additionally, the recovered depth maps align well with visual cues, display-

ing clear transitions between depth planes. For example, the scene on the left shows distinct edges between the box and its background, while the right scene highlights depth variations in the net’s hollows. These results validate the effectiveness of our approach in wide-FoV AiF imaging and depth encoding. Additional real-world results of both camera prototypes are presented in Supplementary Fig. S6.

Comparisons with Simple Lens and Commercial Compound Lens. Although prior works have extensively demonstrated the advantages of DOE-based imaging systems, we also validate our App. 1 prototype against two baselines: (1) an off-the-shelf achromatic doublet with a pre-trained restoration network [55] and (2) a commercial Sony digital single-lens reflex (DSLR) lens. Compared to the first baseline, our DOE-encoded system shows superior detail preservation, particularly in wide-FoV regions. Against the compound lens, our off-aperture design achieves comparable performance in diverse indoor and outdoor scenes, despite its significantly smaller form factor. A simulated non-coded thin lens system is also compared by metrics. Additional details are provided in Supplementary Sec. 6.

7.4 Bells & Whistles: Comparison with A Pinhole

We further evaluate our system’s wide-FoV imaging performance in App. 1 by comparing it with a theoretically aberration-free pinhole setup. In this experiment, an indoor scene was captured under fixed exposure conditions, illustrated in Fig. 10. The counterpart pinhole setup employs the same convex lens but is paired with a 0.5 mm iris aperture. While this configuration minimizes severe off-axis aberrations at wide angles, it also significantly degrades the signal-to-noise ratio. We observe that the measurement appears extremely dim. Even after applying a state-of-the-art low-light image restoration algorithm [56] to the pinhole measurement, the result shows inferior detail recovery compared to our method, which maintains sharp edge fidelity across the entire FoV and ensures closer-to-real color re-

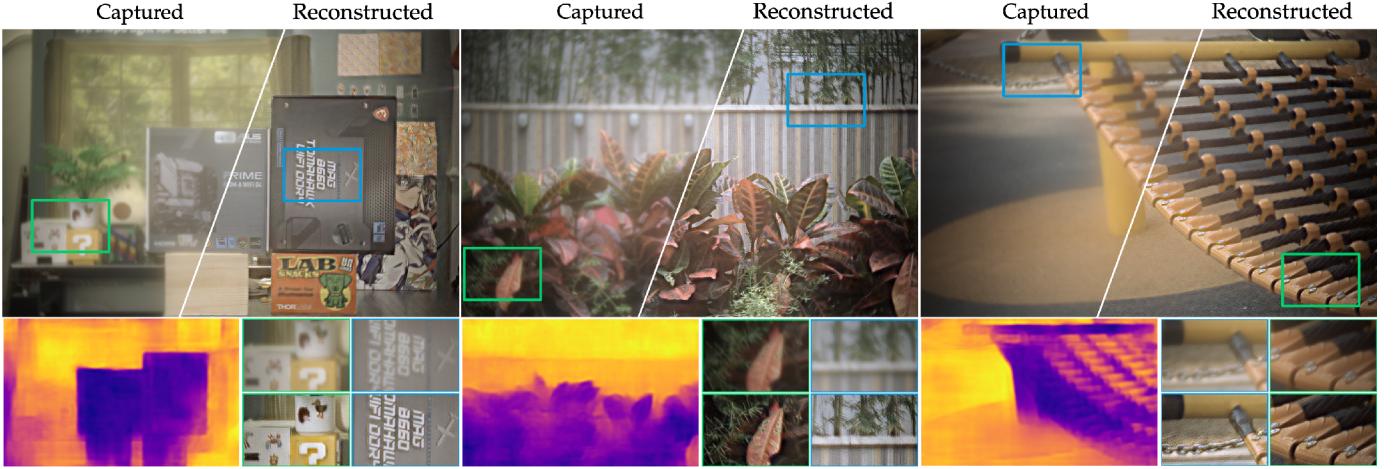


Fig. 9: Indoor and outdoor experimental results for App. 2: wide-FoV AiF and depth imaging. For each scene, we present the captured and reconstructed halves of color images (top), along with their zoom-in patches (bottom-right) and estimated depth map (bottom-left). The equivalent f-number is 7, and the gain set 0 for all captures.



Fig. 10: Experimental results of our wide-FoV RGB imaging prototype compared with denoised [56] results from a pinhole camera setting (0.5 mm diameter). Ours recovers correct color tones (with a reference captured by iPhone 15) and sharper edges. Both are captured with a shutter speed of 0.1 sec and an ISO of 500 under indoor illumination.

production. Importantly, our prototype's measurement preserves a higher fidelity of natural color tone, as evidenced by the color checkerboard captured with a cellphone camera.

Furthermore, while pinholes can reduce off-axis aberrations, large aperture settings remain essential for computational visual tasks that rely on optical encoding [16], such as RGBD imaging demonstrated in App. 2. In this context, pinhole configurations inherently extend the depth of field to near infinity, rendering depth recovery based on PSF encodings less efficient. This limitation can restrict the system to a purely semantic-based depth estimation approach, which is orthogonal to the scope of our off-aperture encoding exploration.

8 DISCUSSION AND CONCLUSION

In conclusion, we have presented an off-aperture co-optimized imaging system designed for wide-FoV imaging

applications. By positioning the DOE at a plane distinct from the aperture or pupil plane, we have achieved enhanced angle-dependent encoding through an additional design parameter. Our analysis of the impact of DOE placement provides valuable insights into how system performance is influenced by the DOE's position and demonstrates a region of jitter-robustness. Supplementary material also includes an optimization curve, which converges to ~ 0.3 in our experimental attempts, validating our choice. By decoupling light from different angles, we can enhance off-axis aberration correction by over 5 dB in PSNR.

Importantly, we have validated our system with two custom-designed and assembled camera prototypes: one featuring a simple lens and the other fusing multiple refractive optics, both equipped with co-optimized, off-aperture-positioned DOEs. These prototypes have demonstrated the system's effectiveness in RGB-only and RGBD imaging applications, achieving a FoV of up to 45° and 28° , respectively. We believe that the proposed off-aperture encoding paradigm holds great potential for broader applications, such as compact and wide-FoV cameras tailored for AR/VR, autonomous systems, and edge devices, where enhanced angular information encoding capability is critical.

Overview of Limitations: While all the individual components of the refractive-diffractive-hybrid, off-aperture encoded imaging framework are differentiable, we present an imaging system that optimizes the refractive lenses and the DOE in two separate steps to accommodate computational constraints, such as limited GPU memory, particularly when facilitating wide-FoV designs. In addition, experimental results also reveal that incorporating a DOE can introduce artifacts such as halos, which are attributable to limitations in diffraction efficiency and fabrication and assembly tolerances, especially under non-uniform and high illuminance conditions. Despite being widely validated on large image datasets, advanced semantic models are more susceptible to visual ambiguities, rendering them a practical yet imperfect choice for depth ground truth in fine-tuning. A more reliable alternative is to utilize active depth-sensing instruments such as a time-of-flight camera as ground truth.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation of China (62322217), the Research Grants Council of Hong Kong (GRF 17208023 & 17201822), the Innovation and Technology Fund of Hong Kong (ITP/062/24AP), and the Hong Kong STEM Fellowship awarded to W. Heidrich. The authors thank Mr. Feifan Qu and Ms. Duolan Huang for assisting with the experimental acquisition.

REFERENCES

- [1] Y. Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghavan, "Phasencam3d—learning phase masks for passive single view depth estimation," in *IEEE International Conference on Computational Photography*, 2019, pp. 1–12.
- [2] S. Shah, S. Kulshrestha, and C. A. Metzler, "Tidy-psfs: computational imaging with time-averaged dynamic point-spread-functions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 657–10 667.
- [3] V. Sitzmann, S. Diamond, Y. Peng, X. Dun, S. Boyd, W. Heidrich, F. Heide, and G. Wetzstein, "End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–13, 2018.
- [4] D. S. Jeon, S.-H. Baek, S. Yi, Q. Fu, X. Dun, W. Heidrich, and M. H. Kim, "Compact snapshot hyperspectral imaging with diffracted rotation," *ACM Transactions on Graphics*, 2019.
- [5] M. Gopakumar, G.-Y. Lee, S. Choi, B. Chao, Y. Peng, J. Kim, and G. Wetzstein, "Full-colour 3d holographic augmented-reality displays with metasurface waveguides," *Nature*, pp. 1–7, 2024.
- [6] R. E. Fischer, B. Tadic-Galeb, P. R. Yoder, R. Galeb, B. C. Kress, S. C. McClain, T. Baur, R. Plympton, B. Wiederhold, and B. G. A. J., *Optical system design*. Citeseer, 2000, vol. 599.
- [7] A. Chakrabarti, "Learning sensor multiplexing design through back-propagation," *Advances in Neural Information Processing Systems*, vol. 29, 2016.
- [8] G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljačić, C. Denz, D. A. B. Miller, and D. Psaltis, "Inference in artificial intelligence with deep optics and photonics," *Nature*, vol. 588, no. 7836, pp. 39–47, 2020.
- [9] Z. Wang, Y. Peng, L. Fang, and L. Gao, "Computational optical imaging: on the convergence of physical and digital layers," *Optica*, vol. 12, no. 1, pp. 113–130, 2025.
- [10] Q. Sun, C. Wang, F. Qiang, D. Xiong, and H. Wolfgang, "End-to-end complex lens design with differentiable ray tracing," *ACM Transactions on Graphics*, vol. 40, no. 4, pp. 1–13, 2021.
- [11] Q. Sun, J. Zhang, X. Dun, B. Ghanem, Y. Peng, and W. Heidrich, "End-to-end learned, optically coded super-resolution spad camera," *ACM Transactions on Graphics*, vol. 39, no. 2, pp. 1–14, 2020.
- [12] X. Liu, L. Li, X. Liu, X. Hao, and Y. Peng, "Investigating deep optics model representation in affecting resolved all-in-focus image quality and depth estimation fidelity," *Optics Express*, vol. 30, no. 20, pp. 36 973–36 984, 2022.
- [13] E. Tseng, S. Colburn, J. Whitehead, L. Huang, S.-H. Baek, A. Majumdar, and F. Heide, "Neural nano-optics for high-quality thin lens imaging," *Nature communications*, vol. 12, no. 1, p. 6493, 2021.
- [14] J. Hu, D. Mengu, D. C. Tzarouchis, B. Edwards, N. Engheta, and A. Ozcan, "Diffractive optical computing in free space," vol. 15, no. 1, p. 1525, 2024.
- [15] X. Yang, M. Souza, K. Wang, P. Chakravarthula, Q. Fu, and W. Heidrich, "End-to-end hybrid refractive-diffractive lens design with differentiable ray-wave model," in *Proceedings of SIGGRAPH ASIA*, 2024, pp. 1–11.
- [16] C. Zhou, S. Lin, and S. K. Nayar, "Coded aperture pairs for depth from defocus and defocus deblurring," *International journal of computer vision*, vol. 93, pp. 53–72, 2011.
- [17] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," *ACM Transactions on Graphics*, vol. 26, no. 3, pp. 70–es, 2007.
- [18] C. Wang, N. Chen, and W. Heidrich, "do: A differentiable engine for deep lens design of computational imaging systems," *IEEE Transactions on Computational Imaging*, vol. 8, pp. 905–916, 2022.
- [19] H. Wei, X. Liu, X. Hao, E. Y. Lam, and Y. Peng, "Modeling off-axis diffraction with the least-sampling angular spectrum method," *Optica*, vol. 10, no. 7, pp. 959–962, 2023.
- [20] H. Ikoma, C. M. Nguyen, C. A. Metzler, Y. Peng, and G. Wetzstein, "Depth from defocus with learned optics for imaging and occlusion-aware depth estimation," in *IEEE International Conference on Computational Photography*, 2021, pp. 1–12.
- [21] O. Cossairt and S. Nayar, "Spectral focal sweep: Extended depth of field from chromatic aberrations," in *IEEE International Conference on Computational Photography*, 2010, pp. 1–8.
- [22] Z. Zhang, C. Deng, Y. Liu, X. Yuan, J. Suo, and Q. Dai, "Ten-mega-pixel snapshot compressive imaging with a hybrid coded aperture," *Photonics Research*, vol. 9, no. 11, pp. 2277–2287, 2021.
- [23] K. Monakhova, K. Yanny, N. Aggarwal, and L. Waller, "Spectral diffusercam: lensless snapshot hyperspectral imaging with a spectral filter array," *Optica*, vol. 7, no. 10, pp. 1298–1307, 2020.
- [24] Z. Shi, X. Dun, H. Wei, S. Dong, Z. Wang, X. Cheng, F. Heide, and Y. Peng, "Learned multi-aperture color-coded optics for snapshot hyperspectral imaging," *ACM Transactions on Graphics*, vol. 43, no. 6, pp. 1–11, 2024.
- [25] B. Seong, W. Kim, Y. Kim, K.-A. Hyun, H.-I. Jung, J.-S. Lee, J. Yoo, and C. Joo, "E2e-bpf microscope: extended depth-of-field microscopy using learning-based implementation of binary phase filter and image deconvolution," vol. 12, no. 1, p. 269, 2023.
- [26] S. Tan, Y. Wu, S.-I. Yu, and A. Veeraraghavan, "Codedstereo: Learned phase masks for large depth-of-field stereo," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7170–7179.
- [27] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, "Dappled photography: mask enhanced cameras for heterodyned light fields and coded aperture refocusing," vol. 26, no. 3, 2007, p. 69.
- [28] B. Ferdman, A. Saguy, D. Xiao, and Y. Shechtman, "Diffractive optical system design by cascaded propagation," *Optics Express*, vol. 30, no. 15, pp. 27 509–27 530, 2022.
- [29] J. Wang, Y. Deng, C. Wang, Y. Lin, Y. Han, J. Liu, X. Liu, H. Li, J. G. Korvink, and Y. Deng, "Portable astronomical observation system based on large-aperture concentric-ring metalens," *Light: Science & Applications*, vol. 14, no. 1, p. 2, 2025.
- [30] Y. Peng, Q. Sun, X. Dun, G. Wetzstein, W. Heidrich, and F. Heide, "Learned large field-of-view imaging with thin-plate optics," *ACM Transactions on Graphics*, vol. 38, no. 6, pp. 219–1, 2019.
- [31] K. Matsushima, "Shifted angular spectrum method for off-axis numerical propagation," *Optics Express*, vol. 18, no. 17, pp. 18 453–18 463, 2010.
- [32] E. Tseng, A. Mosleh, F. Mannan, K. St-Arnaud, A. Sharma, Y. Peng, A. Braun, D. Nowrouzezahrai, J.-F. Lalonde, and F. Heide, "Differentiable compound optics and processing pipeline optimization for end-to-end camera design," *ACM Transactions on Graphics*, vol. 40, no. 2, pp. 1–19, 2021.
- [33] X. Yang, Q. Fu, and W. Heidrich, "Curriculum learning for ab initio deep learned refractive optics," *Nature Communications*, vol. 15, no. 1, p. 6572, 2024.
- [34] A. Teh, I. Gkioulekas, and M. O'Toole, "Aperture-aware lens design," in *Proceedings of SIGGRAPH*, 2024, pp. 1–10.
- [35] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: learning where to look for the pancreas," in *Medical Imaging with Deep Learning*, 2018.
- [36] S. Anwar, Z. Hayder, F. Porikli *et al.*, "Depth estimation and blur removal from a single out-of-focus image," in *The British Machine Vision Conference*, vol. 1, 2017, p. 2.
- [37] S. Nazir, L. Vaquero, M. Mucientes, V. M. Brea, and D. Coltuc, "Depth estimation and image restoration by deep learning from defocused images," *IEEE Transactions on Computational Imaging*, 2023.
- [38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [39] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3828–3838.
- [40] S. Pinilla, S. R. M. Rostami, I. Shevkunov, V. Katkovnik, and K. Egiazarian, "Hybrid diffractive optics design via hardware-in-the-loop methodology for achromatic extended-depth-of-field imaging," *Optics Express*, vol. 30, no. 18, pp. 32 633–32 649, 2022.
- [41] V. V. Evdokimova, V. V. Podlipnov, N. A. Ivliev, M. V. Petrov, S. V. Ganchevskaya, V. A. Fursov, Y. Yuzifovich, S. O. Stepanenko, N. L. Kazanskiy, A. V. Nikonorov *et al.*, "Hybrid refractive-

- diffractive lens with reduced chromatic and geometric aberrations and learned image reconstruction," *Sensors*, vol. 23, no. 1, p. 415, 2022.
- [42] Z. Zhuge, H. Xu, S. Chen, J. Xu, Q. Li, H. Feng, Z. Xu, and Y. Chen, "Calibration-free deep optics for depth estimation with precise simulation," *Optics and Lasers in Engineering*, vol. 180, p. 108313, 2024.
- [43] P. Alfeld, "A trivariate clough—tocher scheme for tetrahedral data," *Computer Aided Geometric Design*, vol. 1, no. 2, pp. 169–181, 1984.
- [44] J. W. Goodman, *Introduction to Fourier optics*, 4th ed. W.H. Freeman and Company, 2017.
- [45] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.
- [46] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1871–1880.
- [47] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4040–4048.
- [48] R. Garg, N. Wadhwa, S. Ansari, and J. T. Barron, "Learning single camera depth estimation using dual-pixels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7628–7637.
- [49] W. Bao, W. Wang, Y. Xu, Y. Guo, S. Hong, and X. Zhang, "In-stereo2k: a large real dataset for stereo matching in indoor scenes," *Science China Information Sciences*, vol. 63, pp. 1–11, 2020.
- [50] P. Dierckx, *Curve and surface fitting with splines*. Oxford University Press, 1995.
- [51] Q. Fu, H. Amata, and W. Heidrich, "Etch-free additive lithographic fabrication methods for reflective and transmissive micro-optics," *Optics Express*, vol. 29, no. 22, pp. 36 886–36 899, 2021.
- [52] X. Dun, H. Ikoma, G. Wetzstein, Z. Wang, X. Cheng, and Y. Peng, "Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging," *Optica*, vol. 7, no. 8, pp. 913–922, 2020.
- [53] Y. Li, Q. Fu, and W. Heidrich, "Extended depth-of-field projector using learned diffractive optics," in *IEEE Conference Virtual Reality and 3D User Interfaces*, 2023, pp. 449–459.
- [54] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 371–10 381.
- [55] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *European Conference on Computer Vision*, 2022, pp. 17–33.
- [56] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3291–3300.



Haoyu Wei is currently a Ph.D. candidate in the Department of Electrical and Electronics Engineering at The University of Hong Kong, Pokfulam, HK. She received the M.S. degree in Computer Science from Northwestern University, Evanston, IL, USA, in 2021, and the B.E. degree in Software Engineering from Sichuan University, Chengdu, China, in 2019. Her research interests include computational imaging systems and optical simulations.



Xin Liu is a Postdoctoral Fellow at The University of Hong Kong. He earned the Ph.D. in Optical Engineering from Zhejiang University in 2023 and the B.E. in Optoelectronic Information Science and Engineering from South China University of Technology in 2018. He has a broad interest in computational optics.



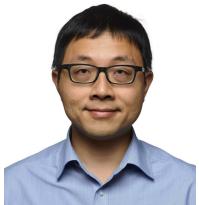
Yuhui Liu is a Ph.D. student in Electrical and Electronic Engineering at The University of Hong Kong (HKU). She received a M.Sc degree from HKU in 2023 and a B.E. degree from Nankai University (China) in 2022. Her research focuses on computational imaging techniques and diffractive optics design for emerging visual systems.



Qiang Fu is a Research Scientist at King Abdullah University of Science and Technology (KAUST) since 2017. Before that, he was a Research Associate Professor at ShanghaiTech University (2016–2017), a Postdoctoral Research Fellow at KAUST (2014–2016), and an Assistant Research Fellow at the Academy of Opto-Electronics, Chinese Academy of Sciences (2012–2014). He received his Ph.D. in Optical Engineering from the University of Chinese Academy of Sciences (2012), and Bachelor's degree in Mechanical Engineering from University of Science and Technology of China (2007).



Wolfgang Heidrich is a Professor of Computer Science and Electrical and Computer Engineering at King Abdullah University of Science and Technology (KAUST). He joined KAUST in 2014, after 13 years as a faculty member at the University of British Columbia (UBC). He received his PhD from the University of Erlangen in 1999, and then worked as a Research Associate in the Computer Graphics Group of the Max-Planck-Institute for Computer Science in Saarbrücken, Germany, before joining UBC in 2000. Prof. Heidrich's research interests lie at the intersection of imaging, optics, computer vision, computer graphics, and inverse problems. His more recent interest is in computational imaging, focusing on hardware-software co-design of the next generation of imaging systems. Prof. Heidrich is a Fellow of the National Academy of Inventors, IEEE, Optica, AAIA, and Eurographics, and the recipient of a Humboldt Research Award as well as the ACM SIGGRAPH Computer Graphics Achievement Award.



Edmund Lam received the B.S., M.S., and Ph.D. degrees in Electrical Engineering from Stanford University. He is currently a Professor of Electrical and Electronic Engineering and the Computer Engineering Program Director at The University of Hong Kong. He is a Founding Member of The Hong Kong Young Academy of Sciences, and a fellow of IEEE, Optica, SPIE, IS&T, IOP, and HKIE. His research interests include computational imaging algorithms, systems, and applications.



Yifan (Evan) Peng is an Assistant Professor at the Department of Electrical & Electronic Engineering and the Department of Computer Science, the University of Hong Kong (HKU). He was a Postdoctoral Fellow at Stanford University before joining HKU. He received his Ph.D. in Computer Science from the University of British Columbia, and Master's and Bachelor's degrees, both in Optical Science & Engineering, from Zhejiang University. His research focuses on incorporating optical and computational techniques to enable new visual computing systems. He is the recipient of the AsiaGraphics Young Researcher Award as well as the IEEE VR Tech Significant New Researcher Award.