This CVPR 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.

Learned Binocular-Encoding Optics for RGBD Imaging Using Joint Stereo and Focus Cues

Yuhui Liu[1] Liangxun Ou[1] Qiang Fu[2] Hadi Amata[2] Wolfgang Heidrich[2,1] Yifan Peng[1,*]
[1]The University of Hong Kong [2]King Abdullah University of Science and Technology

Abstract

Extracting high-fidelity RGBD information from two-dimensional (2D) images is essential for various visual computing applications. Stereo imaging, as a reliable passive imaging technique for obtaining three-dimensional (3D) scene information, has benefited greatly from deep learning advancements. However, existing stereo depth estimation algorithms struggle to perceive high-frequency information and resolve high-resolution depth maps in realistic camera settings with large depth variations. These algorithms commonly neglect the hardware parameter configuration, limiting the potential for achieving optimal solutions solely through software-based design strategies.

This work presents a hardware-software co-designed RGBD imaging framework that leverages both stereo and focus cues to reconstruct texture-rich color images along with detailed depth maps over a wide depth range. A pair of rank-2 parameterized diffractive optical elements (DOEs) is employed to encode perpendicular complementary information optically during stereo acquisitions. Additionally, we employ an IGEV-UNet-fused neural network tailored to the proposed rank-2 encoding for stereo matching and image reconstruction. Through prototyping a stereo camera with customized DOEs, our deep stereo imaging paradigm has demonstrated superior performance over existing monocular and stereo imaging systems in both image PSNR by 2.96 dB gain and depth accuracy in high-frequency details across distances from 0.67 to 8 meters.

1. Introduction

Stereo-empowered three-dimensional (3D) imaging replicates the human binocular vision system to acquire 3D scene information, enabling physical-based depth estimation, image enhancement, and super-resolution by leveraging both left and right frames.

Stereo-matching techniques have been widely adopted for passive depth estimation in applications such as surgical navigation [43], autonomous driving [38], and augmented reality [14]. Despite advancements in stereo depth estimation (SDE) methods—including transformer-based STTR [17], volume-based ACV-Net [42], and iteration-based IGEV [41]—current approaches still face challenges in perceiving high-frequency details while incurring significant computational overhead.

In stereo image enhancement and super-resolution, the strong correlation between left and right views is crucial [11]. Exploiting mutual information and complementary cues [45], such as stereo image correlation [36] and stereo cross-attention [44], enhances individual view quality. Maximizing information from the same scene point across both views is particularly critical over wide depth ranges, where focus considerations are essential. While algorithms like [9, 13, 18] have improved cross-view interactions, the frequency domain's complementarity and interaction remain underexplored, presenting opportunities for innovative optical encoding and systematic design. Prior studies in monocular deep optics [10, 23], that jointly optimize optics and image processing algorithms, have shown that the acquired optical encoding can significantly boost the performance of diverse visual tasks.

This work seeks an end-to-end stereo optics framework that integrates the joint optimization of phase-coded-aperture pairs with a stereo imaging neural network for depth estimation and image enhancement, exploiting both stereo and focus cues, as illustrated in Fig. 1. Specifically, we present two rank-2 DOEs that are jointly optimized to encode complementary high-frequency information while maintaining low computational complexity [1, 32]. The effectiveness of this approach is enhanced through tailored initialization strategies. Overall, we make the following technical contributions:

- We present an end-to-end design paradigm of binocular coded apertures and a stereo imaging network, aiming to resolve high-fidelity RGB and depth images.
- We formulate a rank-2 diffractive optics encoding to facilitate coded apertures for left and right imaging channels, greatly expanding the feature encoding and extraction ca-

*Corresponding author: evanpeng@hku.hk

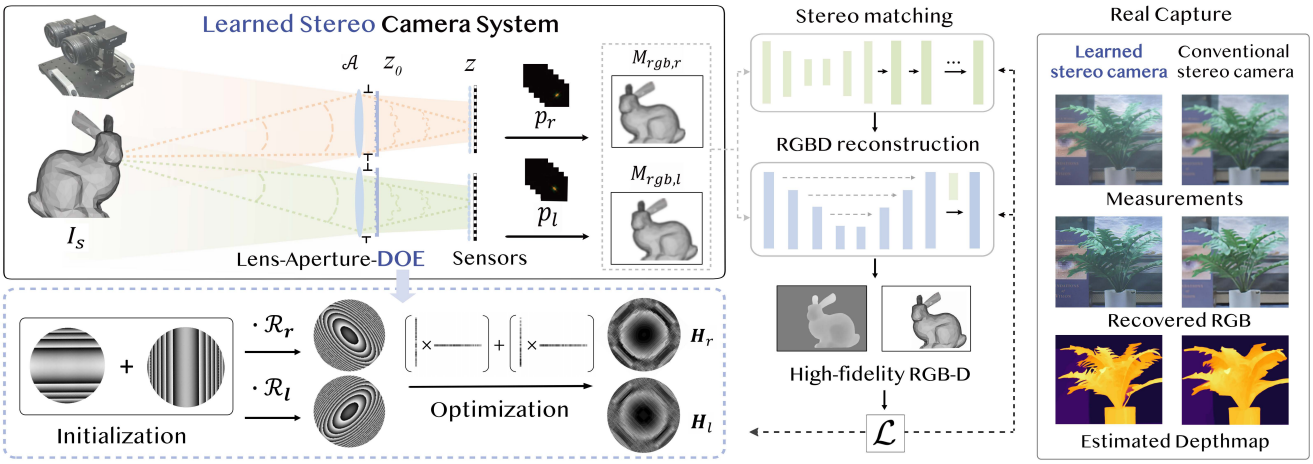


Figure 1. Our end-to-end learned stereo imaging pipeline consists of an accurate differentiable image formation model, an advanced stereo matching algorithm, and a CNN-based RGBD reconstruction network. This model leverages a rank-2 parameterization to efficiently represent and optimize the two DOEs (bottom-left) that are positioned on lenses' apertures, aiming to encode the stereo measurements which is capable to promote the interaction and complementarity of acquired scene information between the left and right imaging channels. Resolved RGBD imaging results from real-captures of our stereo camera prototype are presented (right-most).

pability for subsequent stereo matching.

- We develop a stereo camera prototype with a pair of learned diffractive optical elements placed on apertures, achieving high-fidelity 3D imaging results with a large depth range in various photography scenarios.

2. RELATED WORK

Stereo Depth Estimation. Neural networks excel at extracting features from left and right views, driving the advancement of learning-based stereo methods [19, 40]. PSMNet [4], a prominent example, employs a 3D convolutional encoder-decoder to aggregate and regularize a 4D cost volume. While 4D cost volume-based methods [42] achieve impressive benchmark performance, they require heavy 3D convolutions for aggregation and regularization.

Iterative optimization-based methods [34, 37, 41] have demonstrated remarkable performance on standard benchmarks and high-resolution images. Unlike cost volume-based approaches, these methods iteratively refine the disparity map by retrieving information from a high-resolution 4D cost volume without the need for computational cost aggregation operations. However, due to the computation demands of stereo matching, these methods involve downsampling the depth-map during volume construction updates and subsequently upsampling disparity maps.

Furthermore, when considering stereo depth estimation in real-world scenarios, the fidelity and accuracy are significantly constrained by physical camera capabilities such as depth of field (DoF) and the resolution of stereo cameras. While advanced models can distinguish high-frequency details from stereo image pairs [37], the limitation of physical cameras is that not all captured images are in perfect focus.

Stereo Image Enhancement. When stereo cameras capture left and right images with significant RGB overlap at appropriate working distances, cross-view interactions can enhance RGB quality in challenging scenarios [9, 44] or super-resolution applications [11, 45]. An essential aspect is to appropriately perform cross-view interaction, incorporating the features of the reference image into the target view. Recent studies have explored and utilized such correlations for stereo image enhancement [9, 44].

In addition to algorithmic image fusion and enhancement, a more systematic solution involves hardware-software co-design [27, 31]. In our proposed deep stereo framework, the input stereo pairs are encoded by learned optics within the imaging system to promote more interaction and complementary information in both spatial and frequency domains. Following pre-processing and encoding at physical layer, the learning-based algorithm functions as a decoder and fuser for the hardware measurements. The effectiveness of hardware-software joint optimization in imaging and vision tasks has been showcased in monocular [10, 32] and dual-pixel imaging [29], alleviating the computational burden on software.

Coded-aperture Imaging and Deep Optics. Traditional imaging systems typically utilize compound refractive lenses that are engineered for image quality independently. These refractive lens stacks are constrained by their smooth surface profile, thereby limiting the design flexibility for optically encoding desired task-specific scene information. In recent years, manipulating input light with a coded mask at

the aperture has been extensively explored in various computational imaging applications. The coded aperture can be tailored to manipulate the light wavefront, influencing its amplitude [29], phase [10], and polarization [1]. Diffractive optical elements (DOEs), a common thin lens platform utilized for aperture encoding, allow for fine-grained phase modulation of incident light via diffraction [15, 21], leveraging micron-scale surface profiles.

Thanks to the differentiable optical diffraction and image formation model [8, 39], coded-aperture optical systems can be optimized through back-propagation [31]. A recent development in monocular cameras has demonstrated that the joint optimization of optics and imaging algorithms can lead to superior performance for various visual tasks, including high-quality color photography [27], microscope imaging [20], monocular depth estimation [10, 21], high-dynamic-range imaging [23, 32], hyperspectral imaging [2, 12], and high-level computer vision tasks [28, 35]. This emerging field, dubbed Deep Optics, has been leveraged in dual-pixel camera systems, employing both amplitude- and phase-coded masks for imaging [7, 29]. Recent work by Tan et al. [33] introduced a multi-shot coded stereo system using identical optical encoding with separate RGB/depth processing, achieving extended depth-of-field within a 0.84-diopter range.

Our approach differs in that we utilize a pair of complementary DOEs in a snapshot binocular configuration, achieving superior imaging precision and expansive depth estimation capabilities by fusing aggregated and physical cues. By utilizing asymmetric learned rank-2 PSF encoding and integrating left-right channel image fusion within our stereo framework, we effectively resolve high-fidelity depth and complementary angular information across multiple depth layers, surpassing conventional coded-aperture imaging solutions in both accuracy and versatility.

3. Stereo Phase-coded 3D Imaging

3.1. Monocular Image Formation Model

As shown in Fig. 1, the left and right camera in our binocular setup utilize the same design space, so that we can first describe the image formation process for each camera independently before considering the stereo effect.

Each camera is modeled as an optical stack of an idealized thin lens, an aperture, and a diffractive optical element (DOE), without spacing between the components, i.e. all three components are effectively co-located in the same plane (Section 5 discusses how this model can be approximated with a real optical system). In this imaging system, the ideal lens is tasked with optical power (focusing) while enabling the DOE to perform the tailored encoding operation. An object point at a finite distance z from the camera results in a diverging spherical wavefront incident on the thin lens. Ideal thin lens converts the diverging spherical wave into a converging spherical wave at z, and the resulting wavefront after passing through the thin lens can be denoted as $\boldsymbol{u}_1(x', y', z)$, where (x', y') represents the coordinates on the Lens-Aperture-DOE plane, as shown in the upper left of Fig. 1. Subsequently, this wavefront interacts with the aperture and the height field geometry $\boldsymbol{H}(x', y')$ of the DOE, leading to the generation of the final wave field $\boldsymbol{u}_{z_0}$ immediately after the optical stack [8]:

$$\boldsymbol{u}_{z_0} = \mathcal{A}(x', y')\boldsymbol{u}_1 e^{jk\left[(n(\lambda)-n_0)\boldsymbol{H}(x', y')\right]}, \tag{1}$$

where $\mathcal{A}$ is the aperture, k=$2\pi/\lambda$ denotes the wave number, and n_λ is the wavelength-dependent refractive index.

To obtain the (depth dependent) PSF p_λ of the scene point for a specific wavelength, we can perform free-space propagation of $\boldsymbol{u}_{z_0}$, for example using the angular spectrum method (ASM), and then squaring the resulting wave field to obtain its intensity [8]. These wavelength-dependent PSFs p_λ contribute to the measurement M_c of an RGB image sensor (with c being the color channel) according to the following image formation model:

$$M_c = \int_\Lambda R_c(\lambda)\left[I_s(\lambda) * p_\lambda\right] d\lambda + \eta_c, \tag{2}$$

where Λ denotes the target spectrum, I_s represents the all-in-focus scene image, $*$ represents the 2D convolution, η_c is the corresponding noise, and $R_c(\lambda)$ is the spectral response function of channel c. Incorporating sensor response into measurements can yield more realistic simulation results.

3.2. Rank-2 Parameterized DOE

A critical choice for end-to-end learned diffractive optics is the parameterization of the DOE. Existing choices, such as rotational symmetric models [5, 26] and pixel-wise approaches [24, 31], often induce encoding local minima (e.g., highlight shifts and scaling artifacts [32]) incompatible with our approach. Specifically, pixel-wise representation shows little control over the local smoothness, which increases manufacturing difficulties, usually resulting in a lower diffraction efficiency. On the other hand, the blur encoded by ring-pattern encounters challenges in balancing image and depth performance [21].

Moreover, for our binocular system, we seek encodings that can provide complementary information in the left and right view, while still allowing for robust stereo matching with an easy-to-learn, easy-to-fabricate DOE design space. Inspired by the previous research [32], we choose to parameterize the DOE pair using low-rank matrices. This representation can not only facilitate the encoding of high spatial frequencies but also contribute to parameter reduction during training. A rank-1 height map at coordinates (x', y') can be defined as:

$$\boldsymbol{H}(x', y') = H_{max} \cdot \sigma(\boldsymbol{a}\boldsymbol{b}^T), \tag{3}$$

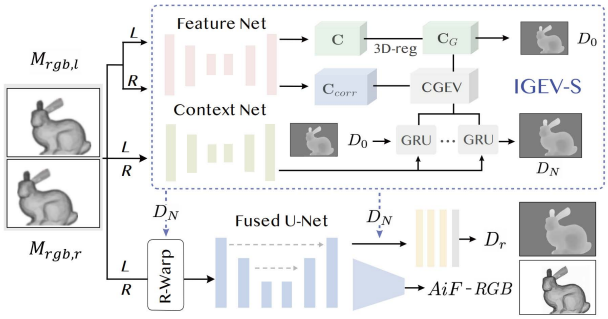


Figure 2. Overview of the stereo imaging network architecture, that consists of two components: a stereo matching network [41] and a RGBD refinement network. After deriving the disparity map D_N through N GRUs, we feed it into the stereo image enhancement network for image warping and depth map enhancement. A Combined Geometry Encoding Volume (CGEV) is constructed, serving as a cost volume by integrating all-pairs correlations C_{corr} with a geometry encoding volume C_G.

where $\boldsymbol{a}, \boldsymbol{b}$ are $m \times 1$ learnable real valued vectors. A sigmoid function σ is applied to constrain the heightmap value.

Instead of using rank-1 representation, we found that a rank-2 design offers enhanced control and reconstruction quality while retaining the benefits of the rank-1 design space. The rank-2 design is formulated as the sum of two rank-1 matrices $\boldsymbol{m}_1 = \boldsymbol{a}_1\boldsymbol{b}_1^T, \boldsymbol{m}_2 = \boldsymbol{a}_2\boldsymbol{b}_2^T$, where $\boldsymbol{a}_i, \boldsymbol{b}_i \in \mathbb{R}^m$ are learnable real number vectors, and m is set 630. We further enforce symmetrical PSFs by optimizing only the parameters for one quadrant, and then replicating this quadrant four times through rotations by 90°, 180°, and 270°, respectively. Finally, we found it beneficial to have the axes of the low-rank DOE rotated by 45° relative to the pixel grid of the cameras (as well as the baseline between the cameras) to facilitate simultaneous frequency sampling in the x and y directions, as illustrated in Fig. 7. The final DOE height map $\boldsymbol{H}(x', y')$ can then be described as:

$$\boldsymbol{H}(x', y') = \mathcal{R}_{\frac{\pi}{4}}\left\{\text{Quad}\left[H_{max} \cdot \sigma\Big(\sum_{i=1}^{2} \boldsymbol{a}_i\boldsymbol{b}_i^T\Big)\right]\right\}, \quad (4)$$

where Quad represents the quadrant replication operator, and $\mathcal{R}_{\frac{\pi}{4}}$ is the rotation by 45°. The modeling process for left and right DOE is illustrated in the bottom-left Fig. 1.

3.3. Network Architecture

In the stereo matching network (refer to Fig. 2), we leverage the context network following the RAFT-Stereo [19] and a multi-scale feature extractor to extract features $\mathbf{f}_{l,i}$ ($\mathbf{f}_{r,i}$) $\in \mathbb{R}^{Ci \times H_i \times W_i}$, where i=4, 8, 16, and C_i indicates feature channels. Besides, a simplified iterative geometry encoding volume (IGEV) [41] is utilized for stereo matching, consisting of a group-wise correlation volume and processing the volume through a lightweight 3D regularization network to derive the geometry encoding volume (GEV). We then pass the GEV through ConvGRU-based operators to iteratively update the disparity. The resulting disparity map D_N is a predetermined set of disparity indices at 1/4 resolution, linked with a spatial upsampler. In the stereo image recovery network, we warp the right image to the left view using the estimated depth map and incorporate a U-Net with a cross-cue mechanism [16] to fuse images encoded by complementary PSFs from our rank-2 modeling.

3.4. End-to-end Optimization

The proposed stereo RGBD imaging pipeline jointly optimizes three primary parts — a differentiable optics model, a robust stereo matching network, and an RGBD reconstruction network, as illustrated in Fig 1. This integrated system seeks to resolve the all-in-focus RGB image (AiF-RGB) and detailed depthmap $\boldsymbol{D}_{\{l,r\}}$, from blurred measurements.

DOE Initialization. The objective is to ensure that the left and right DOEs are optimized to offer complementary encoding and induce non-smooth phase variations within our rank-2 model. For the left and right channels, the symmetric-like nature of network architectures may pose challenges in achieving complementary encoding with standard initializations, such as zeros or random matrices. Alternatively, in this work, the DOE is initialized to emulate a pair of compound cylindrical lenses, as illustrated in Fig. 3 (left-most), enforcing different spreading directions for left and right PSFs. This special initialization not only enables rich phase variation but also imparts focal power to our stereo camera system. We have conducted a comprehensive analysis of optimized sampling outcomes using different initialization methods in the frequency domain. The impact of various DOE models and initialization strategies is presented in the supplementary material.

Image Simulation. We integrate the state-of-the-art LS-ASM [39] into our image simulation process to determine the minimal number of samplings essential for imaging simulations. When generating 3D PSFs, it is crucial to model the realistic defocus blur for each region in an image based on its depth value, especially in areas with depth transitions. Taking into account multiple depth layers within the working-distance range, we utilize a nonlinear differentiable image formation model proposed by Ikoma et al. [10] based on alpha compositing. In this model, the input RGBD image is quantized into K depth layers l_k, with $k = 0$ representing the furthest layer. As it is computationally demanding to forward and back-propagate the entire full visible spectrum during model training, we consider the sensor responses as δ functions and employ a simplified image formation model. Detailed analysis on the PSF behaviors across full spectrum is presented in the supplementary material. Thus, the left and right sensor images $I_i(\lambda)$, where $i \in l, r$ and $\lambda \in (632, 550, 450)$nm, can be expressed as:

$$I_i(\lambda) = \sum_{k=0}^{K-1} \frac{p_{i,k}(\lambda) * l_k}{E_k(\lambda)} \prod_{k'}^{K-1} \left[1 - \frac{p_{i,k'}(\lambda) * \alpha_{i,k'}(\lambda)}{E_{i,k'}(\lambda)}\right] + \eta_i, \tag{5}$$

where $\alpha_{i,k}$ denotes the binary mask at a particular depth layer k, and E serves as a normalization factor.

Loss Function. We train the network using a feature similarity loss [30] for the RGB image $\mathcal{L}_{\text{RGB}}$, and calculate the $\mathcal{L}_1$ loss on initial disparity D_0 regressed from GEV and all predicted disparities from ConvGRU as $\mathcal{L}_{D_0}$ and $\mathcal{L}_{D_i}$. We also include the $\mathcal{L}_1$ loss of refined depthmap $\mathcal{L}_{D_r}$ after passing through the depth refinement network. For PSF regularization, we introduce a concentration mask M to enforce constraints on its divergence, as follows:

$$\mathcal{L}_{\text{PSF}} = \sum_{\lambda} \sum_{k} \text{M} \cdot |\text{PSF}_{\lambda,\text{k}}|^2. \tag{6}$$

As such, the total loss is defined as:

$$\begin{aligned} \mathcal{L} = \psi_{\text{PSF}}\mathcal{L}_{\text{PSF}} + \psi_{\text{RGB}}\mathcal{L}_{\text{RGB}} + \\ \psi_{\text{D}}(\mathcal{L}_{d_0} + \sum_{i}^{N} w_i \mathcal{L}_{D_i}) + \psi_{\text{D}_\text{r}}\mathcal{L}_{D_r}, \end{aligned} \tag{7}$$

where ψ_i denotes the weights of each loss, empirically set as $\psi_{\text{PSF}} : \psi_{\text{RGB}} : \psi_{\text{D}} = 1 : 1 : 1$, and weight w_i is computed in each GRU iteration to prioritize earlier iterations: $w_i = \left(\gamma^{15/(i-1)}\right)^{N-i-1}$, where N is the total iteration number and γ is set 0.9.

Our model demonstrates reduced dependency on the performance trade-off between RGB recovery and depth estimation that often appears in monocular deep optics models [10, 21]. This is because both RGB imaging and stereo matching prioritize extracting high-frequency information, and the fused depth cues are not solely dependent on the PSF distribution but also on sharp edges for matching.

4. Simulation Assessment

4.1. Data Preparation

Datasets. During initial training, we utilize a cleanpass subset of Sceneflow Flyingthings3D [22], which respectively comprises 22K and 8K pairs of synthetic RGB images and corresponding depth maps for training and testing. The input patch size is (320, 736) and the maximum disparity set in IGEV is 192. While the FlyingThings3D dataset offers variable depth maps aligned with RGB images, it is synthetic and does not reflect natural scenes accurately. Therefore, we additionally incorporate the InStereo2K dataset [3] to test the robustness of trained model. This dataset contains 2k captured images for training and 50 for testing.

Data Augmentation. To incorporate noise into simulated measurements (Eq. 6), we utilize a normal distribution with a maximum noise value of 0.005. To account for tolerances during DOE assembly and calibration, we introduce randomness to the DOE position at the Lens-Aperture-DOE plane. This randomness involves a 2D translation (x, y) by up to 4×4 pixels and a rotation by up to (θ) by 2°. In addition, we have introduced a simple but effective data augmentation technique involving mirrored sub-batches for the stereo imaging pipeline. Our network input comprises not only the measurements generated by the image formation model but also their mirrored counterparts. Consequently, two pairs of stereo images are included in each sub-batch, enabling the training of left-right RGB-D images and enhancing network training concurrently. Furthermore, since our model is only trained on synthetic datasets, we utilize the data augmentation methods from [19] in stereo matching, including horizontal image stretching, saturation adjustment and vertical perturbation of the right image.

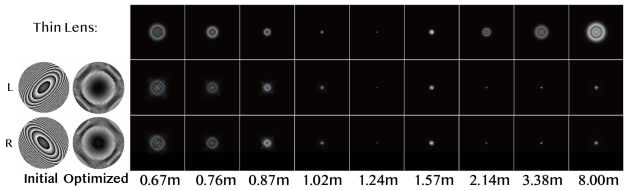


Figure 3. PSF visualization at varying depths. We simulate PSFs at 9 depths for the ideal thin lens (top row) and our DOE + thin lens optical system (middle and bottom rows). The left and right DOEs are initialized by enforcing the phase of compound cylindrical lenses with perpendicular spreading-out direction (left-most). After the end-to-end optimization, we can obtain learned DOE profiles (second column from left).

4.2. Implementation Details

In the simulation, we employ three principal wavelengths $\Lambda(\text{nm})=\{632, 550, 450\}$ across 7 depth layers $\mathcal{D}(\text{m})=\{0.67, 0.79, 0.96, 1.24, 1.72, 2.83, 8.00\}$, which are uniformly sampled in the diopter domain. As depicted in the right of Fig. 3, the rank-2 encoded depth-dependent PSF exhibits a more focused distribution compared to that of a thin-lens model, highlighting its ability to capture a broader frequency spectrum. In addition, these PSFs exhibit distinct extension directions to the left and right, perpendicular to each other across all depths except at the focal points, thereby providing complementary spatial sampling.

In this research endeavor, we have developed and trained two distinct sub-branch models tailored to different work distances. The medium-shot model exhibits exceptional precision in 3D imaging within the range of 1–5m, corresponding to approximately 0.8 diopter. In contrast, the long-shot model excels in imaging spanning 0.67–8m, approximately equivalent to 1.4 diopter.

Table 1. Ablation on varying neural network architectures. Evaluated models: depth from defocus with learned optics [10], simplified IGEV stereo matching [41] (IGEV-S), our stereo imaging network without DOEs (Baseline), and deep stereo models (D-S) using different DOE modeling methods, including pixel-wise (PW), rotational symmetric (Ring), rank-1 (Rank1), and rank-2 (Rank2) encoding. d denotes the diopter of working distance (unit: m^{-1}).

d (m^{-1})	Model	DOE	Image		Depth	
			PSNR	SSIM	RMSE	EPE (px)
0.8	DfD	Ring	31.02	0.905	0.132	--
	IGEV-S	w/o	--	--	0.078	1.52
	Baseline	w/o	29.83	0.897	0.076	1.38
	D-S	PW	31.88	0.915	0.071	1.23
	D-S	Ring	32.10	0.922	0.069	1.16
	D-S	Rank1	32.52	0.925	0.072	1.19
	D-S	**Rank2**	**32.96**	**0.926**	**0.066**	**1.12**
1.4	Baseline	w/o	29.17	0.832	0.090	1.49
	D-S	PW	31.08	0.890	0.082	1.37
	D-S	Ring	31.24	0.912	0.078	1.28
	D-S	Rank1	31.65	0.906	0.079	1.33
	D-S	**Rank2**	**32.13**	**0.917**	**0.071**	**1.21**

4.3. Simulation Results

We compare our deep stereo (D-S) framework with the simplified version of advanced stereo depth estimation algorithm IGEV [41] and monocular deep-optics imaging on the Scene Flow testset, as shown in Table 1. Detailed comparisons with recent deep optics methods [10, 33] are provided in the supplementary material. We further assess the imaging performance utilizing four different optical encoding methods in the DOE optimization: pixel-wise(PW), Ring, Rank-1, and our Rank-2 parameterization. As listed in Table 1, we assess benchmarks for depth estimation using RMSE and end-point error (EPE) matrices, while image recovery quality using PSNR and SSIM. The baseline is our proposed RGBD reconstruction framework without optical encoding (thin lens only). Furthermore, by utilizing the proposed end-to-end learned encoding D-S framework, optimized results were obtained through applying varying DOE representations. Results tested on the Sceneflow dataset demonstrate that our Rank-2 encoding, characterized by complementary encoding and superior focusing properties, yields superior performance in both image recovery and depth estimation. Specifically, we have achieved a PSNR of 32.12 dB in RGB reconstruction, surpassing the benchmark by 2.96 dB, and 1.21 px in EPE, decreased by 0.28 px. Figure 4 shows the qualitative simulation results.

5. Experimental Assessment

Prototype Implementation. We fabricate a pair of optimized DOEs and construct a stereo camera prototype, as illustrated in Fig. 5. The DOE fabrication involves iterative photolithography and dry etching processes on a fused silica wafer [6, 32] to achieve 2^4-level phase profiles. These DOEs have a clear diameter of 4.4 mm and utilize

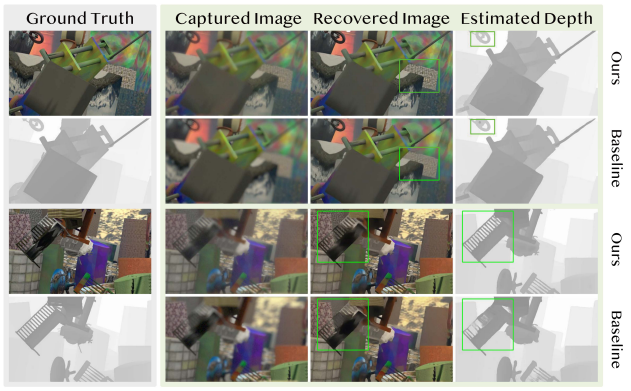


Figure 4. Comparison in simulation between the proposed method and baseline (w/o optical encoding).

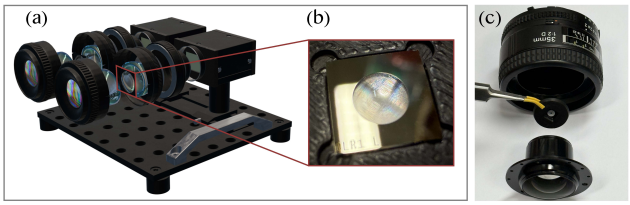


Figure 5. Diagrams of our learned stereo camera prototype and DOE assembling. (a) shows the 3D model of our prototype, consisting of four lens groups, two sensors, associated adapters, and two optimized DOEs placed at aperture planes (b–c).

a Chromium layer as an optical baffle. The equivalent f-number of our prototype is 8. Our camera setup consists of a pair of Nikkon EF 35mm f/1.8 SLR lenses and FLIR Grasshopper3 1/1.2" sensors.

To simulate the co-planar arrangement of lens, aperture, and DOE, we insert the DOE in the middle of the optical system, near the pupil plane. The manual SLR lenses allow for detaching the front and rear lens groups, facilitating the easy replacement of the aperture with custom DOEs [10]. To position the DOEs accurately, we have designed and 3D printed a pair of custom DOE holders with slots to aid in the 45° angle calibration. We construct our prototype by fixing the DOE plane at the aperture inside the compound lens, illustrated in Fig. 5(c). The f-number of an optical system should be determined at the entrance pupil, located at a specific distance from the aperture plane. Consequently, we delineate the magnification between the pupil and aperture and then fine-tune our pre-trained model by incorporating the magnified aperture. Our experiment shows that the magnification from pupil to aperture plane is only 1.031, which can be easily adjusted through network fine-tuning.

PSF Calibration and Model Fine-tuning. After prototyping the stereo camera consisting of left and right SLR lenses assembled with learned DOEs, as illustrated in Fig. 5, we need to capture the PSFs of our optical system. We

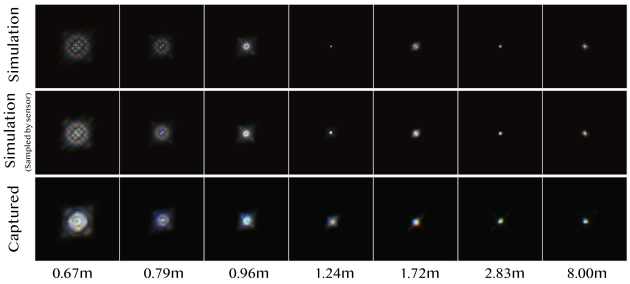


Figure 6. Captured and simulated depth-dependent PSFs. The designed PSF (top row) is optimized with our end-to-end simulator. Center row simulates the PSFs sampled and captured by our sensor. Optical imperfections, light source spectrum and size result in the captured PSFs (bottom row), slightly deviating from the simulated sensor images. The PSF patch size is 50×50px.

use a 25μm pinhole and a tunable laser-driven white light source as the point light to measure the depth-variant PSFs at depth layers $\mathcal{D}$. At each depth in $\mathcal{D}$, we captured ten images for average calculation, and five extra background images to be averaged and subtracted as noise. The bottom row in Fig. 6 shows the measured depth-variant PSFs, which slightly differ from the three-wavelength PSF simulation (top row). This discrepancy may originate from the spectrum distribution and size of the point light source, the misalignment, and the rotation error of left and right DOE positions. Since we considered the PSF jittering and rotating in the augmentation and end-to-end optimization, it has a tolerance and robustness within 5° and 40μm. Given calibrated PSFs, we incorporate the real PSF distribution and pupil-aperture magnification into our pre-trained model and fin-tune it for 3 epochs under the imaging size of (320, 736).

Real-world Results. We demonstrate the effectiveness of our stereo camera system with learned optics through four captured indoor and outdoor scenes (Fig. 8). The first row of each scene displays the captured images and RGBD reconstruction results from our prototype, while the second row compares the performance with a conventional stereo camera employing our baseline model. All processed RGB and depth maps in our experiments are at full resolution of 1,200×1,920 pixels.

Results indicate that our model excels in far-field scenarios, overcoming challenges encountered by conventional cameras such as shallow depth-of-field and potential degradation of image SNR with aperture adjustments. Leveraging complementary optical information encoding in the left and right channels enables us to retain edge details in coded blurry images, mitigating concerns like detail loss, distortion, and deformation in RGBD imaging. Our prototype performs effectively in demanding scenarios such as capturing intricate small features like the panda toy's head (Indoor-2) and the distant plants (Outdoor-1), and mitigates distortions caused by significant blur, as evidenced by the color checkerboard (Indoor-1) and shapes of the box and doll (Outdoor-2).

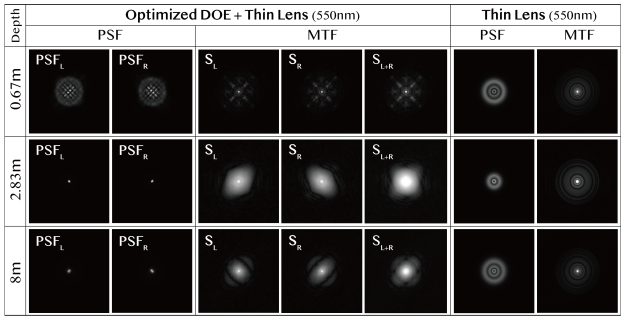


Figure 7. Ablation on sampling capabilities in the frequency domain of the DOE-lens hybrid optics. We present the PSFs at 3 depths: (0.67, 2.83, 8) m. MTF, indicated as S, represents the 2D modulation transfer function of the corresponding optics. It is evident that compared to traditional stereo camera with same lenses, our optimized stereo DOEs-lens system offers enhanced and comprehensive sampling in frequency domain, allowing our proposed camera to extract more spatial information.

6. Discussion and Conclusion

Unlike prior studies on stereo imaging, this work delves into evaluating the detailed interactions between hardware and software binocular vision systems, especially the interaction and optimization of stereo and focus cues. We have devised an encoding method that provides complementary spatial frequency information in the left and right camera to significantly improve both color and depth results over a large depth range in real world settings.

The complementarity of the MTFs between the two camera allows for a broader range of frequency components to be collectively encompassed [25]. This insight implies that a greater number of frequency components from the input signal can be preserved and transmitted, thereby enhancing the systems' ability to reconstruct the original signal. Conversely, in cases where the MTFs of two coded-aperture cameras are identical but orthogonal in the frequency domain, they can address complementary frequency components. Fig. 7 illustrates the simulated MTFs for the left and right cameras, denoted as $S_L(f)$ and $S_R(f)$, respectively. Ideally, the combined MTF for our stereo camera can be represented as $S_{L+R} = \mathbf{max}(S_L + S_R, 1)$ across most points (5^{th} column), covering a broader frequency spectrum, facilitating capturing more frequency information compared to a conventional stereo camera pair.

Limitations and Future Work. One limitation of the current system is that the computational complexity of our model prevents real-time image reconstruction. Additionally, experiments reveal artifacts in scenes with high dynamic range, e.g. directly visible light sources or specular highlights. For instance, circular DOEs can produce circular halos, while rank-1 and rank-2 DOEs may generate

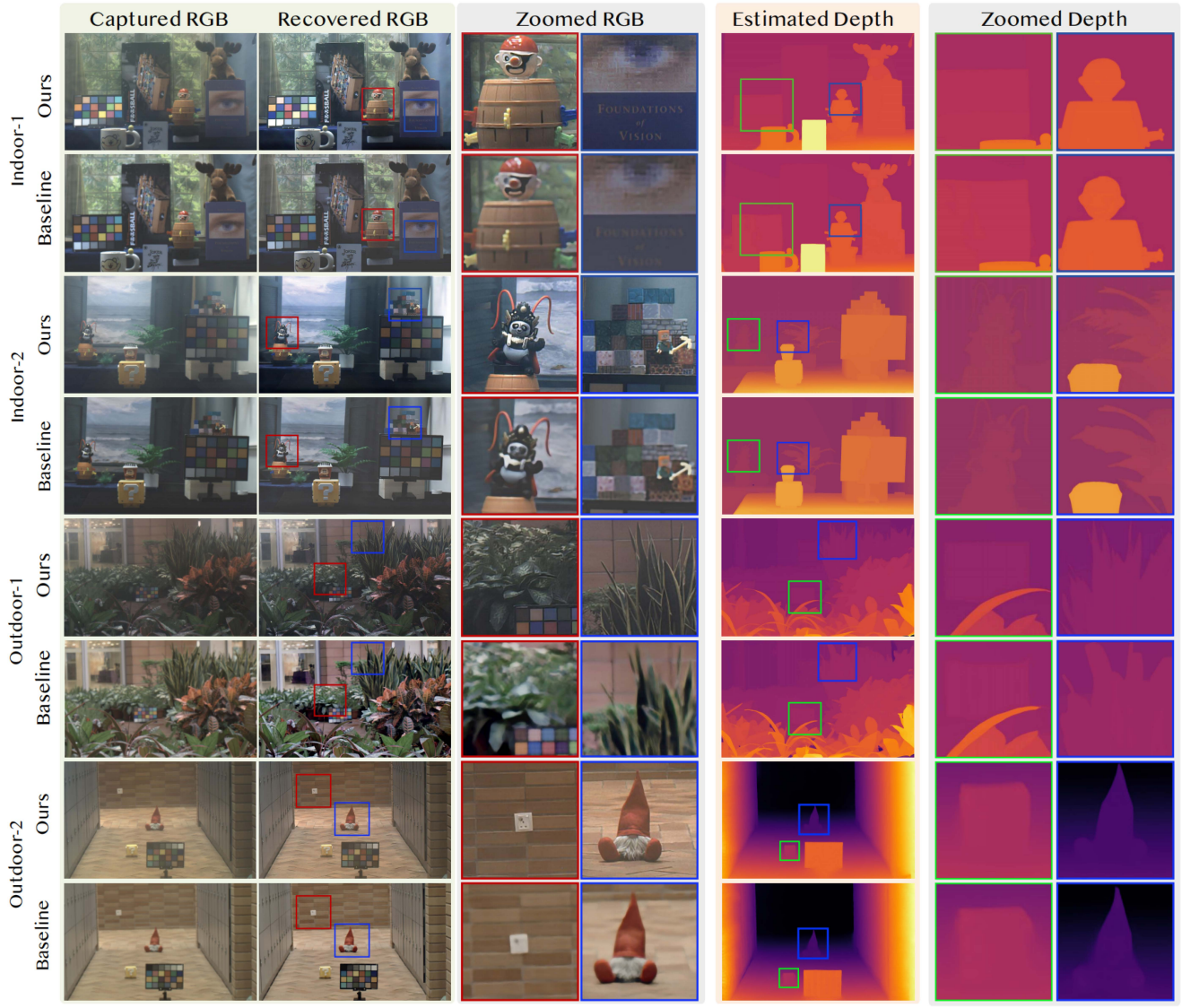


Figure 8. Experimental results of our learned stereo camera prototype. Baseline is the proposed stereo RGBD reconstruction network without optical encoding, aka., thin lens only. From left: Images captured by our stereo camera (Row 1 in each scene) and conventional stereo camera (Row 2 in each scene), AiF images recovered by our reconstruction network, zoomed-in comparison for recovered RGB images, and zoomed-in comparison for estimated depth maps.

cross-shaped due to "long tail" PSFs from limited diffraction efficiency of the DOE. This is at least in part due to prototyping challenges that could be substantially reduced with commercial grade fabrication and assembly.These challenges could also be mitigated by implementing optimization constraints and adjusting exposure settings appropriately. Future avenues also include exploring more advanced network architectures to improve the scalability of our end-to-end deep stereo framework, particularly targeting imaging tasks in extreme environments. Advancements in differentiable optical designs that are more resource-efficient and offer greater design flexibility are also worth investigating.

Conclusion. We presented a deep stereo imaging framework that jointly optimizes neural networks and a pair of complementary phase-coded apertures. By utilizing an appropriate initialization and a rank-2 encoding scheme, our stereo imaging model can acquire focus cues and complementary dual-channel information. Through intensive simulations and experiments, we observed significant enhancement in high-frequency image recovery and detailed depth estimation, particularly in far-field scenarios. Our stereo imaging framework also paves the way towards enabling various task-specific computational camera configurations, such as low-light and hyperspectral imaging scenarios.

Acknowledgment This work was partially supported by the National Science Foundation of China (62322217), the Research Grants Council of Hong Kong (GRF 17208023), the Innovation and Technology Fund of Hong Kong (ITP/062/24AP), and the Hong Kong Global STEM Professorship.

References

[1] Jorge Bacca, Tatiana Gelvez-Barrera, and Henry Arguello. Deep coded aperture design: An end-to-end approach for computational imaging tasks. *IEEE Transactions on Computational Imaging*, 7:1148–1160, 2021. 1, 3

[2] Seung-Hwan Baek, Hayato Ikoma, Daniel S Jeon, Yuqi Li, Wolfgang Heidrich, Gordon Wetzstein, and Min H Kim. Single-shot hyperspectral-depth imaging with learned diffractive optics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2651–2660, 2021. 3

[3] Wei Bao, Wei Wang, Yuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. Instereo2k: a large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 63:1–11, 2020. 5

[4] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 2

[5] Xiong Dun, Hayato Ikoma, Gordon Wetzstein, Zhanshan Wang, Xinbin Cheng, and Yifan Peng. Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging. *Optica*, 7(8):913–922, 2020. 3

[6] Qiang Fu, Hadi Amata, and Wolfgang Heidrich. Etch-free additive lithographic fabrication methods for reflective and transmissive micro-optics. *Opt. Express*, 29(22):36886–36899, 2021. 6

[7] Bhargav Ghanekar, Salman Siddique Khan, Pranav Sharma, Shreyas Singh, Vivek Boominathan, Kaushik Mitra, and Ashok Veeraraghavan. Passive snapshot coded aperture dual-pixel rgb-d imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25348–25357, 2024. 3

[8] Joseph W Goodman. *Introduction to Fourier optics*. Roberts and Company publishers, 2005. 3

[9] Jie Huang, Xueyang Fu, Zeyu Xiao, Feng Zhao, and Zhiwei Xiong. Low-light stereo image enhancement. *IEEE Transactions on Multimedia*, 25:2978–2992, 2022. 1, 2

[10] Hayato Ikoma, Cindy M Nguyen, Christopher A Metzler, Yifan Peng, and Gordon Wetzstein. Depth from defocus with learned optics for imaging and occlusion-aware depth estimation. In *2021 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2021. 1, 2, 3, 4, 5, 6

[11] Daniel S Jeon, Seung-Hwan Baek, Inchang Choi, and Min H Kim. Enhancing the spatial resolution of stereo images using a parallax prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1721–1730, 2018. 1, 2

[12] Daniel S. Jeon, Seung-Hwan Baek, Shinyoung Yi, Qiang Fu, Xiong Dun, Wolfgang Heidrich, and Min H. Kim. Compact snapshot hyperspectral imaging with diffracted rotation. *ACM Trans. Graph.*, 38(4), 2019. 3

[13] Zhicheng Ji, Huan Zheng, Zhao Zhang, Qiaolin Ye, Yang Zhao, and Mingliang Xu. Multi-scale interaction network for low-light stereo image enhancement. *IEEE Transactions on Consumer Electronics*, 70(1):3626–3634, 2023. 1

[14] Wonwoo Lee, Nohyoung Park, and Woontack Woo. Depth-assisted real-time 3d object detection for augmented reality. In *ICAT*, pages 126–132, 2011. 1

[15] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. *ACM transactions on graphics (TOG)*, 26 (3):70–es, 2007. 3

[16] Rui Li, Dong Gong, Wei Yin, Hao Chen, Yu Zhu, Kaixuan Wang, Xiaozhi Chen, Jinqiu Sun, and Yanning Zhang. Learning to fuse monocular and multi-view cues for multi-frame depth estimation in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21539–21548, 2023. 4

[17] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X Creighton, Russell H Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6197–6206, 2021. 1

[18] Jianxin Lin, Lianying Yin, and Yijun Wang. Steformer: Efficient stereo image super-resolution with transformer. *IEEE Transactions on Multimedia*, 25:8396–8407, 2023. 1

[19] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 2, 4, 5

[20] Guoxuan Liu, Ning Xu, Huaidong Yang, Qiaofeng Tan, and Guofan Jin. Miniaturized structured illumination microscopy with diffractive optics. *Photonics Research*, 10(5):1317–1324, 2022. 3

[21] Xin Liu, Linpei Li, Xu Liu, Xiang Hao, and Yifan Peng. Investigating deep optics model representation in affecting resolved all-in-focus image quality and depth estimation fidelity. *Optics Express*, 30(20):36973–36984, 2022. 3, 5

[22] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016. 5

[23] Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1375–1385, 2020. 1, 3

[24] Elias Nehme, Daniel Freedman, Racheli Gordon, Boris Ferdman, Lucien E Weiss, Onit Alalouf, Tal Naor, Reut Orange, Tomer Michaeli, and Yoav Shechtman. Deepstorm3d: dense 3d localization microscopy and psf design by deep learning. *Nature methods*, 17(7):734–740, 2020. 3

[25] Stephen K Park, Robert Schowengerdt, and Mary-Anne Kaczynski. Modulation-transfer-function analysis for sampled image systems. *Applied optics*, 23(15):2572–2582, 1984. 7

[26] Yifan Peng, Qiang Fu, Hadi Amata, Shuochen Su, Felix Heide, and Wolfgang Heidrich. Computational imaging using lightweight diffractive-refractive optics. *Optics express*, 23(24):31393–31407, 2015. 3

[27] Yifan Peng, Qilin Sun, Xiong Dun, Gordon Wetzstein, Wolfgang Heidrich, and Felix Heide. Learned large field-of-view imaging with thin-plate optics. *ACM Trans. Graph.*, 38(6): 219–1, 2019. 2, 3

[28] Zheng Shi, Yuval Bahat, Seung-Hwan Baek, Qiang Fu, Hadi Amata, Xiao Li, Praneeth Chakravarthula, Wolfgang Heidrich, and Felix Heide. Seeing through obstructions with diffractive cloaking. *ACM Transactions on Graphics (TOG)*, 41(4):1–15, 2022. 3

[29] Zheng Shi, Ilya Chugunov, Mario Bijelic, Geoffroi Côté, Jiwoon Yeom, Qiang Fu, Hadi Amata, Wolfgang Heidrich, and Felix Heide. Split-aperture 2-in-1 computational cameras. *ACM Transactions on Graphics (TOG)*, 43(4):1–19, 2024. 2, 3

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[31] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018. 2, 3

[32] Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1386–1396, 2020. 1, 2, 3, 6

[33] Shiyu Tan, Yicheng Wu, Shoou-I Yu, and Ashok Veeraraghavan. Codedstereo: Learned phase masks for large depth-of-field stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7170–7179, 2021. 3, 6

[34] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2

[35] Ethan Tseng, Ali Mosleh, Fahim Mannan, Karl St-Arnaud, Avinash Sharma, Yifan Peng, Alexander Braun, Derek Nowrouzezahrai, Jean-Francois Lalonde, and Felix Heide. Differentiable compound optics and processing pipeline optimization for end-to-end camera design. *ACM Transactions on Graphics (TOG)*, 40(2):1–19, 2021. 3

[36] Longguang Wang, Yulan Guo, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, and Wei An. Parallax attention for unsupervised stereo correspondence learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):2108–2125, 2020. 1

[37] Xianqi Wang, Gangwei Xu, Hao Jia, and Xin Yang. Selective-stereo: Adaptive frequency information selection for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19701–19710, 2024. 2

[38] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8445–8453, 2019. 1

[39] Haoyu Wei, Xin Liu, Xiang Hao, Edmund Y Lam, and Yifan Peng. Modeling off-axis diffraction with the least-sampling angular spectrum method. *Optica*, 10(7):959–962, 2023. 3, 4

[40] Gangwei Xu, Junda Cheng, Peng Guo, and Xin Yang. Attention concatenation volume for accurate and efficient stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12981–12990, 2022. 2

[41] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023. 1, 2, 4, 6

[42] Gangwei Xu, Yun Wang, Junda Cheng, Jinhui Tang, and Xin Yang. Accurate and efficient stereo matching via attention concatenation volume. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2

[43] Menglong Ye, Edward Johns, Ankur Handa, Lin Zhang, Philip Pratt, and Guang-Zhong Yang. Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery. *arXiv preprint arXiv:1705.08260*, 2017. 1

[44] Minghua Zhao, Xiangdong Qin, Shuangli Du, Xuefei Bai, Jiahao Lyu, and Yiguang Liu. Low-light stereo image enhancement and de-noising in the low-frequency information enhanced image space. *arXiv preprint arXiv:2401.07753*, 2024. 1, 2

[45] Xiangyuan Zhu, Kehua Guo, Hui Fang, Liang Chen, Sheng Ren, and Bin Hu. Cross view capture for stereo image super-resolution. *IEEE Transactions on Multimedia*, 24:3074–3086, 2021. 1, 2