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

Supplementary Material

A. Details of Image Formation

A.1. Differentiable Propagation Modeling

As mentioned in Section 3.1 and shown in Figure 1 of the main text, we approximate the optical system of each camera with a model consisting of a thin lens, a DOE, and an aperture stop, all co-located at the same plane. This model approximates our physical prototype, in which the DOE and aperture stop are placed in the pupil plane of a compound lens. The approximation is valid under two assumptions: (1) the compound lens is well-corrected, i.e., its optical aberrations are negligible; and (2) the magnification of the pupil plane by the lens is accounted for in the model. For the latter issue, we note that we have experimentally verified that the Nikon lens in our setup offers close to unit magnification of the pupil plane; i.e., the entrance pupil is approximately 1.03 times the diameter of the physical aperture stop. Due to the small magnification factor and the measurement uncertainty, we therefore design our DOE with unit magnification, and rely on network fine-tuning (Section D.1) to compensate for the deviation.

With this in mind, we now derive the details for the simplified model using a co-located thin lens. As in most camera systems, the point spread function (PSF) for our optical system is depth-dependent. In this work, we empirically focus the lens at a nominal distance of 1.23m, given the target scene depth range from 0.67m to 8m, which is equivalent to 1.4 diopter.

Our optical modeling of the PSF begins with a point source placed in front of the camera system. At each depth, the point source generates a spherical wave. Upon the arrival of the wavefront at the Lens-Aperture-DOE plane, the wavefront can be expressed as:

$$\boldsymbol{u_0} = A_0 e^{jk\sqrt{(x'^2+y'^2+z'^2)}}, \tag{1}$$

where A_0 denotes the amplitude of the input light, $k = 2\pi/\lambda$ is the wave number related to wavelength λ, and (x', y') represents the 2D spatial coordinates at the Lens-Aperture-DOE plane, z indicates the distance from the point source and aperture center.

According to the simplified model, the light wave first encounters a thin lens, which is responsible for focusing the image and allows the DOE to undertake the optimized encoding operation independently. Thus, the phase delay $\Delta\phi_f$ introduced by the lens with the focal length f is defined as:

$$\Delta\phi_f = \frac{k}{2f}(x'^2 + y'^2). \tag{2}$$

Subsequently, the wavefront after passing through the thin lens, denoted as $\boldsymbol{u_1}$, can be represented as:

$$\boldsymbol{u_1} = \mathcal{A}(x', y')\boldsymbol{u_0} e^{j\Delta\phi_f}, \tag{3}$$

where $\mathcal{A}(x', y')$ represents the aperture of this optical system. The complex wave field after the phase modulation of DOE, as described in [3], can be expressed as:

$$\boldsymbol{u_{z_0}} = \mathcal{A}(x', y')\boldsymbol{u_1} e^{jk(n(\lambda)-n_0)\boldsymbol{H}(x',y')}, \tag{4}$$

where n_λ is the wavelength-dependent refractive index of the substrate, and $\boldsymbol{H}(x', y')$ is the DOE's height map.

After traversing the aperture plane, the incident light has accumulated phase variation:

$$\Delta\phi = k\Big[(n(\lambda) - n_0)\boldsymbol{H}(x', y') + \frac{1}{2f}(x'^2 + y'^2)\Big]. \tag{5}$$

We apply the vanilla angular spectrum method (ASM) [3] to simulate the wave propagation from the scene to the sensor, traversing both the learnable DOE and the thin lens. The ASM can strictly portray the scalar wave propagation process, which guarantees the simulation accuracy for the scene from different fields of view. The diffractive wave field at the sensor imaging plane (x, y, z) can be calculated by the following:

$$\boldsymbol{u_z}(x, y) = \mathcal{F}^{-1}\Big\{\mathcal{F}\{\boldsymbol{u_{z_0}}(f_X, f_Y)\} \cdot \mathcal{H}_z(f_X, f_Y)\Big\}, \tag{6}$$

where $\mathcal{F}$ is the Fourier transformation operator, and $\mathcal{H}$ is the light transport term. Therefore, the PSF of this imaging system captured by sensors can be specified as:

$$p = \big|\boldsymbol{u_z}(x, y)\big|^2, \tag{7}$$

which represents the amplitude of the wave field at the imaging plane z.

To enforce an accurate, fast, and flexible wave propagation in DOE optimization, we apply the LS-ASM [8] in simulation, which effectively specifies the least samplings in the imaging process for simulation and optimization.

A.2. Measurements with Sensor Responses

In our deep stereo optics framework, we can derive the PSFs (p_l, p_r) for the left and right cameras, expressed as:

$$p_{l,r} = \Big|\mathcal{F}^{-1}\Big\{\mathcal{F}\{\boldsymbol{u_0}'^{+}_{l,r}(f_X, f_Y)\} \cdot \mathcal{H}_z(f_X, f_Y)\Big\}\Big|^2. \tag{8}$$

In the real-world scenarios, light sources are generally incoherent. Therefore, when the light waves emitted from different points in the scene contribute to the same pixel on the image plane, the measured light intensity of the pixel is the linear summation of the intensities of these point sources [8]. Considering the image formed by the target scene through an ideal optical system, its light intensity distribution is $I(x, y, \lambda)$, where (x, y) represents the coordinates of the image plane, and λ is the wavelength. In general, the sensor captures images with N_c channels, and the image of the c^{th} channel ($c = 1, 2, ..., N_c$) can be defined as:

$$M_c(x, y) = \sum_{\Lambda} R_c(\lambda)\Big\{\Omega\big[I(x, y, \lambda) * p_\lambda(x, y)\big]\Big\} + \eta_c(x, y), \tag{9}$$

where $\Lambda = \{632, 550, 450\}$nm is the considered spectrum, and the wavelength $\lambda \in \Lambda$, $*$ represents the 2D convolution, $\eta_c(x, y)$ is the corresponding noise, N_c is the number of channels, Ω represents the image plane modulation, which linearly transforms $I(x, y, \lambda)$ into another 3D function, $p_\lambda(x, y)$ is PSF at the wavelength λ, and $R_c(\lambda)$ is the spectral response function of channel c. The last three parts correspond to the three types of wavefront encoding: image plane encoding, PSF encoding, and spectral response encoding.

$I' = I(x, y, \lambda) * p_\lambda$ describes the forward imaging model and the depth-variant PSF p_λ facilitates different encodings in the imaging process. Note that the sampling process of the sensor discretizes the data and herein we ignore the image plane variation and assume Ω as an identity matrix.

Therefore, for a given scene $I_{sce.}(x, y, \lambda)$, the stereo camera measurement can be modeled as:

$$I_{l,r} = \sum_{\Lambda} R_c(\lambda)\big[I_{sce.}(x, y, \lambda) * p_{l,r}(x, y, \lambda)\big] + \eta_{l_c, r_c}(x, y), \tag{10}$$

where the left PSF $p_l(x, y, \lambda)$ describes the image formation process of the left capture while $p_r(x, y, \lambda)$ is its counterpart for the right capture. After optimizing the PSFs of the imaging system, we apply the sensor response curves to derive the captured PSFs at the sensor plane.

A.3. Sampling Details in Simulation

In our sampling strategy at the aperture plane, we utilized an oversampling factor of 1.04 at a wavelength of 450nm using the LS-ASM approach. This oversampling factor corresponds to a spatial sampling number of (1,260 $\times$ 1,260) for a 35mm lens with the f-number of 8, resulting in a frequency sampling number of (1,600 $\times$ 1,600). Analysis of the aliasing situation in the frequency spectrum indicates that our sampling is sufficient and appropriate. The spatial sampling pitch is $3.5\mu\text{m} \times 3.5\mu\text{m}$ in physical scale. At the DOE aperture plane, we set the upsampling factor to 2, indicating that the size of optimizable height-map is (630 $\times$ 630), and the DOE pixel pitch is $7\mu\text{m} \times 7\mu\text{m}$.

B. Comparison of DOE Modeling Methods

In our simulations, we have optimized two sub-branch models targeting different depth ranges, as illustrated in Fig. 1. The PSF distribution optimized for the range of 1 to 5 meters exhibits a more concentrated distribution within the range of 0.8 to 3 meters. This concentration results in a faster convergence speed of the optimized DOE. On the other hand, the branch model optimized for the $0.67 - 8\text{m}$ (1.4 diopters) range demonstrates a relatively concentrated state within the $1 - 8\text{m}$ range, significantly enhancing the capability to capture high-frequency information in far-field scenarios, thereby offering practical utility for a wide range of scenes. The subsequent experimental comparisons are all based on discussions conducted using the 1.4-diopter branch model, with the focal distance set as 1.24m. Overall, we have implemented four DOE encoding schemes in our Deep-Stereo model (Fig. 2) and optimized them in an end-to-end manner.

B.1. Analysis of DOE Initialization

The asymmetric initialization plays a crucial role in providing complementary sampling for the left and right channels. As shown in Fig. 3, we have optimized the left and right DOEs using five different initialization approaches under our Deep-Stereo framework. When using symmetric initialization methods like zeros, it becomes challenging for the networks to escape a local minimum and break symmetry to generate the necessary complementary sampling for the left and right channels. This behavior is evident in the cylindrical and perpendicular cylindrical initialization, as illustrated in Row 2 and Row 4 in Fig. 3. Average PSNR and EPE metrics are presented. Compared with zero initialization, the cylindrical phase introduces the optical focal power of the DOEs, resulting in increased phase variation.

On the other hand, perpendicular cylindrical initialization combines two directional cylindrical phases with different phase variation ranges to provide focal power in perpendicular directions and achieve asymmetric initialization. The optimization results in the simulation indicate that both the rotated cylindrical and perpendicular cylindrical phases can be optimized into DOEs capable of generating complementary PSFs for the left and right channels. However, the RGB imaging quality of the rotated cylindrical initialization method appears more dispersed and blurred compared to the perpendicular cylindrical approach.

B.2. Analysis of DOE Parameterization

The Rank-2 modeling has the capability of initializing the matrix using a perpendicular cylindrical phase with the low-

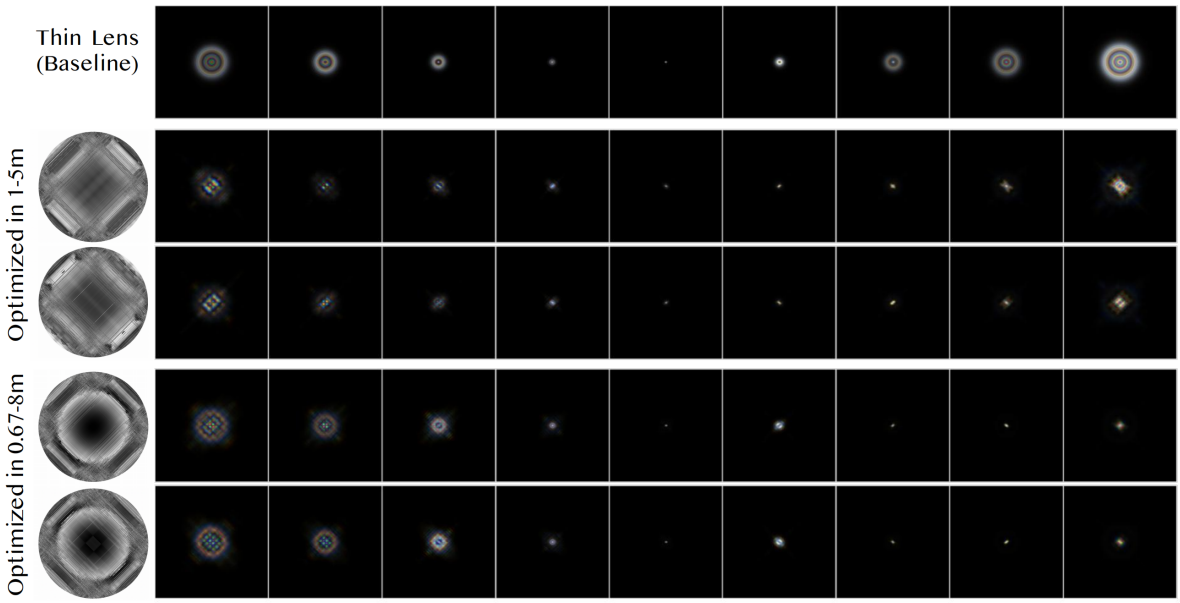

Figure 1. The PSFs of two optimized optical systems across varying depth ranges are presented. Row 1 represents the baseline thin lens system. Rows 2 and 3 depict the optimization of the left and right imaging channels within the range of $1\mathrm{m} - 5\mathrm{m}$. Rows 4 and 5 showcase the learned PSFs for the left and right cameras covering the range of $0.67\mathrm{m} - 8\mathrm{m}$.

est parameter count compared with other methods we have tested. For a DOE with a sampling size of $m \times m$, a pixel-wise design space offers m^2 degrees of freedom (DoF), which can be challenging for optimization when m is a large value. Apart from Rank-2 modeling, we optimized two frequently utilized DOE encoding approaches for comparison, namely the ring design with rotational symmetry [4] and the Rank-1 design [6]. By comparison, the ring model and the low-rank model are encoded by a small number of k vectors with a length of the mask, corresponding to $k \cdot m$ DoF, with $k = 1$, 2, or 4 for the ring, rank-1 and rank-2 models, respectively. These models are limited in their ability to encode matrices with asymmetric or perpendicular cylindrical phases. We have evaluated the PSFs in Fig. 2 and RGBD imaging results optimized via different DOE modeling schemes in Fig. 11, where the real-captured results were all directly generated by simulation-optimized models without further model fine-tuning.

C. Ours vs. Other Deep-Optics Methods

In this supplementary section, we provide a comprehensive comparison between our proposed Rank-2 Stereo method and several state-of-the-art deep-optics approaches, namely Monocular Depth-from-Defocus (Mono-DfD) [4], Coded Stereo [7], and Ring-coded Stereo using our proposed network architechure. Our analysis is based on both qualitative observations and quantitative simulation results, as summarized in Table 1 and illustrated in Fig. 5 and 6.

Our simulation studies (see Fig. 5) were conducted using the same stereo matching algorithm and UNet architecture across all methods for fairness. In these simulations, the Coded Stereo approach was modeled with identical rotational encoding for both cameras—omitting the benefits of view warping and RGB-depth fusion—which our Rank-2 Stereo method fully exploits.

Structural & Hardware Design: Unlike methods that extend the depth-of-field or rely solely on defocus cues, our Rank-2 Stereo employs a complementary encoding scheme that enables physical interaction between the left and right channels. In contrast, Coded Stereo [7] uses identical rotational encoding for both cameras. The identicalsymmetrical DOE encoding (Ring), despite achieving comparable extending DoF results, suffers from hazier RGB reconstructions due to lower diffraction efficiency and lacks physical information interaction between left and right channel, particularly affecting high-frequency details, as evidenced in Fig. 6. The same rotational encoding on both channels can not exploit the mutual benefits available through RGB-depth interaction and view warping.

Algorithm Design: Our method leverages joint stereo and focus information to reconstruct RGB and estimate depth. A key difference from [7] is the warping of the right view to the left perspective, which facilitates the fusion of overlapping regions for enhanced image restoration and improved depth estimation. By comparison, Coded Stereo [7] decouples the processes for RGB reconstruction and depth estimation, while Mono-DfD relies exclusively on defocus

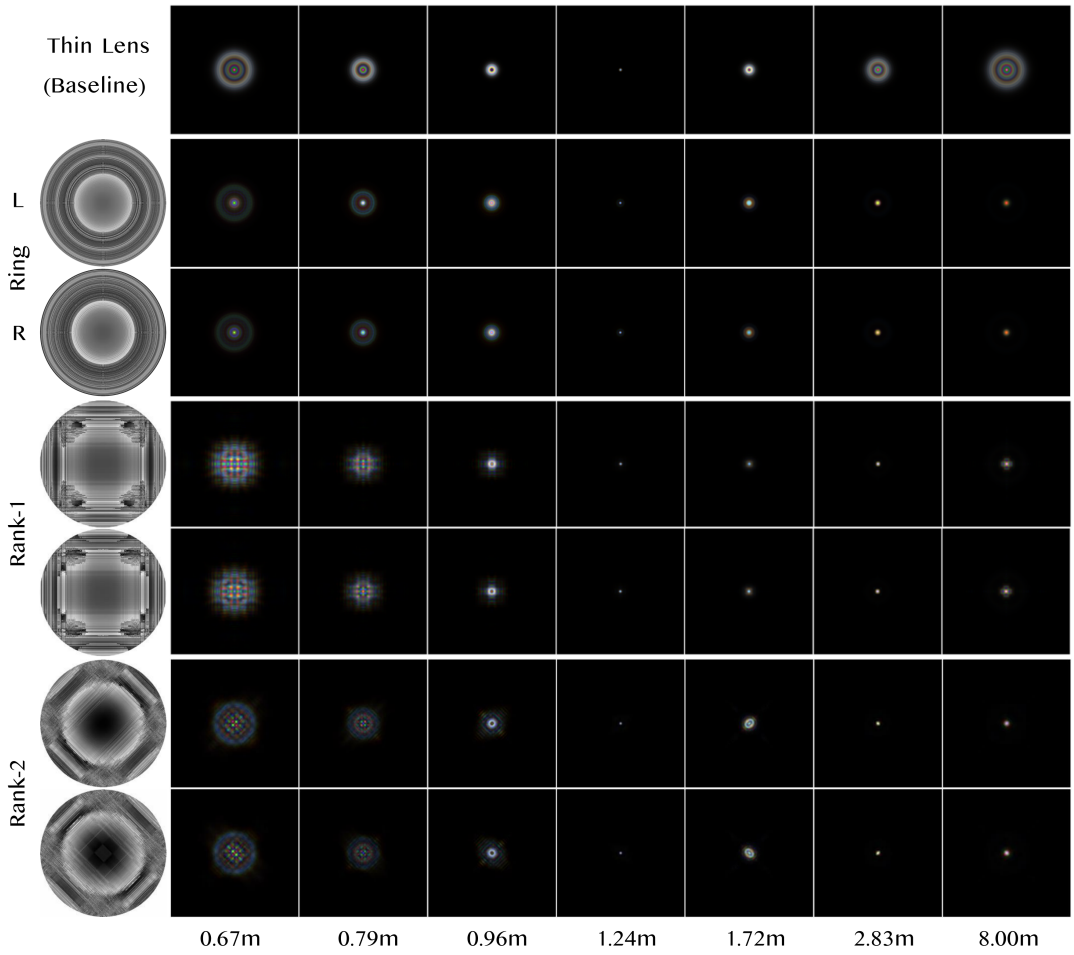

Figure 2. PSF distributions under different learned DOE profiles, ranging from 0.67m to 8m. Row 1 represents our baseline thin-lens system without DOE, which is consistent for both the left and right cameras. The subsequent rows depict thin lens+DOE imaging systems optimized using various encoding approaches. Rows 2–3 and Rows 4–5 are optimized using Ring modeling and Rank-1 modeling, demonstrating small differences and complementary characteristics between the left and right channels. Rows 6–7 showcase our proposed DOEs optimized through the Rank-2 modeling method. It is evident that they exhibit differentiation between the left and right channels for acquiring complementary information while simultaneously displaying far-field focusing properties.

Table 1. Comparison of Deep-Optics Methods.

Method	**DOE**	**Algorithm**	**Depth range**	**RGB**-PSNR/SSIM	**D**-RMSE/EPE
Mono-DfD [4]	Ring	Defocus-based U-Net	0.8 diopter	29.13/0.889	0.139/ – –
Ring-coded stereo [7]	Ring	Separate SDE + RGB recon	0.84 diopter	31.06/0.902	0.090/1.41
Our Ring-coded	Ring	Fused SDE+ RGBD	1.4 diopter	31.24/0.912	0.078/1.28
Our rank2-coded	Rank2	Fused SDE+ RGBD	1.4 diopter	**32.13/0.917**	**0.071/1.21**

cues from a monocular view. Moreover, our Rank-2 Stereo method extends the effective diopter range to 1.4 diopter, in contrast to the 0.84 diopter range demonstrated by Coded Stereo [7].

Overall, the integration of complementary hardware design and a unified algorithmic framework yields superior quantitative and qualitative RGBD performance.

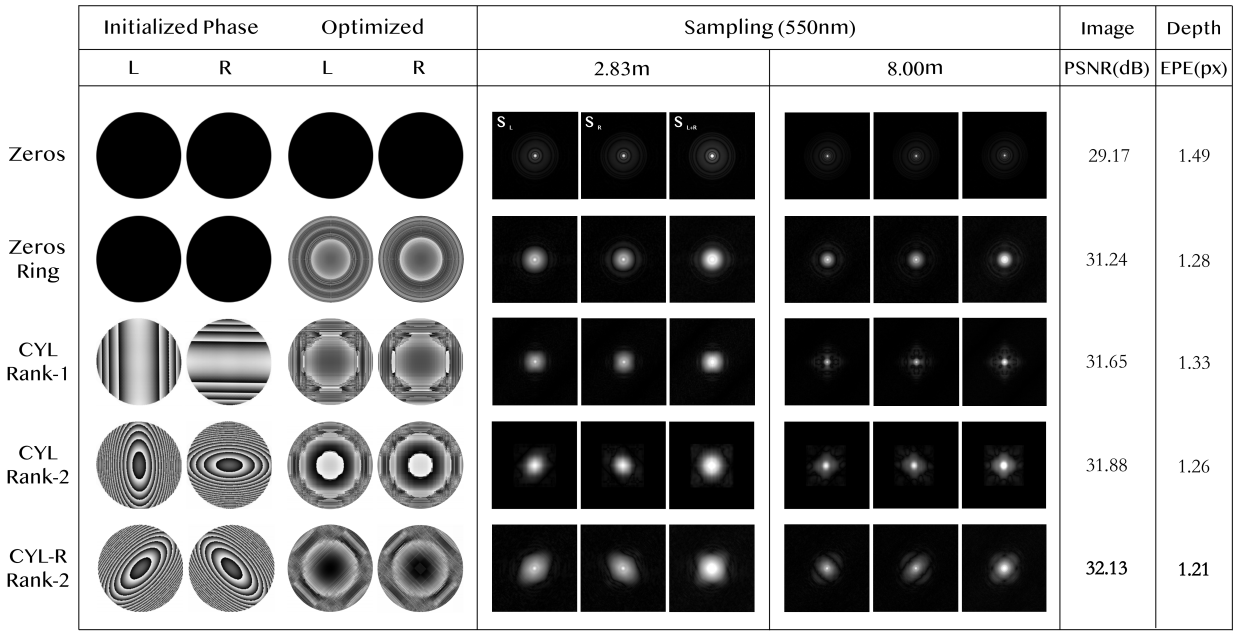

Figure 3. Optimized DOE profiles with different initialization schemes. Row 1 represents the thin-lens baseline, Row 2 is the Ring modeling with zero initialization. Row 3–5 are the cylindrical phase initialization (CYL). We present the Peak Signal-to-Noise Ratio (PSNR) and depth Endpoint Error (EPE) values for each initialization and encoding method on the rightmost side, as evaluated using the Scene Flow datasets[5].

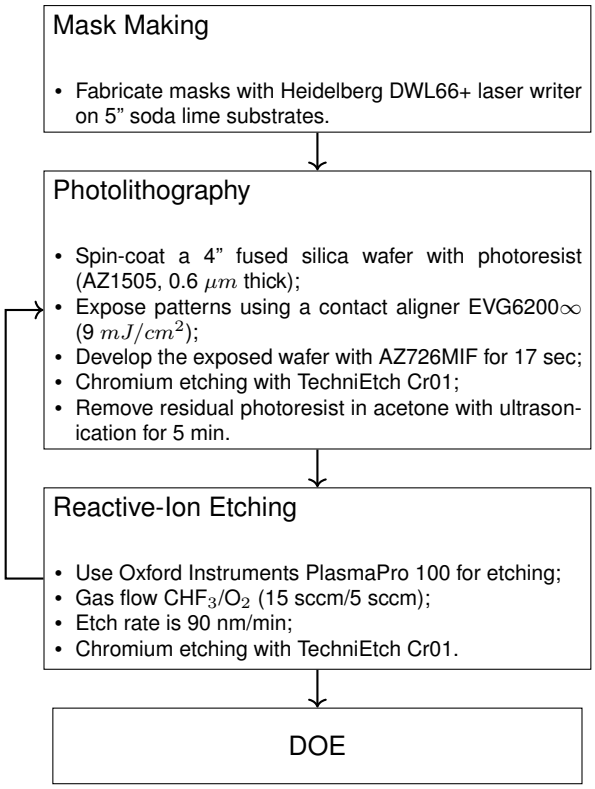

Figure 4. DOE fabrication workflow.

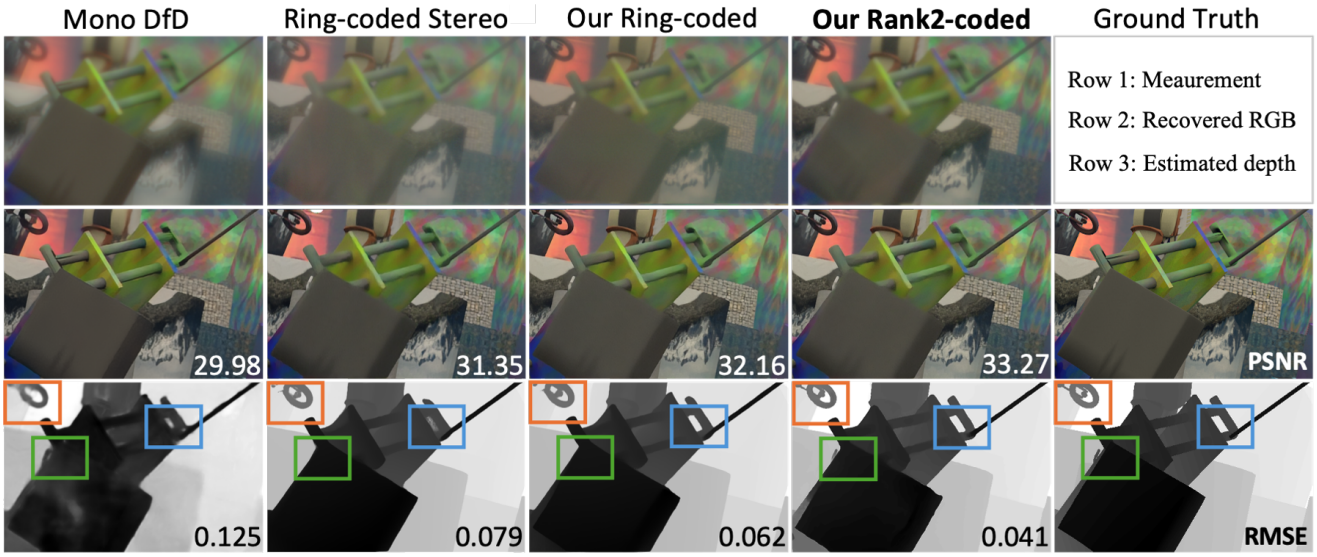

Figure 5. Simulation comparisons of RGB and depth estimation between monocular depth-from-defocus (Mono DfD), identical coded stereo (Ring-coded Stereo)[7], ring-coded stereo using our architecture (Our Ring-coded), and our proposed Rank2-coded Stereo. Zoom-in views highlight the improved reconstruction of high-frequency details by our Rank-2 Stereo method compared to Ring-coded and Coded Stereo approaches.

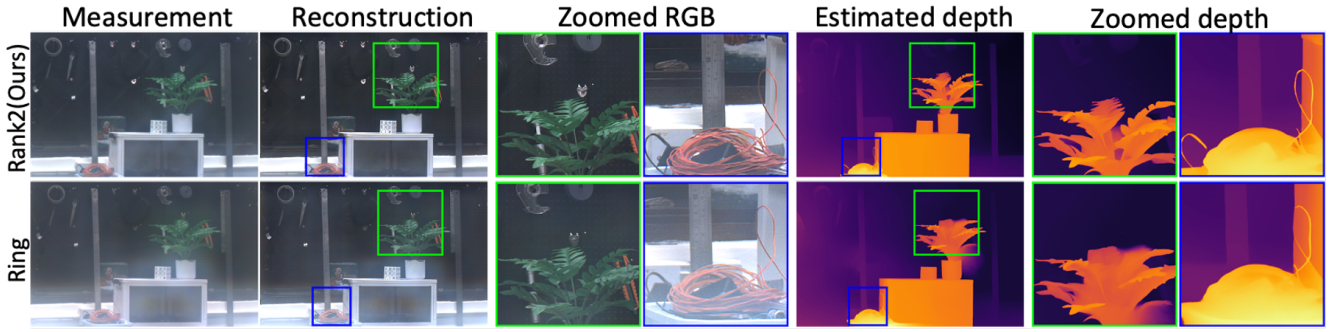

Figure 6. Experiment comparisons of RGB and depth estimation between rank2-coded and ring-coded stereo.

D. Model Fine-tuning and Prototype Details

D.1. Neural Network Model Fine-tuning

Our ability to create phase profiles on fused silica wafers is limited to 2^4 levels due to constraints imposed by photolithography and dry etching techniques. Prior to fabrication, it is essential to quantize the optimized DOEs, as depicted in Fig. 7. It is worth noting that there are minimal discrepancies between the PSFs of continuously designed DOEs and their quantized versions, as illustrated in Fig. 8. This indicates that the quantization process does not significantly impact the performance of the optimized DOEs in terms of PSF quality.

After the optimization process, a two-step fine-tuning approach is implemented. Firstly, following the end-to-end training, the model undergoes fine-tuning using the quantized DOEs. Furthermore, to simulate wide-spectrum PSFs that are more representative of real-world photography scenarios, Gaussian blur is applied. This additional step enhances the model's ability to generate more realistic and versatile PSFs across a wider spectrum of wavelengths, improving the overall performance and applicability of the model.

In the second phase of fine-tuning, which occurs after DOE fabrication, a uniform white light source with a 25μm aperture is employed to capture the actual PSF distribution of the stereo DOE-thin lens system. Given the asymmetrical and $45°$-rotated nature of our DOE design, which differs from conventional lens and ring-coded DOEs, post-assembly calibration is necessary when integrating the DOEs into standard lens groups. During the calibration process, near-field PSFs are captured using high exposure levels. Overexposure leads to the captured PSF of a point light source assuming a cross-star shape, facilitating straightfor-

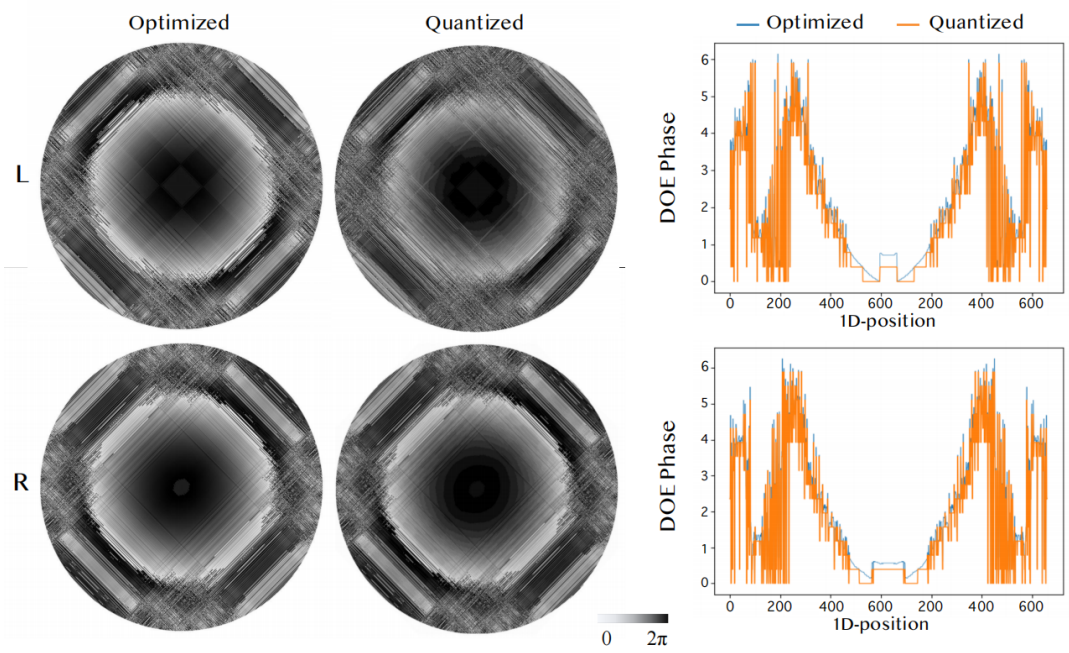

Figure 7. Quantization of DOE patterns for fabrication.

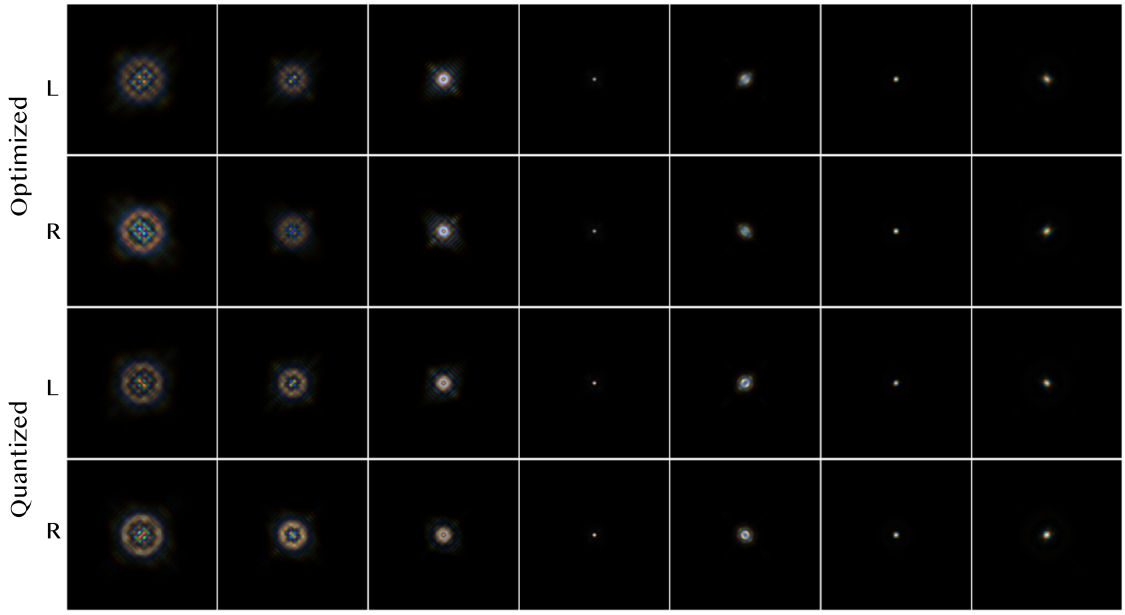

Figure 8. PSFs generated using optimized DOEs with the continuous (top-group) and the 16-level quantized (bottom-group) heightmaps.

ward calibration in the $x - y$ direction. After acquiring 3D PSFs, each PSF is centered based on its maximum value and used to refine our pre-trained model.

The final model undergoes training for 3 epochs, with images sized at (320 $\times$ 736). In Fig. 10, we compare the imaging results before and after fine-tuning the trained model. The comparison shows that there is slight enhancement in the quality of RGBD imaging by reducing halos and restoring colors in RGB image recovery. The observed differences in these effects fall within an acceptable range. It is evident that the end-to-end trained model exhibits strong generalization capabilities, showcasing its ability to adapt and improve performance through fine-tuning processes.

Figure 9. Comparison of captured and simulated PSFs for our learned stereo camera across the depth range from 0.67m to 8m. Row 1 and 4 show the PSFs captured by our stereo camera prototype, while Row 2 and 5 showcase the simulated PSFs with a measured patch size of 50×50 pixels. Row 3 and 6 indicate the simulated PSFs with a higher sampling at three principal wavelengths: 632nm, 550nm, 450nm.

D.2. Additional Prototype Details

To address the challenge of precisely placing the DOE plates at the entrance pupil plane, it is essential to measure and calculate the magnification between the entrance pupil and the DOE plane (aperture). For our Nikon 35mm f/2D lens, the magnification factor is $M_{E \to A} = 0.97$, and the dimensions of our DOE are $7\mu\text{m} \times 630 = 4.41\text{mm}$. With the DOE positioned at the aperture plane, the equivalent F number is calculated to be 8.18.

We are providing a supplementary video clip, **prototyperesults.mp4**, which presents an overview of our learned stereo camera prototype, particularly illustrating the positioning of DOEs within a pair of lenses, and the image-capturing process.

D.3. DOE Fabrication

Various photolithography techniques are available [1, 2] for DOE fabrication. The micro-structures in the optimized DOEs consist of mainly high-frequency spatial features that require depth-preservation in the vertical direction. Therefore, we adopt the well-established etching-based fabrication techniques that have been widely used in similar diffractive imaging tasks [6]. The fabrication workflow is illustrated in Fig. 4. The fabrication cycle consists of three major blocks, including mask making, photolithography, and reactive-ion etching. Key parameters in each step are listed in each corresponding block. By repeating the photolithography and reactive-ion etching for 4 times, a 16-level DOE can be fabricated.

E. Additional Experimental Results

The captured scene in Fig. 11 presents challenges in lighting and color distribution due to the dark lighting of the pure black background and uneven foreground lighting conditions. While Ring and Rank-1 encodings can capture more high-frequency information compared to the thin-lens baseline model, they are more susceptible to variations in light source illumination and object reflections. This sensitiv-

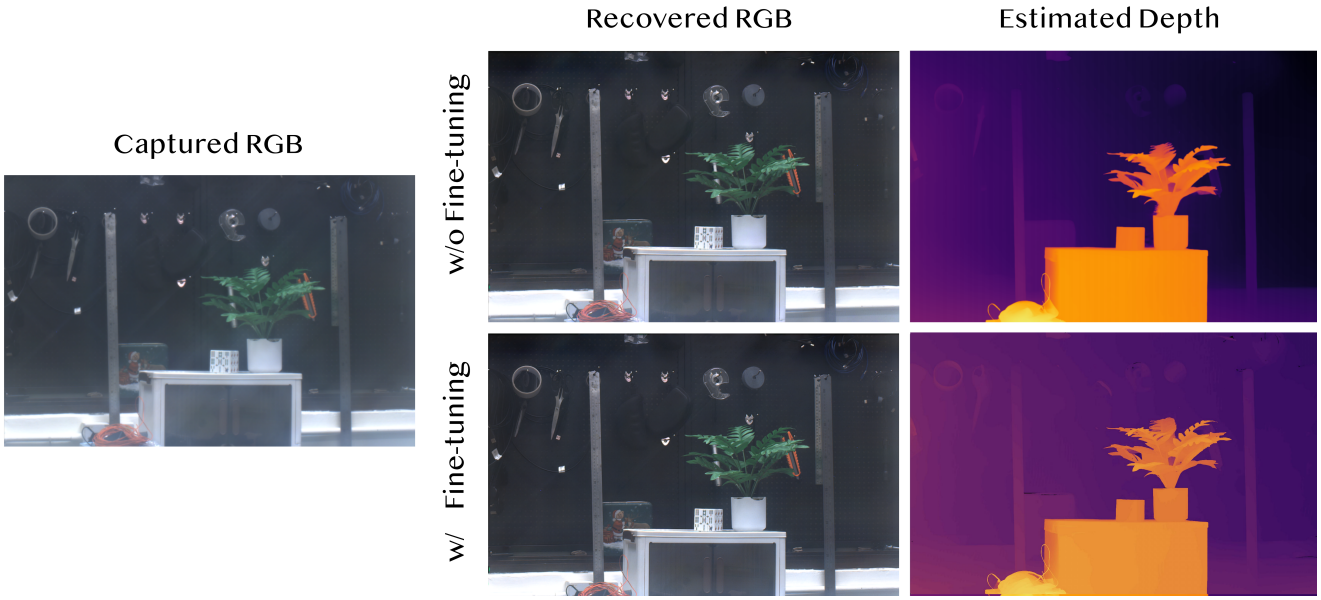

Figure 10. Comparison of RGBD images recovered using the original trained model and the model after fine-tuning for 3 epochs.

ity can result in the appearance of circular or cross-shaped scattering spots in high-energy regions, which were not observed in simulation results, necessitating careful scene selection. This highlights one of the unresolved issues with DOEs in imaging. On the other hand, our Rank-2 encoding exhibits relatively lower sensitivity to the uniformity of lighting in both the light source and the scene being captured, enabling the generation of high-quality all-in-focus and depth images with improved color reproduction.

Additionally, the last row of real-world scenes depicted in Fig. 11 showcases the outcomes captured using a 1.6 mm pinhole (f/22). We adjusted the exposure time for the image captured with the small pinhole to 990ms, which is three times longer than that of all other captured images. While this setting allows for capturing some high-frequency information, the RGB image quality still lags behind our system, and the depth estimation is prone to generating erroneous disparities. We note that a smaller pinhole diameter can provide a broader depth of field, albeit with reduced light and energy transmission through the aperture. This reduction necessitates higher exposure levels, leading to increased noise in the captured images. In real-world experiments, excessively high exposure can result in a significantly reduced frame rate during image capture, rendering video shooting and processing unfeasible. Furthermore, operating in challenging conditions such as low-light environments presents significant challenges. Under such circumstances, even increasing the exposure settings may not be sufficient to achieve accurate scene recovery.

We present additional results acquired using our Rank-2 Deep-Stereo framework in Fig. 12. By employing the mirror data augmentation method when training the model, we can obtain high-quality left and right RGBD images with snapshots, as shown in Row 1–2 in Fig. 12. Our observations indicate that our model performs reasonably well for both indoor and outdoor scenes at various distance scales. Details and edges of objects are better resolved, compared to baselines. Nevertheless, certain constraints persist due to the diffraction efficiency of the DOE.

We have noticed that the imaging quality of the DOE can be affected in indoor environments featuring strongly divergent light sources and light-reflective objects, as opposed to outdoor settings and uniform light sources. As illustrated in the upper-right corner of scenes in Fig. 12, higher-energy light sources can result in star-shaped light patterns appearing in the coded images. We leave this issue to future endeavor.

F. Supplementary Dynamic Acquisition Demo

In the provided supplementary video clip, **prototyperesults.mp4**, we also present the measurements and recovery RGBD imaging results of our learned stereo camera under dynamic acquisition configurations. The exposure is set 20 ms.

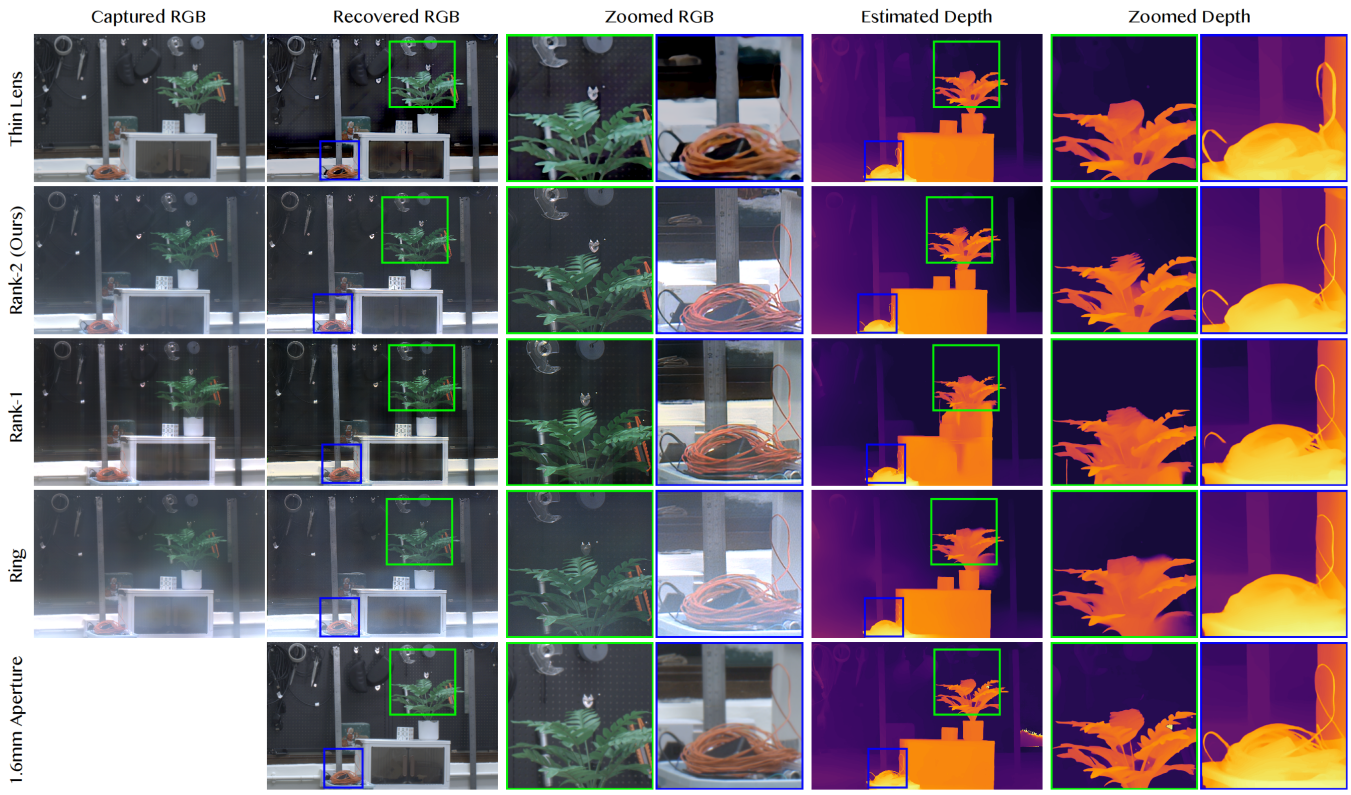

Figure 11. Sanity-check experimental results acquired using various optics, including the thin lens, our Rank-2 representation, the vanilla Rank-1 representation, the rotational symmetric representation (Ring), and a clear aperture with a small diameter size. We have captured images of the same scene under identical illuminance conditions. The first row represents the results from a traditional stereo camera, serving as our baseline model. The second row exhibits our proposed stereo camera prototype utilizing a pair of Rank-2 encoding DOEs. Rows 3 and 4 demonstrate the RGBD reconstruction results achieved using DOEs optimized with Ring and Rank-1 encoding schemes, respectively. The final row presents the results of approximate all-in-focus imaging obtained by employing a small aperture of f/22, captured with three-fold higher exposure time compared to the other setups. The scales of our estimated depth map for one scene are all the same.

References

[1] 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. 8

[2] 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. 8

[3] Joseph W Goodman. *Introduction to Fourier optics*. Roberts and Company publishers, 2005. 1

[4] 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. 3, 4

[5] 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

[6] 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. 3, 8

[7] 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, 4, 6

[8] 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. 1, 2

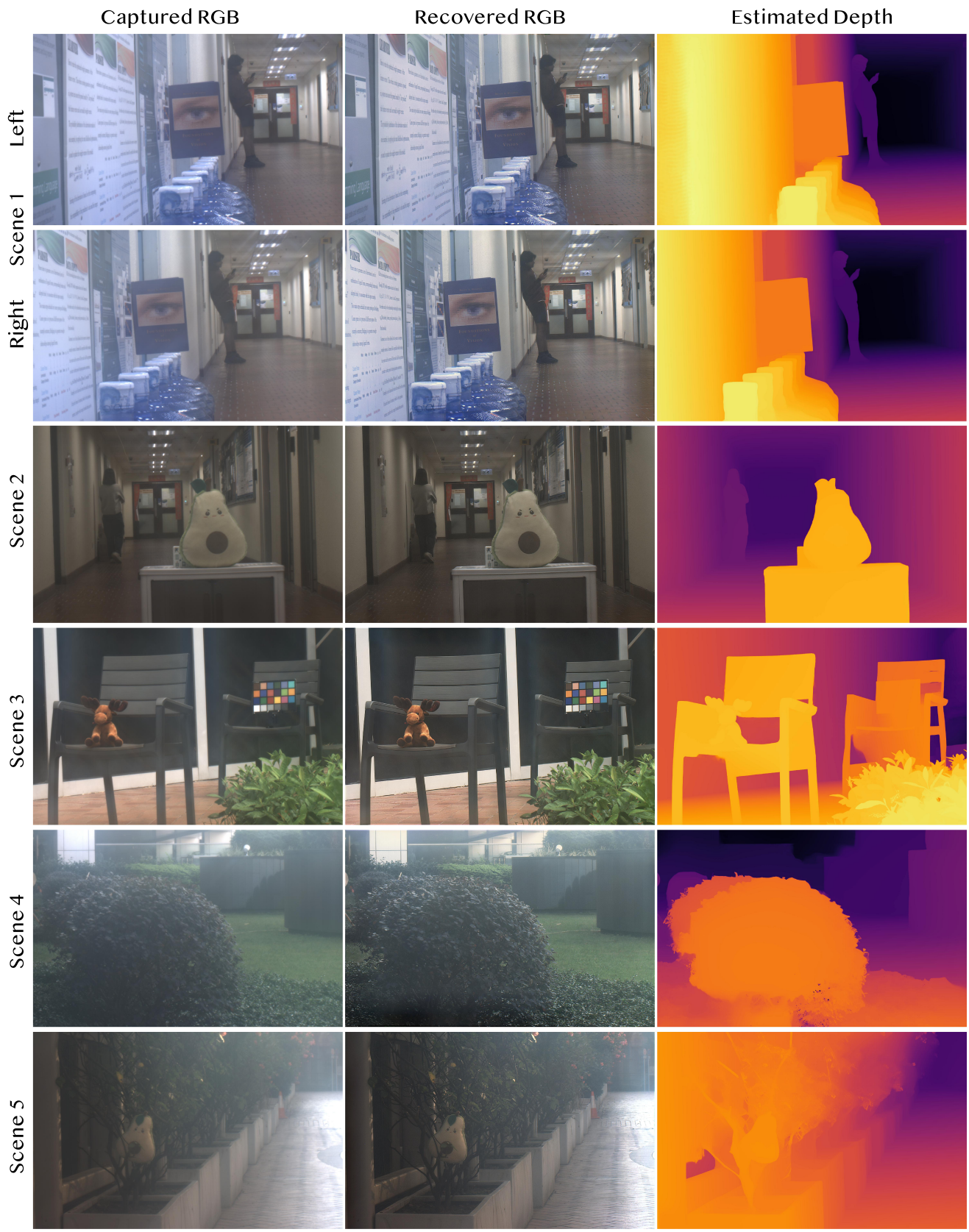

Figure 12. Additional experimental results acquired from our Rank-2 learned stereo camera. From left to right: Captured, Recovered AiF RGB, and Estimated Depth.