Depth from Defocus with Learned Optics for Imaging and Occlusion-aware Depth Estimation – Supplemental Material

Hayato Ikoma, Cindy M. Nguyen, Christopher A. Metzler, *Member, IEEE,* Yifan Peng, *Member, IEEE,* and Gordon Wetzstein, *Senior Member, IEEE*

1 Image Formation Model

1.1 Point Spread Function

An accurate model for the point spread function (PSF) of an imaging system is crucial for simulation. In our configuration, we use a conventional photographic compound lens to focus the light on the sensor, and we modify this lens by adding a diffractive optical element (DOE) into its aperture plane. This phase-coded aperture allows us to directly control the depth-dependent PSF of the imaging system. With the camera lens focusing at a distance d, the PSF is modeled as [3]

$$\overline{\text{PSF}}(x,y,z,\lambda) = \left| \frac{1}{\lambda s} \iint_{-\infty}^{\infty} \bar{D}(u,v,\lambda,z)\bar{P}(u,v,\lambda) e^{-i\frac{2\pi}{\lambda s}(ux+vy)}\, du\, dv \right|^2 \tag{1}$$

$$= \left| \frac{1}{\lambda s} \underbrace{\iint_{-\infty}^{\infty} \bar{D}(u,v,\lambda,z)\bar{P}(u,v,\lambda) e^{-i2\pi(u\tilde{x}+y\tilde{y})}\, du\, dv}_{\mathcal{F}\{\cdot\}} \right|^2, \tag{2}$$

where $\mathcal{F}\{\cdot\}$ represents a 2D-Fourier transform, $(\tilde{x},\tilde{y}) = \left(\frac{x}{\lambda s}, \frac{y}{\lambda s}\right)$, and s is the distance between the lens and the sensor which satisfies the Gaussian thin lens formula $\frac{1}{f} = \frac{1}{d} + \frac{1}{s}$, where f is the focal length of the lens. The defocus factor $\bar{D}(u,v,\lambda,z)$, which models the depth variation of the PSF and, for some point at a distance z from the lens, is given by

$$\bar{D}(u,v,\lambda,z) = \frac{z}{\lambda(u^2+v^2+z^2)} e^{i\frac{2\pi}{\lambda}\left(\sqrt{u^2+v^2+z^2}-\sqrt{u^2+v^2+d^2}\right)}. \tag{3}$$

A phase-coded mask delays the phase of the light in a per-pixel manner by an amount that is proportional to the height profile of the DOE surface $h(x,y)$ and the refractive index of the DOE material $n(\lambda)$ as

$$\phi(u,v) = \frac{2\pi}{\lambda}\left(n(\lambda) - n_{\text{air}}\right) h(u,v), \tag{4}$$

where n_{air} is the refractive index of air. In Eq. 2, this phase delay is modeled by the factor $\bar{P}(u,v,\lambda) = a(u,v)e^{i\phi(x,y)}$, where $a(u,v)$ is the transmissivity of the phase mask, which is usually 1 except when it is masked by some optical blocking filter, where it would be 0.

Following the recent proposal by Dun et al. [2], it is computationally very efficient to constrain the DOE to be radially symmetric. This approach not only reduces the number of unknown DOE surface elements by an order of magnitude, but it also allows us to reduce the dimensionality of the diffraction integral in Eq. 2 by one, which makes computing the PSF from the DOE significantly faster. To leverage the radial symmetry, we apply the Hankel transform of order zero, or the Fourier-Bessel transform [3], to derive the radially symmetric PSF PSF as

$$\text{PSF}(\rho,z,\lambda) = \left| \frac{2\pi}{\lambda s} \int_0^{\infty} r D(r,\lambda,z) P(r,\lambda) J_0(2\pi\rho r)\, dr \right|^2, \tag{5}$$

where D and P are the radially symmetric defocus and phase factors, respectively, and $\rho \coloneqq \sqrt{\tilde{x}^2+\tilde{y}^2}$ and $r \coloneqq \sqrt{u^2+v^2}$ are the radial distances on sensor and aperture, respectively. The function $J_0(\cdot)$ is the zero-th order Bessel function of the first kind.

- *H. Ikoma, C. M. Nguyen, Y. Peng, and G. Wetzstein are with the Department of Electrical Engineering, Stanford University. C. A. Metzler is with the Department of Computer Science, University of Maryland.*

- *Project website: https://www.computationalimaging.org*

Inspired by Dun et al. [2], we derive a efficient formula to evaluate the integral of Eq. 5 as

$$\begin{aligned}
\text{PSF}(\rho, z, \lambda) &= \left| \frac{2\pi}{\lambda s} \int_0^R r P(r,\lambda) D(r,\lambda,z) J_0(2\pi\rho r)\, dr \right|^2 && (6)\\
&= \left| \frac{2\pi}{\lambda s} \sum_{l=0}^{L-1} \int_{l\Delta}^{(l+1)\Delta} r P(r,\lambda) D(r,\lambda,z) J_0(2\pi\rho r)\, dr \right|^2 && (7)\\
&\approx \left| \frac{2\pi}{\lambda s} \sum_{l=0}^{L-1} P(n\Delta,\lambda) D(l\Delta,\lambda,z) \int_{l\Delta}^{(l+1)\Delta} r J_0(2\pi\rho r)\, dr \right|^2 && (8)\\
&= \left| \frac{2\pi}{\lambda s} \sum_{l=0}^{L-1} P(l\Delta,\lambda) D(l\Delta,\lambda,z) \left(\int_0^{(l+1)\Delta} r J_0(2\pi\rho r)\, dr - \int_0^{l\Delta} r J_0(2\pi\rho r)\, dr \right) \right|^2 && (9)\\
&= \begin{cases} \left| \frac{2\pi}{\lambda s} \sum_{l=0}^{L-1} P(l\Delta,\lambda) D(l\Delta,\lambda,z) \left(\frac{((l+1)\Delta)^2}{2} - \frac{(l\Delta)^2}{2} \right) \right|^2 & \text{if } \rho = 0 \\ \left| \frac{2\pi}{\lambda s} \sum_{l=0}^{L-1} P(l\Delta,\lambda) D(l\Delta,\lambda,z) \left(\frac{(l+1)\Delta}{2\pi\rho} J_1(2\pi(l+1)\Delta\rho) - \frac{l\Delta}{2\pi\rho} J_1(2\pi l\Delta\rho) \right) \right|^2 & \text{otherwise} \end{cases}, && (10)
\end{aligned}$$

where R is the radius of the aperture, L is the number of discrete radial elements on the DOE, $\Delta := R/L$, the function J_1 is the first order Bessel function of the first kind. The approximation follows $D(l\Delta, z) \approx D((l-1)\Delta, z)$ which holds for a sufficiently small DOE feature size of Δ. The fifth equality uses the properties $\int_0^a x J_0(x)\, dx = a J_1(a)$ and $J_0(0) = 1$. Note that Dun et al. [2] derived another variant of this scheme for their optical setup. To the best of our knowledge, the above formulation for a phase-coded aperture with a conventional compound lens is novel.

1.2 Comparison of Image Formation Models

As discussed in the primary text, prior work on end-to-end depth estimation used variants of a simple linear image formation model that adds the image contributions from different scene depths convolved with their respective PSF as

$$b(\lambda) = \sum_{k=0}^{K-1} \left(\text{PSF}_k(\lambda) * l_k(\lambda) \right) + \eta. \tag{11}$$

While this naive model is accurate for textured image parts in regions with locally constant depth, it fails to model the out-of-focus blur of a camera at depth discontinuities. We validate this quantitatively in Table S1 and qualitatively in Figure S1.

In Table S1, we apply several metrics, including mean squared error (MSE), mean absolute error (MAE), peak signal-to-noise ratio (PSNR), and structural similarity (SSIM), to the simulated sensor images when compared to a ray-traced ground truth image. Note that all of the approaches calculate the sensor image given only an RGB image and a per-pixel depth map. Thus, in all cases, the out-of-focus blur will only be approximated, especially at depth discontinuities where hidden parts of the scene may contribute to the blurred sensor image. These hidden scene parts are not included in either the RGB image or depth map. Due to the fact that our image formation model (see primary text for details) is computed in the linear intensity domain, we compare these image formation models both in the linear domain (lin.) and also in sRGB space. For all metrics, our image formation model, introduced in Section 3.2 of the primary text, achieves the best results.

TABLE S1
Quantitative evaluation of depth-dependent image formation models. Our nonlinear image formation model (see Sec. 3.2) achieves the best quality when compared to a ray-traced ground truth sensor image.

Model	MSE (lin.) ↓	MAE (lin.) ↓	MSE (sRGB) ↓	MAE (sRGB) ↓	PSNR (sRGB) ↑	SSIM (sRGB) ↑
Linear	0.062e-2	0.766e-2	0.557e-3	0.988e-2	32.54	0.972
Wu [4]	1.091e-2	1.054e-2	0.741e-3	1.017e-2	31.30	0.970
Chang [1]	1.088e-2	1.014e-2	0.713e-3	0.975e-2	31.47	0.973
Ours	**0.045e-2**	**0.529e-2**	**0.211e-3**	**0.683e-2**	**36.76**	**0.983**

Figure S1 also shows qualitative comparisons of these image formation models. Again, all of these are directly computed from the input RGB image and depth map (top row), so they only approximate the ground truth image generated with ray tracing (top right). As seen in the synthesized images (second row) and the corresponding error maps w.r.t the ground truth image, the biggest challenge for all of these methods is to accurately reproduce the blur at depth discontinuities. While the linear model (left column) simply ignores depth discontinuities, it actually performs reasonably well. Both Wu et al.'s [4][1] and Chang et al.'s [1] method introduce normalization factors that ensure smooth transitions between the individual layers of the depth map by unrealistically manipulating the local brightness, thereby introducing additional error. The linear model, on the other hand, ignores depth discontinuities, but it does not do any brightness manipulation such that it mostly preserves the energy, leading to a relatively low error. Our proposed occlusion-aware image formation model (right column) achieves the best and most faithful approximation of the target image.

1. Note that Wu et al.'s [4] actually report that they use the linear model in their paper, but implemented a slightly different variant of this in their code. Hence, we evaluate both the linear model and also their implemented method, labeled as Wu et al., here.

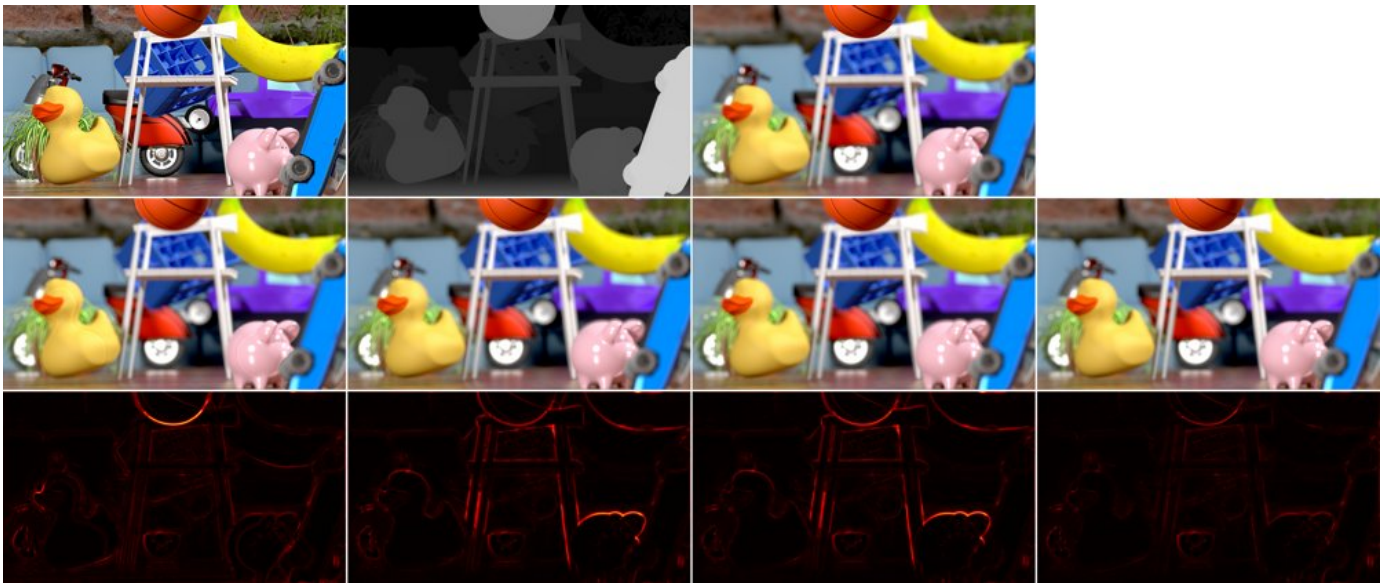

Fig. S1. Simulating accurate defocus blur (top center right) given only an RGB image (top left) and a depth map (top center left) is a challenging task, because scene information that is partly occluded is missing from the input data but does contribute to the blur. We compare a simple linear image formation model (see primary text for details), the variants of the linear model proposed by Wu et al. [4] and Chang et al. [1], and our occlusion-aware nonlinear model. As seen in the simulated sensor images (second row), corresponding error maps (third row), and close-ups (bottom row), our image formation model produces the most accurate simulation for out-of-focus blur, especially around depth discontinuities.

1.3 Approximate Inverse of Image Formation Model

As an approximate inverse for our occlusion-aware, nonlinear image formation model, we use the pseudo-inverse of the simpler linear model (Eq. 11). This is most efficiently formulated as an optimization problem of the form

$$l^{(est)} = \underset{l \in \mathbb{R}^{M \times N \times K}}{\arg\min} \left\| b - \sum_{k=0}^{K-1} \text{PSF}_k * l_k \right\|^2 + \gamma \| \, l \, \|^2 \quad \Leftrightarrow \quad \widehat{l}^{(est)} = \underset{\widehat{l} \in \mathbb{C}^{M \times N \times K}}{\arg\min} \left\| \widehat{b} - \sum_{k=0}^{K-1} \widehat{\text{PSF}}_k \circ \widehat{l}_k \right\|^2 + \gamma \| \, \widehat{l} \, \|^2, \tag{12}$$

where we make use of Parseval's identity to formulate the problem in the discrete Fourier transform (DFT) domain. Here, $l \in \mathbb{R}^{M \times N \times K}$ is our multiplane image with K layers, $\widehat{\cdot}$ denotes the DFT of a variable, $*$ is a 2D convolution and $\circ$ an element-wise multiplication. Because the measurements $b \in \mathbb{R}^{M \times N}$ have K times fewer elements as the unknowns, this is an underdetermined equation system, and we add a Tikhonov regularizer weighted by γ to account for this.

The formulation in the DFT domain makes this problem separable and allows us to solve it for each spatial frequency f_x, f_y separately using the following closed-form solution:

$$\widehat{l}^{(est)}\,[f_x, f_y, 1:K] = \underset{\widehat{l}[f_x, f_y, k] \in \mathbb{C}^K}{\arg\min} \left\| \widehat{b}\,[f_x, f_y] - \sum_{k=0}^{K-1} \widehat{\text{PSF}}\,[f_x, f_y, k]\, \widehat{l}\,[f_x, f_y, k] \right\|^2 + \gamma \| \, \widehat{l}\,[f_x, f_y] \, \|^2 \tag{13}$$

$$= \underset{\widehat{\mathbf{l}} \in \mathbb{C}^K}{\arg\min} \left\| \widehat{\mathbf{b}} - \mathbf{P}\widehat{\mathbf{l}} \right\|^2 + \gamma \| \, \widehat{\mathbf{l}} \, \|^2 \tag{14}$$

$$= \left(\mathbf{P}^H \mathbf{P} - \gamma \mathbf{I} \right)^{-1} \mathbf{P}^H \, \widehat{\mathbf{b}} \tag{15}$$

$$= \frac{1}{\gamma} \left(\mathbf{I} - \frac{1}{\gamma + \mathbf{P}\mathbf{P}^H} \mathbf{P}^T \mathbf{P} \right) \mathbf{P}^H \, \widehat{\mathbf{b}}. \tag{16}$$

Here, $\widehat{l}\,[f_x, f_y, 1:K] = \widehat{\mathbf{l}} \in \mathbb{C}^K$ is a column vector with the values of a single spatial frequency of the multiplane image across all layers $1 \ldots K$, $\mathbf{P} \in \mathbb{C}^{1 \times K}$ is a complex-valued matrix with just a single row but K columns, each corresponding to the value of the DFT of the PSF $\widehat{\text{PSF}}$ (i.e., the optical transfer function) at f_x, f_y at layers $k = 1 \ldots K$, and $\mathbf{I} \in \mathbb{R}^{K \times K}$ is the identity matrix. Going from Eq. 13 to Eq. 14 is just a change of notation, going to Eq. 15 outlines the normal equations that minimize this objective, and going to Eq. 16 applies the Woodbury formula to derive the closed-form solution that is independently computed per spatial frequency and does not require a matrix inverse to be computed. An example of the solution is visualized in Fig. S2.

REFERENCES

[1] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *International Conference on Computer Vision (ICCV)*, 2019.

[2] 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.

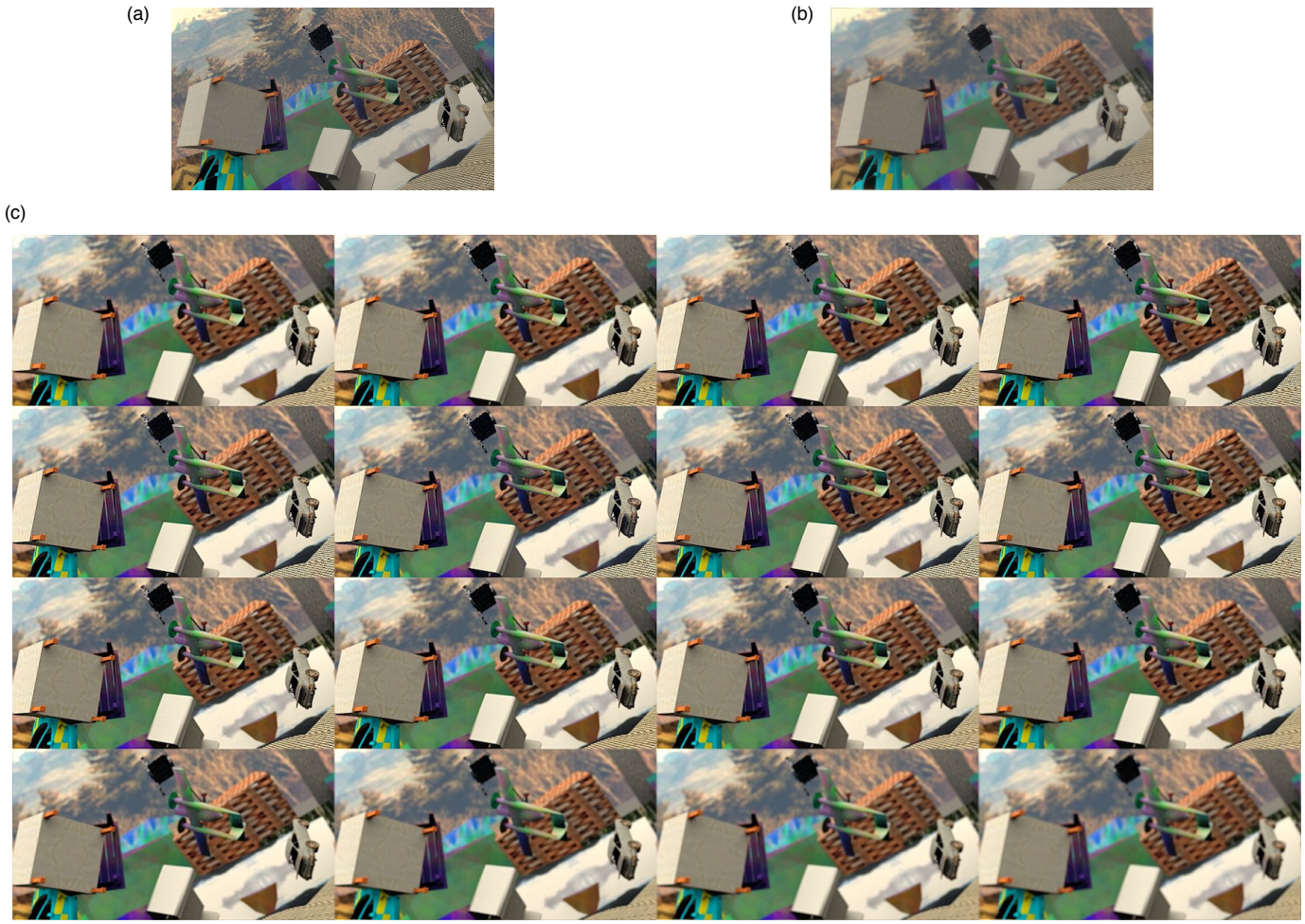

Fig. S2. (a) All-in-focus image. (b) A shallow depth-of-field image. (c) A layered representation of the least-squares solution (16). The layers are visualized sequentially from the closest plane (top-left) to the furthest plane (bottom-right).

[3] Joseph W Goodman. *Introduction to Fourier optics*. Roberts and Company Publishers, 2005.
[4] Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. Phasecam3d—learning phase masks for passive single view depth estimation. In *International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2019.

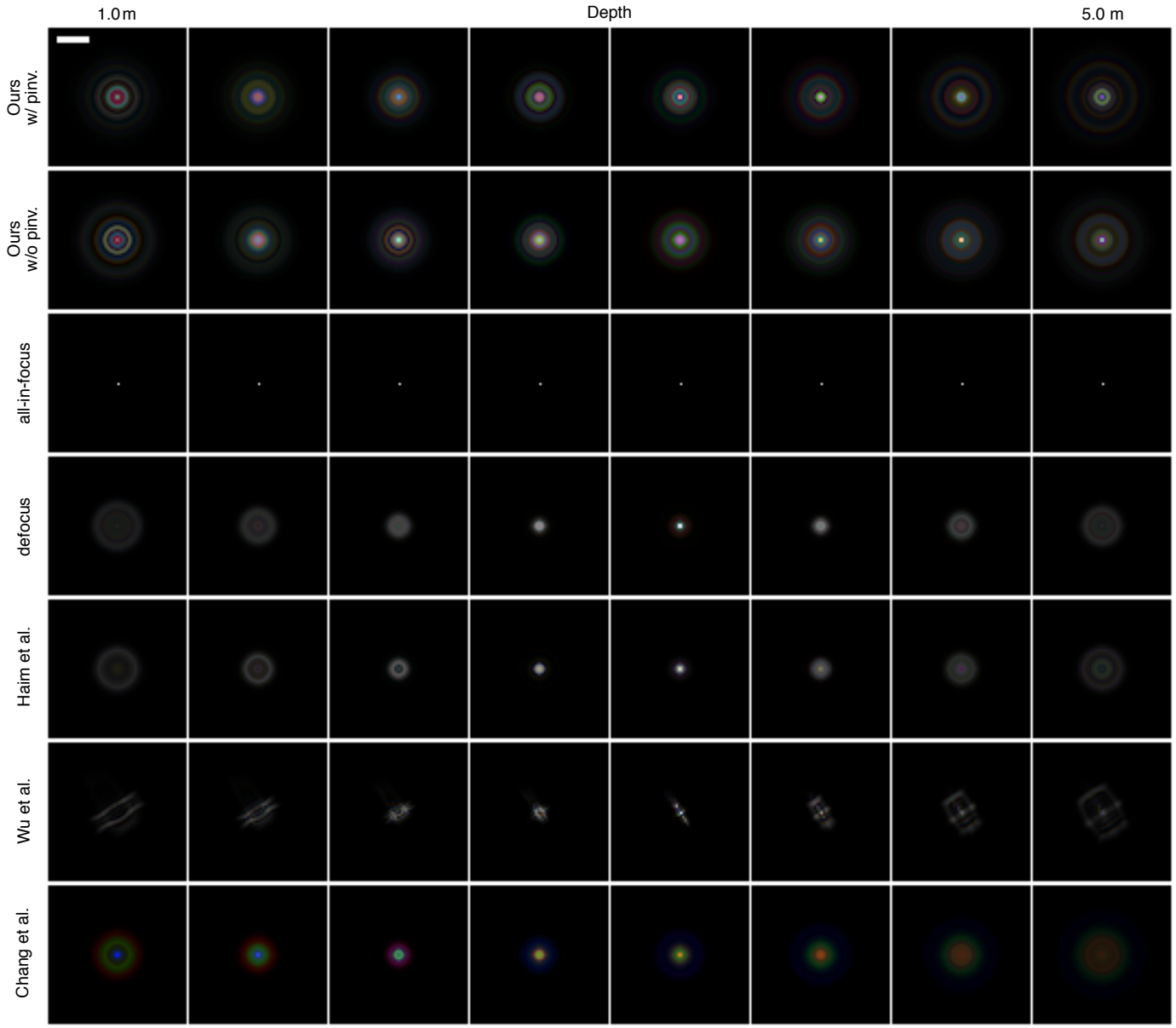

Fig. S3. Comparison of the PSFs used for the ablation study (Fig. 3 and Table 1). The scale bar is $100\,\mu\mathrm{m}$

Fig. S4. Comparison of the simulated captured images in the ablation study (Fig. 3).

Fig. S5. Top view of the toy example scene.

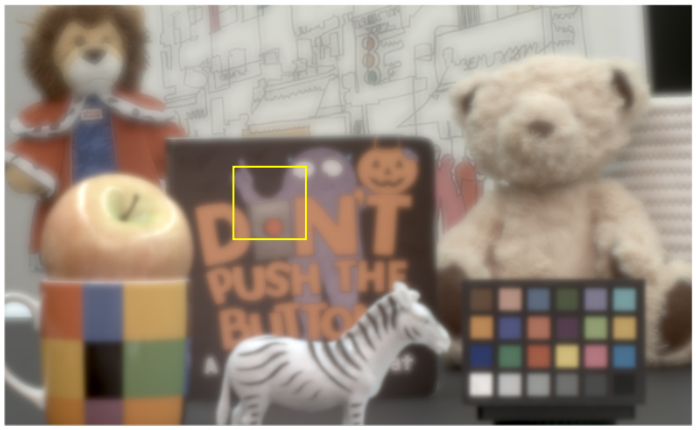

Fig. S6. Visualization of the receptive field. The receptive field of our network is 205×205. The yellow rectangle in the figure demonstrates the field-of-view size of our camera prototype.

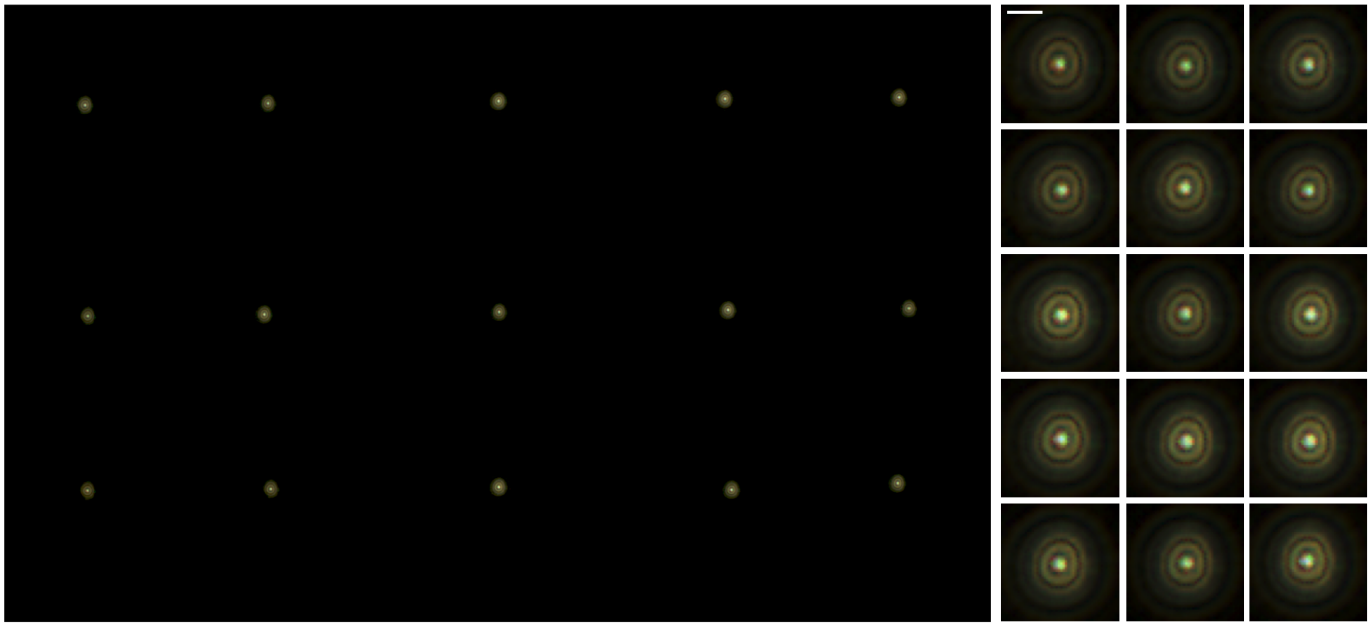

Fig. S7. Spatial variance of the PSF of our camera prototype. The PSF at the depth of $1.7\,\mathrm{m}$ is captured at different locations within a field of view. The right 15 images are the magnified view of each PSF. While all PSFs are mostly similar, some spatial variance is observed. For visualization purposes, the intensity is scaled with square root. The scale bar is $100\,\mu\mathrm{m}$.