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

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

Abstract—Monocular depth estimation remains a challenging problem, despite significant advances in neural network architectures that leverage pictorial depth cues alone. Inspired by depth from defocus and emerging point spread function engineering approaches that optimize programmable optics end-to-end with depth estimation networks, we propose a new and improved framework for depth estimation from a single RGB image using a learned phase-coded aperture. Our optimized aperture design uses rotational symmetry constraints for computational efficiency, and we jointly train the optics and the network using an occlusion-aware image formation model that provides more accurate defocus blur at depth discontinuities than previous techniques do. Using this framework and a custom prototype camera, we demonstrate state-of-the-art image and depth estimation quality among end-to-end optimized computational cameras in simulation and experiment.

Index Terms—Computational Photography, Computational Optics

1 INTRODUCTION

Robust depth perception is a challenging, yet crucial capability for many computer vision and imaging problems in robotics [1], [2], autonomous driving [3], [4], [5], [6], augmented reality [7], and 3D photography [8]. Existing approaches building on time-of-flight, stereo pairs, or structured illumination require high-powered illumination and complex hardware systems, making monocular depth estimation (MDE) from just a single 2D image [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19] one of the most attractive solutions.

MDE approaches typically rely on pictorial depth cues, such as perspective, partial occlusions, and relative object sizes learned from a dataset of training images in a supervised manner. These contextual cues reliably help estimate the relative ordering of objects within a scene [10], [16]. Defocus blur is another complementary depth cue, which has been exploited in depth from defocus (DfD) approaches [20], [21], [22], [23], [24]. Recent DfD methods also propose network architectures that learn both pictorial and depth cues simultaneously [25]. Defocus cues, however, are ambiguous, which is why many computational photography approaches use coded apertures to engineer the defocus blur to optimally encode more information than the conventional defocus blur contains [26], [27], [28], [29]. Hand-crafted aperture designs have recently been improved using an end-to-end (E2E) optimization of optics and image processing [30], [31], [32].

While existing E2E coded aperture MDE techniques have proven to work well, these methods do not take full advantage of the available monocular depth cues. Specifically, the linear optical image formation models employed by these approaches [30], [31], [32] do not model defocus blur at occlusion boundaries accurately. Thus, prior works exclusively rely on defocus information in image regions of locally constant depth. It is well known in the vision science community, however, that defocus blur and the spatial relationships implied by occluding edges provide an even stronger depth cue than pictorial cues for human vision [33], [34], [35].

To alleviate this shortcoming in DfD, we propose a nonlinear occlusion-aware optical image formation that models defocus blur at occlusion boundaries more accurately than previous E2E approaches. Moreover, we adopt a rotationally symmetric design of our optimized phase-coded aperture, reducing the computational complexity and memory requirements of the optimization by an order of magnitude. Finally, we derive an effective preconditioning approach that applies an approximate inverse of the optical image formation model to the sensor measurements. This approximate inverse makes it significantly easier for the MDE network to robustly infer a depth map from the coded sensor image. Our approach is uniquely robust in estimating not only the depth map, but also an all-in-focus image from a single optically coded sensor image, which is crucial for direct view or downstream tasks that rely on image and depth, such as classification or object detection.

Specifically, our contributions are the following:

- We formulate the E2E optimization of a phase-coded aperture and MDE network using an occlusion-aware image formation model, a rotationally symmetric aperture, and an effective preconditioning approach.
- We analyze the proposed framework and demonstrate that it outperforms standard and E2E MDE approaches with comparable network architectures.
- We build a camera prototype with a custom-fabricated diffractive optical element (DOE) in its aperture and demonstrate its performance for indoor and outdoor scenes along with high-quality RGBD video capture.

2 RELATED WORK

2.1 Monocular Depth Estimation (MDE)

Deep learning is an attractive approach for MDE as networks can identify features unknown to humans in depth estimation.
A variety of deep learning methods for MDE have been proposed using custom loss functions [10], [13], local and global constraints [16], [36], [37], and varying levels of supervision [38], [39], [40]. Geometrically-driven approaches learn surface normal estimation in conjunction with depth estimation using conditional random fields [41], two-stream CNNs [42], and 3D reconstruction from videos [43]; all showing high performance on datasets such as KITTI [44] and NYU Depth [45]. Other approaches include estimating relative depth maps [46] and using the spectral domain to augment estimation [47]. To generalize better across datasets, past works have also taken to incorporating physical camera parameters, such as defocus blur [25], [48], focal length [49], or other sensor information [50] to utilize their implicit encoding of depth cues. We propose a computational optics approach to jointly optimize a phase-coded aperture and neural network for passive 3D imaging from a single image.

2.2 Computational Imaging for Depth Estimation

Instead of relying on a single 2D image, several variants of DfD capture and process two or more images using a sum-modified-Laplacian operator [51], spatial-domain convolution transforms [52], and quadratic likelihood functions [53]. Dual-pixel sensors have also been demonstrated to capture a stereo pair with sufficient disparity to estimate depth [54]. Amplitude- [26], [27], [55] and phase-coded [56], [57] apertures have also been extensively studied as having depth estimation techniques that utilize chromatic aberrations [23]. Most of these approaches, however, use conventional lenses or hand-crafted aperture designs and algorithms, which do not optimize the system performance in an E2E fashion.

2.3 Deep Optics

Jointly designing optics or sensor electronics and networks have been explored for color filter design [58], spectral imaging [59], superresolution localization microscopy [60], superresolution single-photon imaging [61], extended depth of field [62], achromatic imaging [63], HDR imaging [64], [65], image classification [66], and video compressive sensing [67], [68]. A recent survey of the use of artificial intelligence in optics can be found in [69].

Principled approaches to jointly optimizing camera optics and depth estimation networks have also recently been proposed. For example, Haim et al. [31] use concentric rings in a phase mask to induce chromatic aberrations, while Wu et al. [30] rely on defocus cues in their jointly optimized phase mask and CNN-based reconstruction. Chang et al. [32] use E2E optimization to design a freeform lens for the task. Deep optics has also been extended to extract a depth map and multispectral scene information from a sensor measurement [70].

Inspired by the idea of deep optics, we propose a novel approach to E2E depth imaging that makes several important improvements over existing approaches in this area [30], [31], [32]. First, we introduce an occlusion-aware image formation model that significantly improves our ability to model and optically encode defocus blur at occlusion boundaries. Second, we introduce a preconditioning approach that applies an approximate inverse of our nonlinear image formation model before feeding the data into the depth estimation network. Finally, we tailor a rotationally symmetric optical design, which was recently introduced for achromatic imaging with a single DOE [63], to the application of MDE with a phase-coded aperture. Our framework enables us to recover both an RGB image and a depth map from a single coded sensor image, providing significantly higher resolution and accuracy compared to estimates from related work.

3 Phase-coded 3D Imaging System

This section describes our E2E training pipeline from the image formation model to the neural network-based reconstruction algorithm. We consider a camera with a learnable phase-coded aperture and a CNN that estimates both an all-in-focus (AiF) RGB image and a depth map from a raw sensor image with coded depth of field. This pipeline is illustrated in Fig. 1.

3.1 Radially Symmetric Point Spread Function

As in most cameras, ours is comprised of a sensor and a conventional photographic compound lens that focuses the scene on the sensor. We modify this optical system by adding a DOE into its aperture plane. This phase-coded aperture allows us to directly control the depth-dependent point spread function (PSF) of the imaging system using variations in the surface height of the DOE. The goal of the E2E optimization procedure described in this section is to find a surface profile, which shapes the PSF in a way that makes it easy and informative for the CNN to estimate per-pixel scene depth and color from a single image.

The PSF is modeled as [71]

$$\text{PSF}(\rho, z, \lambda) = \frac{2\pi}{\lambda s} \int_0^\infty r D(r, \lambda, z) P(r, \lambda) J_0(2\pi \rho r) \, dr \quad . \quad (1)$$

Here, $\rho$ and $r$ are the radial distances on the sensor and aperture planes, respectively, $\lambda$ is the wavelength, and $J_0(\cdot)$ is the zeroth order Bessel function of the first kind. In this formulation, the camera lens with focal length $f$ is focused at some distance $d$. The Gaussian thin lens formula $\frac{1}{s} = \frac{1}{d} + \frac{1}{f}$ relates these quantities to the distance between lens and sensor $s$. The defocus factor $D(r, \lambda, z)$, which models the depth variation of the PSF for a point at some distance $z$ from the lens, is given by

$$D(r, \lambda, z) = \frac{z}{\lambda(v^2 + z^2)} e^{\frac{2\pi i}{\lambda}(\sqrt{v^2 + z^2} - \sqrt{r^2 + z^2})} . \quad (2)$$

We employ a radially symmetric DOE design [63], which reduces the number of DOE parameters to be optimized, memory requirements, and compute time of the PSF by an order of magnitude compared to the requirements of a nonsymmetric design. Finally, the phase delay on the aperture plane $P$ is related to the surface profile $h$ of a DOE with refractive index $n(\lambda)$ as

$$P(r, \lambda) = a(r) e^{\frac{2\pi i}{\lambda}(n(\lambda) - n_{\text{air}}) h(r)} , \quad (3)$$

where $n_{\text{air}} \approx 1.0$ is the refractive index of air, and $a$ is the transmissivity of the phase mask, which is typically 1, but can also include light-blocking regions which set the transmissivity locally to 0.

We include a more detailed derivation of these formulations in our Supplemental Material. Although these equations are based on standard optical models [71], in the Supplement, we derive a novel formulation that allows us to evaluate the integral of Eq. 1 efficiently.
3.2 Image Formation Model with Occlusion

Prior work on E2E optimized phase-coded apertures for snapshot 3D imaging [30], [31], [32] used variants of a simple linear image formation model of the form

$$b(\lambda) = \sum_{k=0}^{K-1} \text{PSF}_k(\lambda) * l_k(\lambda) + \eta,$$

where $*$ is the 2D convolution operator, $b(\lambda)$ is a single wavelength of the sensor image, and $\eta$ is additive noise. For this model, the input RGBD image is quantized into $K$ depth layers $l_k$, with $k = 0$ being the farthest layer.

A linear model can accurately reproduce defocus blur for image regions corresponding to a locally constant depth value. However, this approach is incapable of accurately modeling defocus blur at depth discontinuities. Defocus blur at these depth edges is crucial for human depth perception [33], [35]—we argue that a MDE network would similarly benefit from more accurate defocus blur at depth edges. To this end, we adopt a nonlinear differentiable image formation model based on alpha compositing [72], [73], [74] and combine it with our wavelength- and depth-dependent PSF as

$$b(\lambda) = \sum_{k=0}^{K-1} \tilde{l}_k \prod_{k'=k+1}^{K-1} (1 - \tilde{\alpha}_{k'}) + \eta,$$

where $\tilde{l}_k := (\text{PSF}_k(\lambda) * l_k(\lambda))/E_k(\lambda)$ and $\tilde{\alpha}_{k'} := (\text{PSF}_k(\lambda) * \alpha_k(\lambda))/E_k(\lambda)$. The depth map is quantized into $K$ depth layers to compose binary masks $\alpha_k$. As the convolution with the PSFs are naively performed with the sub-images $l_k$ and the binary masks $\alpha_k$, the energy or the brightness is unrealistically reduced at the transition of depth layers. Therefore, to recover it, we apply a normalization with a factor $E_k(\lambda) := \text{PSF}_k * \sum_{k'=0}^{K-1} \alpha_{k'}$. We implement the convolutions with fast Fourier transforms (FFTs) and crop 32 pixels at the boundaries to reduce possible boundary artifacts.

As seen in Fig. 2, our nonlinear model produces a more realistic defocused image from RGBD input than previously used linear models. Compared with Wu et al. [30] and Chang et al. [32], our model’s improvements are especially noticeable around depth discontinuities, which provide the downstream network superior defocus information. Compared to the direct linear model, our model produces more accurate defocus blur around texture and depth edges. The error maps shown in Fig. 2 are computed with respect to the ray-traced ground truth sensor image. Note that ray tracing is a valuable tool for verifying these different images formation models, but it is not a feasible tool for training our system. It takes too long to ray trace images on the fly during training, and it is infeasible to pre-compute every ray-traced image for every possible phase-coded aperture setting. Please refer to the Supplemental Material for additional discussions.

3.3 CNN-based Estimation of Image and Depth

In the E2E training, we utilize a CNN to jointly estimate an all-in-focus image and a depth map or an RGBD image. We describe its architecture and training details in the following.

3.3.1 Preconditioning with Approximate Inverse

Although the linear image formation model outlined in Eq. 4 is not accurate at occlusion boundaries, it provides a simple-enough framework to serve as a preconditioning step for our network. Specifically, we formulate the inverse problem of finding a multiplane representation $l^{(est)} \in \mathbb{R}^{M \times N \times K}$ from a single
2D sensor image as a Tikhonov-regularized least squares problem with regularization parameter \( \gamma \)

\[
l^{(\text{est})} = \arg\min_{l \in \mathbb{R}^{M \times N \times K}} \left\| b - \sum_{k=0}^{K-1} \text{PSF}_k \ast l_k \right\|^2 + \gamma \left\| l \right\|^2 .
\] (6)

We omit the wavelength dependence for notational simplicity here. In our Supplemental Material, we derive a closed-form solution for this inverse problem in the frequency domain. It is implemented with FFTs, and edge-tapering is applied as a pre-processing step to reduce ringing artifacts [75]. This closed-form inverse of the linear image formation model maps the 2D sensor image into a layered 3D representation that has the sharpest details on the layer corresponding to the ground truth depth, even though it is incorrect at depth edges. Thus, in simplified terms, our CNN then has to find the layer with the sharpest details or highest gradients at each pixel. This pixel value is close to the sought after RGB layer corresponding to the ground truth depth, even though it is linear image formation model maps the 2D sensor image intowith FFTs, and edge-tapering is applied as a pre-processing step to reduce ringing artifacts [75]. This closed-form inverse of the linear image formation model maps the 2D sensor image into a layered 3D representation that has the sharpest details on the layer corresponding to the ground truth depth, even though it is incorrect at depth edges. Thus, in simplified terms, our CNN then has to find the layer with the sharpest details or highest gradients at each pixel. This pixel value is close to the sought after RGB layer corresponding to the ground truth depth, even though it is linear image formation model maps the 2D sensor image into

The E2E model was trained for 100 epochs with the Adam optimizer \((\beta_1 = 0.9, \beta_2 = 0.999)\) with a batch size of 3 and evaluated on the validation set at the end of every epoch. Among the 100 checkpoints, the one achieving the lowest validation loss is used for evaluating on the test set. Source code and pre-trained network models and phase masks are available on the project website: https://www.computationalimaging.org/publications/deepopticsdfd.

4 Analysis and Evaluation

In this section, we describe a number of qualitative and quantitative experiments we performed to evaluate our method and compare it to related work.

4.1 Datasets

For our simulated results, we use the cleanpass subset of the FlyingThings3D dataset for training [79], [80]. This dataset contains 22K and 8K pairs of an RGB image and corresponding depth maps for training and testing, respectively. The training set is divided into 18K and 4K pairs for training and validation, respectively. During training, we performed random cropping with window sizes of 384 x 384 pixels and random horizontal/vertical flipping to augment the training set. The target depth range was set to 0.1 m to 5.0 m, and the camera is focused at 1.7 m with an f-number of 6.3. When the depth map was converted to an alpha channel volume, it was resampled with the inverse perspective sampling scheme [54].

4.2 Baseline Comparisons

We compare our method to several alternative approaches:

- **AiF**: Applying the same depth estimation CNN we use in our model directly to a ground truth (GT) all-in-focus (AiF) image.
- **DFF**: Applying the same depth estimation CNN we use to a sensor image with a conventional (non-learned) defocus blur with a similar f-number as our setting.
- **Haim et al. [31]**: A three-ring phase-coded aperture design implemented with our radially symmetric PSF model. The step function representing the rings is implemented with \( \tanh(100\rho) \) as proposed in that work.
An ablation study and comparison to previous work in simulation. Top: all methods are implemented as described in their respective papers and use their respective sensor image as input. The output of each network is compared to the ground truth depth map, and we additionally compare either the estimated RGB image or, if an algorithm does not directly compute that, the sensor image to the all-in-focus reference image. Bottom: an ablation of different variants of the proposed rotationally symmetric DOE design for the linear image formation, a linear image formation with nonlinear refinement [30], and the proposed nonlinear model. Using a variety of different metrics on estimated RGB images and depth maps, we demonstrate that the proposed approach is the best when using a comparable CNN architecture for all methods.

### Table 1

<table>
<thead>
<tr>
<th>Model</th>
<th>Refinement</th>
<th>Image</th>
<th>MAE↓</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
<th>MAE↓</th>
<th>RMSE↓</th>
<th>log_{10}δ↓</th>
<th>δ&lt;1.25↑</th>
<th>δ&lt;1.25^2↑</th>
<th>δ&lt;1.25^3↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>All in focus (AiF)</td>
<td>—</td>
<td>GT</td>
<td>0.357</td>
<td>0.500</td>
<td>0.099</td>
<td>0.658</td>
<td>0.807</td>
<td>0.874</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depth from def. (DfD)</td>
<td>—</td>
<td>GT</td>
<td>0.357</td>
<td>0.500</td>
<td>0.099</td>
<td>0.658</td>
<td>0.807</td>
<td>0.874</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haim et al. [31]</td>
<td>—</td>
<td>GT</td>
<td>0.357</td>
<td>0.500</td>
<td>0.099</td>
<td>0.658</td>
<td>0.807</td>
<td>0.874</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wu et al. [30]</td>
<td>—</td>
<td>3.24e-2</td>
<td>7.111</td>
<td>0.097</td>
<td>0.228</td>
<td>0.039</td>
<td>0.692</td>
<td>0.692</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chang et al. [32]</td>
<td>—</td>
<td>3.28e-2</td>
<td>0.708</td>
<td>0.297</td>
<td>0.635</td>
<td>0.109</td>
<td>0.803</td>
<td>0.803</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear w/o pinv</td>
<td>—</td>
<td>3.49e-2</td>
<td>0.704</td>
<td>0.207</td>
<td>0.521</td>
<td>0.090</td>
<td>0.865</td>
<td>0.865</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear w/ pinv</td>
<td>—</td>
<td>3.62e-2</td>
<td>0.694</td>
<td>0.205</td>
<td>0.490</td>
<td>0.077</td>
<td>0.888</td>
<td>0.888</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonlin. w/o pinv</td>
<td>Nonlin. w/o pinv</td>
<td>3.20e-2</td>
<td>0.870</td>
<td>0.268</td>
<td>0.598</td>
<td>0.108</td>
<td>0.845</td>
<td>0.845</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonlin. w/ pinv</td>
<td>Nonlin. w/ pinv</td>
<td>3.19e-2</td>
<td>0.891</td>
<td>0.258</td>
<td>0.554</td>
<td>0.103</td>
<td>0.856</td>
<td>0.856</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rot. Symmetric</td>
<td>—</td>
<td>3.18e-2</td>
<td>0.900</td>
<td>0.127</td>
<td>0.264</td>
<td>0.065</td>
<td>0.901</td>
<td>0.901</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>3.17e-2</td>
<td>0.902</td>
<td>0.095</td>
<td>0.203</td>
<td>0.038</td>
<td>0.931</td>
<td>0.931</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>3.16e-2</td>
<td>0.903</td>
<td>0.104</td>
<td>0.237</td>
<td>0.041</td>
<td>0.925</td>
<td>0.925</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>3.15e-2</td>
<td>0.905</td>
<td>0.089</td>
<td>0.191</td>
<td>0.034</td>
<td>0.941</td>
<td>0.941</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Wu et al. [30]:** The PhaseCam3D approach implemented with a DOE size of $256 \times 256$ features and 55 Zernike coefficients. The DOE was initialized with the DOE which minimizes the mean of the Cramér-Rao lower bound for single-emitter localization as described in that work.
- **Chang et al. [32]:** A singlet lens introducing chromatic aberrations implemented with our radially symmetric PSF model. All optical parameters match our setup.

Related works are typically trained using only $L_{Depth}$ with regularization of depth maps and the PSFs. For a fair comparison of optical models, we used only $L_{Depth}$ for the respective works and the baselines. We reimplemented their image formation model by following their respective papers or the code provided by the authors.

### 4.3 Comparisons to Prior Work

Fig. 3 shows qualitative and quantitative results for one example from our test dataset. The ground truth RGB image and depth maps are shown on the left, followed by all baselines described above. Due to the fact that none of the baselines attempt to estimate an AiF RGB image, we evaluate their RGB image quality on the captured sensor image. Unsurprisingly, our estimated RGB image is significantly better than all of these sensor images when compared to the reference AiF image. When comparing the quality of the estimated depth maps, the conventional DfD approach does surprisingly well, much better than any of the optimized methods. This is likely due to the fact that all of these approaches use variants of the linear image formation model, which provide inaccurate defocus blur around depth discontinuities, whereas our implementation of DfD is trained with the nonlinear image formation model that all methods are tested against. Nevertheless, our approach outperforms all of these baselines when implemented with the proposed preconditioning using the approximate inverse ($pinv$). Without the preconditioning, our approach does slightly worse on the depth map than the DfD approach, which is understandable because our approach needs to recover both depth map and RGB image whereas DfD only estimates the depth map with the same CNN architecture. These trends are confirmed by the quantitative results shown in Table 1 (top).

### 4.4 Additional Ablations

We also ablate the proposed rotationally symmetric DOE design in more detail in Table 1 (bottom) by analyzing the importance
of the nonlinear image formation model over the linear one with optional nonlinear refinement, as proposed by Wu et al. [30]. For all of these variants of our DOE design, using the \( \psi \text{pinv} \) improves both image and depth quality compared to the results not using the \( \psi \text{pinv} \). Moreover, the nonlinear model also performs better than linear variants.

In Table 2, we evaluate the effect of the relative weights of image and depth terms of the loss function (Eq. 8). As expected, the relative weights between the two loss terms directly trade RGB image quality and depth quality. For simulating the PSF, however, we upsample these 400 features to 4,000 pixels using nearest-neighbor upsampling to reasonably well. For simulating the PSF, however, we upsample these 400 features to 4,000 pixels using nearest-neighbor upsampling to ensure the accuracy.

To evaluate the impact of the limited DE of a physical DOE, we performed additional simulations analyzing the performance of various combinations of diffraction efficiencies for training and testing (Tab. 3). Unsurprisingly, optimizing the correct DE is always best, with mismatches degrading performance. Reducing the DE also decreases the overall performance.

5.1.2 PSF Model with Limited Diffraction Efficiency

As often observed in practice, our fabricated DOEs have an imperfect diffraction efficiency (DE), which means that some amount of the incident light passes straight through them without being diffracted. In this scenario, the measured PSF of the imaging system comprises a superposition of the native PSF of the focusing lens and the designed PSF created by the phase-coded aperture. With a DE of \( \mu \), we model the resulting PSF as

\[
\text{PSF} = \mu \cdot \text{PSF}_{\text{design}} + (1 - \mu) \cdot \text{PSF}_{\text{native}}.
\]

To quantify our DE, we fabricated a diffraction grating and determined that the DE of our fabrication process is \( \sim 70 \% \). With this DE, the DOE and the network were jointly optimized for our physical prototype.

We parameterized the DOE height using 400 learnable parameters which matches the accuracy of our fabrication technique reasonably well. For simulating the PSF, however, we upsample these 400 features to 4,000 pixels using nearest-neighbor upsampling to ensure the accuracy.

To evaluate the impact of the limited DE of a physical DOE, we performed additional simulations analyzing the performance of various combinations of diffraction efficiencies for training and testing (Tab. 3). Unsurprisingly, optimizing the correct DE is always best, with mismatches degrading performance. Reducing the DE also decreases the overall performance.
1.0 m 5.0 m Depth

Fig. 5. Depth-dependent point spread functions (PSFs). The designed PSF (top row) is optimized with our end-to-end simulator. Optical imperfections result in the captured PSF (center row), slightly deviating from the design. Instead of working directly with the captured PSF, we fit a parametric model to it (bottom row), which is then used to refine our CNN. The scale bar represents 100 µm. For visualization purposes, we convert the linear intensity of the PSF to amplitude by applying a square root.

5.1.4 Training Details

The camera settings, optimizer, and loss function are the same as in the ablation study except for the change of the weighting for loss functions. We set $\psi_{\text{RGB}} = \psi_{\text{Depth}} = \psi_{\text{PSF}} = 1$.

5.1.5 Fabrication and Hardware Implementation

The trained DOE is fabricated using the imprint lithography technique. For this purpose, the designed phase mask is patterned on a positive photosensitive layer (AZ-1512, MicroChemicals) with a grayscale lithography machine (MicroWriter ML3, Durham Magneto Optics), and its 3D structure is then replicated onto a UV-curable optical adhesive layer (NOA61, Norland Products) on a glass substrate. The glass substrate is also coated with a chromium-gold-chromium layer to block the incoming light around the DOE. Additional details on this fabrication procedure are described in [63].

The glass substrate with the DOE is mounted in the aperture plane of a compound lens (Yongnuo, 50 mm, f/1.8) with a custom 3D-printed holder. To reduce multiple reflections inside the lens, a black nylon sheet is also inserted between the DOE and the lens. The DOE has a diameter of 5.6 mm which corresponds to f/6.3 for the compound lens. The lens is mounted on a machine vision camera (FLIR Grasshopper3), and images are captured in 16-bit raw mode. The fabricated DOE and our mounting system are shown in Fig. 4. Since we manually align the DOE and the light-blocking annulus (Fig. 4, b), these two are not perfectly aligned, partly contributing to the undiffracted light. Specifically, we measured a misalignment of $\sim$140 µm between these two components.
5.2 Model Refinement with PSF calibration

After fabricating and mounting the DOE in our camera, we record depth-dependent PSFs of this system by capturing a white LED with a 15 µm pinhole at multiple depths. For each depth, ten camera images are averaged to reduce capture noise, and the averaged image is demosaiced with bilinear interpolation. As shown in Fig. 5 (center row), the captured PSF is slightly different from the designed one (top row). This difference originates from various factors, including optical aberrations, misalignment of the DOE inside the compound lens, and fabrication errors. To accommodate for this difference with our RGB and depth estimation CNN, a PSF model is fitted with the MSE loss to the captured PSF by optimizing a rotationally symmetric height map and the diffraction efficiency in post-processing. With the fitted PSF (Fig. 5, bottom row), we refine our CNN with the same training procedure described before but with a fixed PSF for inference with captured images.

To optimize the robustness of our method during inference, we feed a set of horizontally and vertically flipped sensor images into our pre-trained network and take the average of their outputs as the final estimation. This inference-time augmentation is possible due to the rotational symmetry of the PSF.

5.3 Experimental Results

We show experimentally captured results in Fig. 6 and in the supplemental movies. These examples include scenes captured in both indoor and outdoor settings. The sensor images captured with our phase-coded aperture camera prototype (column 3) look more blurry than those of a conventional camera image of the same scenes (column 1). Notably, this depth-dependent blur encodes the optimized information that is used by our pre-trained and refined CNN to estimate all-in-focus images (column 4) and depth maps (column 5). The image quality of our estimated RGB images is very good and comparable to the reference images. Our depth maps show accurately estimated scene depth with fine details, especially in challenging areas like the plants in the bottom rows and the toys in the top row. Compared to depth maps estimated from the conventional camera images with a CNN architecture similar to that used by our approach (column 2), our depth maps are significantly more detailed. They can easily segment high-frequency objects apart, and they show an overall higher quality than this baseline does.

In Fig. 7 and the supplemental movies, we compare our estimated RGBD images against a baseline model trained on AiF images and a state-of-the-art MDE method (MiDaS) [19]. For MiDaS, we used the code with a trained checkpoint provided by the authors (v2.1). While MiDaS estimates a qualitatively good depth map, their estimation remains relative and is not consistent between different frames. On the other hand, our method estimates accurate depth in a temporally consistent manner.

Finally, we show experiments that help quantify the depth accuracy achieved by our prototype in Fig. 8. In this experiment, we capture five photographs of a scene where one object, i.e., the book, is moved to different distances of known values. We extract a region of interest (ROI) of size 50 × 50 pixels in each of the estimated depth maps and report the estimated depth as the mean value of the ROI. The estimated depth values (shown in the labels of the individual depth maps) are in good agreement with the calibrated ground truth distances with a total root mean square error of 0.17 m for all five depth planes.

6 Discussion

In summary, we present a new approach to jointly optimize a phase-coded aperture implemented with a single DOE and a CNN that estimates both an all-in-focus image and a depth map of a scene. Our approach is unique in leveraging a nonlinear image formation model that more accurately represents the defocus blur observed at depth discontinuities than previous approaches do. Our model also leverages a rotationally symmetric DOE/PSF design, which makes the training stage computationally tractable by reducing both memory consumption and the number of optimization variables by an order of magnitude compared to those of previous works. Although our nonlinear image formation model is marginally more computationally expensive than the linear model during training time, it is not part of the test/inference time where this operation is performed physically by the optics.

We note that other parameterizations of the DOE could also provide computational benefits. For example, similar to Sitzmann et al. [62] and Wu et al. [30], we could use a Zernike representation of the DOE that matches the small number of parameters of our rotationally symmetric model. Although these two options would have the same number of parameters to optimize, the Zernike representation would be smooth and still require an order of magnitude higher memory resources, which is the primary problem the rotationally symmetric model solves. The latter requires exclusively 1D computations to evaluate the whole rotationally symmetric 2D PSF. For the Zernike representations, all of these calculations need to be done in 2D at full resolution. Because we use an E2E-differentiable model, the huge amount of intermediate variables that need to be stored in the computational graph for these 2D calculations make a Zernike-based option as memory intensive as other options.

6.1 Limitations and Future Work

One of the primary limitations of our phase-coded aperture includes the limited diffraction efficiency as well as some amount of shift variance of the measured PSFs (see the Supplemental.
Material). In this project, we were able to successfully work around these issues by optimizing a DOE, taking the limited diffraction efficiency into account, and by randomly jittering the PSF during training, making it robust to slight shifts. Yet, the performance of similar types of computational imaging systems could be greatly improved by optimizing the fabrication processes and diffraction efficiency of the optical elements as well as the alignment and calibration of the fully integrated systems.

In our captured results (Fig. 6), we also see some edges of textured regions appearing in the estimated depth maps. These remaining imperfections could be introduced by any difference between our image formation model and the physical system, including a small amount of spatial variation of the PSF, optical aberrations, or a slight mismatch of the estimated and true diffraction efficiency of the DOE. Moreover, we only simulate the PSF at three discrete wavelengths, to keep memory usage reasonable, whereas the physical sensor integrates over a reasonably broad spectrum. Finally, we discretize the depth of the scene into layers whereas the physical model is continuous. We account for some of these issues by jittering the PSF during the training, but not all of these physical effects can be perfectly modeled. Thus, although our approach shows significant improvements over related methods, there is further room for improving experimental results.

Network architectures and training procedures for MDE have greatly improved in performance at the cost of increased complexity (e.g., [18], [19]). These software-only approaches are very successful in estimating relative depth information of a scene, but they are unable to reliably estimate absolute scene depth. Depth-from-learned-defocus-type approaches have the ability to estimate robust absolute scene depth in regions where texture and depth edges are available, but our work and previous approaches in this area use relatively small networks that lack the capacity of modern monocular depth estimators and thus may not be able to learn contextual cues as effectively as those methods do. Therefore, it is important to explore different network architectures that are optimized to capture both the physical information provided by (coded) defocus blur as well as the contextual cues encoded by the pictorial scene information. Finally, treating the image and depth reconstruction tasks with separate networks could further improve the network capacity, but at the cost of increased memory consumption.

6.2 Conclusion

The emerging paradigm of E2E optimization of optics and image processing has great potential in various computational optics applications. We believe that depth-dependent PSF engineering in particular, for example to passively estimate the depth of a scene, is among the most promising directions of this paradigm with potential impacts on robotics, autonomous driving, human-computer interaction, and beyond. With our work, we make significant progress towards making jointly optimized hardware-software systems practical in these applications.

ACKNOWLEDGMENTS

C.M.N. was supported by an NSF Graduate Research Fellowship under award DGE-1656518. C.A.M. was supported by an appointment to the Intelligence Community Postdoctoral Research Fellowship Program at Stanford University administered by Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy and the Office of the Director of National Intelligence (ODN). G.W. was further supported by NSF awards 1553333 and 1839974, a Sloan Fellowship, and a PECASE by the ARO. Part of this work was performed at the Stanford Nano Shared Facilities (SNSF), supported by the National Science Foundation under award ECCS-2026822. We would like to thank the following Blend Swap users for models used in our Blender rendering: pujiyanto (wooden chair), wawanbreton (blue sofa), bogiva (kettle), danikreuter (vespa), animatedheaven (basketball), tikiteyboo (bottle crate), TowerCG (spider plant), oenvoyage (piggy bank), JSantel (banana), Rohit Miraje (purple jeep), mStuff (rubber duck), and sudeepsingh (blue car).


Hayato Ikoma is a Ph.D. student in Electrical Engineering at Stanford University (USA). He received a B.E. at University of Tokyo (Japan), and M.S. degrees at Kyoto University (Japan), Massachusetts Institute of Technology (USA), and École Normale Supérieure de Cachan (France). His research focuses on the development of computational imaging techniques for cameras and fluorescence optical microscopy.

Cindy M. Nguyen received her B.S. in Bioengineering at Stanford University, Stanford, CA, USA in 2019. She is currently a Ph.D. student in Electrical Engineering at Stanford University. Stanford, CA, USA. Her interests lie in applying optimization methods to problems in computer vision and computational imaging. She is a recipient of the NSF Graduate Research Fellowship.
Christopher A. Metzler (Member, IEEE) is an Assistant Professor of Computer Science (and Electrical and Computer Engineering by courtesy) at the University of Maryland, College Park. He received his B.S., M.S., and Ph.D. degrees in Electrical and Computer Engineering from Rice University, Houston, TX, USA in 2013, 2014, and 2019, respectively, and recently completed a two-year postdoc in the Stanford Computational Imaging Lab. He was an Intelligence Community Postdoctoral Research Fellow, an NSF Graduate Research Fellow, a DoD NDSEG Fellow, and a NASA Texas Space Grant Consortium Fellow. His research uses machine learning and statistical signal processing to develop data-driven solutions to challenging imaging problems.

Yifan (Evan) Peng (Member, IEEE) received a Ph.D. in Computer Science from the University of British Columbia, Canada, in 2018, and an M.Sc. and B.S., both in Optical Science & Engineering, from Zhejiang University, China, in 2013 and 2010, respectively. He is currently a Postdoctoral Fellow in the Stanford Electrical Engineering Department. His research focuses on incorporating optical and computational techniques for enabling new imaging modalities. He is working on computational cameras & displays with wave optics.

Gordon Wetzstein (Senior Member, IEEE) received the graduation (with Hons.) degree from the Bauhaus-Universität Weimar, Weimar, Germany and the Ph.D. degree in computer science from the University of British Columbia, BC, Canada, in 2011. He is currently an Assistant Professor of Electrical Engineering and, by courtesy, of Computer Science, with Stanford University, Stanford, CA, USA. He is the Leader of Stanford Computational Imaging Lab and a Faculty Co-Director of the Stanford Center for Image Systems Engineering. At the intersection of computer graphics and vision, computational optics, and applied vision science, his research has a wide range of applications in next-generation imaging, display, wearable computing, and microscopy systems. He is the recipient of an NSF CAREER Award, an Alfred P. Sloan Fellowship, an ACM SIGGRAPH Significant New Researcher Award, a Presidential Early Career Award for Scientists and Engineers (PECASE), an SPIE Early Career Achievement Award, a Terman Fellowship, an Okawa Research Grant, the Electronic Imaging Scientist of the Year 2017 Award, an Alain Fournier Ph.D. Dissertation Award, Laval Virtual Award, and the Best Paper and Demo Awards at ICCP 2011, 2014, and 2016 and at ICIP 2016.