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 optically 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.

- *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*

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]

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

Here, ρ and r are the radial distances on the sensor and aperture planes, respectively, λ is the wavelength, and $J_0(\cdot)$ is the zero-th 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 $\frac{1}{f} = \frac{1}{d} + \frac{1}{s}$ 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

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

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

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

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.

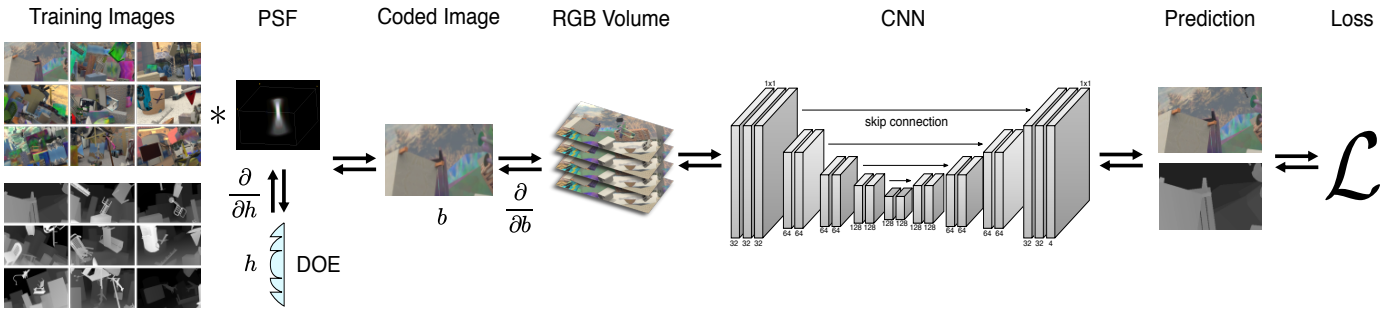


Fig. 1. Illustration of E2E optimization framework. RGBD images of a training set are convolved with the depth-dependent 3D PSF created by a lens surface profile h and combined using alpha compositing. The resulting sensor image b is processed by an approximate-inverse-based preconditioner before being fed into the CNN. A loss function $\mathcal{L}$ is applied to both the resulting RGB image and the depth map. The error is backpropagated into the CNN parameters and the surface profile of the phase-coded aperture.

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

$$b\left(\lambda\right) = \sum_{k=0}^{K-1} \mathrm{PSF}_k\left(\lambda\right) * l_k\left(\lambda\right) + \eta, \tag{4}$$

where $*$ is the 2D convolution operator, $b\left(\lambda\right)$ is a single wavelength of the sensor image, and η 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

$$b\left(\lambda\right) = \sum_{k=0}^{K-1} \widetilde{l}_k \prod_{k'=k+1}^{K-1} \left(1 - \widetilde{\alpha}_{k'}\right) + \eta, \tag{5}$$

where $\widetilde{l}_k := \left(\mathrm{PSF}_k\left(\lambda\right) * l_k\right) / E_k(\lambda)$ and $\widetilde{\alpha}_k := \left(\mathrm{PSF}_k\left(\lambda\right) * \alpha_k\left(\lambda\right)\right) / E_k(\lambda)$. The depth map is quantized into K depth layers to compose binary masks α_k. As the convolution with the PSFs are naively performed with the sub-images l_k and the binary masks α_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 $E_k(\lambda) := \mathrm{PSF}_k * \sum_{k'=0}^{k} \alpha_{k'}$. 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

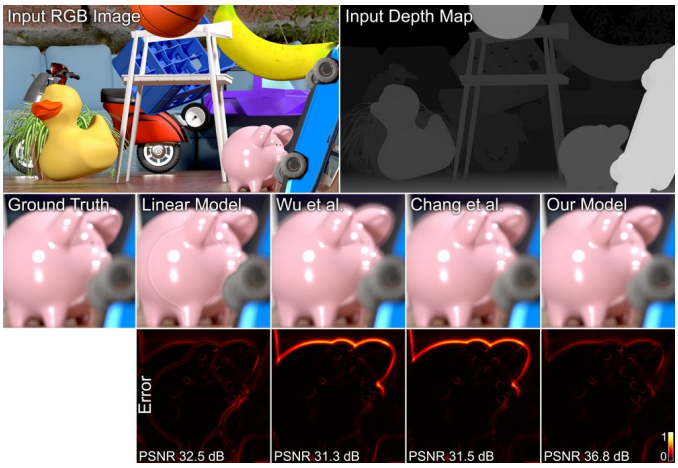


Fig. 2. Comparing image formation models that simulate defocus blur from an RGB image (top left) and a depth map (top right). Existing linear models, including Wu et al.'s [30] and Chang et al.'s [32] variants of it, do not model blur at depth discontinuities adequately. Our nonlinear occlusion-aware model achieves a more faithful approximation of a ray-traced ground truth image.

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 γ

$$l^{(est)} = \underset{l \in \mathbb{R}^{M \times N \times K}}{\arg\min} \left\| b - \sum_{k=0}^{K-1} \text{PSF}_k * l_k \right\|^2 + \gamma \| l \|^2. \quad (6)$$

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 value, and the corresponding layer index is representative of its depth. It is therefore intuitive that the CNN will have an easier time learning the mapping from a layered depth representation to an RGB image and depth map, rather than having to learn the "full" inverse starting from the sensor image. These arguments are further discussed and experimentally validated in Section 4. The closed-form solution is fully differentiable, and its computational cost is dominated by the Fourier transform, so it is $O(N^2 \log N)$ for an image with N^2 pixels.

3.3.2 CNN Architecture

We use the well-known U-Net style network architecture [76] to estimate the 4-channel RGBD image (Fig. 1). The input is the channel-wise concatenation of the captured image and the multilayer deconvolved image (Eq. 6) and is transformed to a 32-channel feature map with a 1×1 convolution. Our CNN has skip connections and five scales with four consecutive downsamplings and upsamplings. Each scale has two consecutive convolutions to output features, and the number of channels for the features is set to $[32, 64, 64, 128, 128]$ respectively. All convolution layers are followed by a batch normalization and a rectified linear unit (ReLU). The downsamplings and upsamplings are performed with maxpool and bilinear interpolation. Our CNN has $\sim$1M trainable parameters, which is significantly smaller than conventional MDE networks.

3.4 PSF Regularization

Due to memory constraints, the E2E pipeline has to be trained using image patches, which are significantly smaller than the full sensor. Therefore, the PSF is optimized only over the size of the image patch and has no constraints outside the patch. However, the PSF may create non-zero energy outside the patch, which would reduce the contrast of captured images in practice. To prevent these artifacts, we penalize the energy of the PSF with the regularizer

$$\mathcal{L}_{\text{PSF}} = \sum_{\lambda \in (\text{R,G,B})} \sum_{k=0}^{K-1} \sum_{\rho > \rho_{\text{target}}} \left| \text{PSF}_k(\rho, \lambda) \right|^2, \quad (7)$$

where $\text{PSF}_k(\cdot, \lambda)$ is a 1D PSF evaluated over a full sensor size and ρ_{target} is a target PSF radius. Although the evaluation of the multi-color 3D PSF over a full sensor is computationally expensive, the regularizer (Eq. 7) is inexpensive to evaluate while having the same effectiveness due to the rotational symmetry of the PSF. In our training, we used the target radius of 32 pixels.

3.4.1 Training Loss Function

We train the network using feature similarity for the RGB image $\mathcal{L}_{\text{RGB}}$, an L1 loss for the depth map $\mathcal{L}_{\text{Depth}}$ and a regularizer for the PSF $\mathcal{L}_{\text{PSF}}$:

$$\mathcal{L} = \psi_{\text{RGB}} \mathcal{L}_{\text{RGB}} + \psi_{\text{Depth}} \mathcal{L}_{\text{Depth}} + \psi_{\text{PSF}} \mathcal{L}_{\text{PSF}}, \quad (8)$$

where ψ is the regularization weight. The feature similarity is evaluated on *input*, *conv1_2*, *conv2_2* and *conv3_3* features of a pre-trained VGG-16 network [77]. The losses for the features are weighted with $[0.250, 0.080, 0.250, 0.208]$ by following the fine-tuned weight used in [78]. Since the preconditioning with the approximate inverse has worse performance at the boundaries due to edge-tapering, the 32 pixels at the boundaries are excluded for the evaluation of the loss. We set $\psi_{\text{RGB}} = \psi_{\text{Depth}} = 1$ and $\psi_{\text{PSF}} = 45$, all of which are manually tuned.

3.5 Training Details

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 384×384 pixels and random horizontal/vertical flipping to augment the training set. The target depth range was set to $1.0\,\text{m}$ to $5.0\,\text{m}$, and the camera is focused at $1.7\,\text{m}$ 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.
- **DfD:** 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.

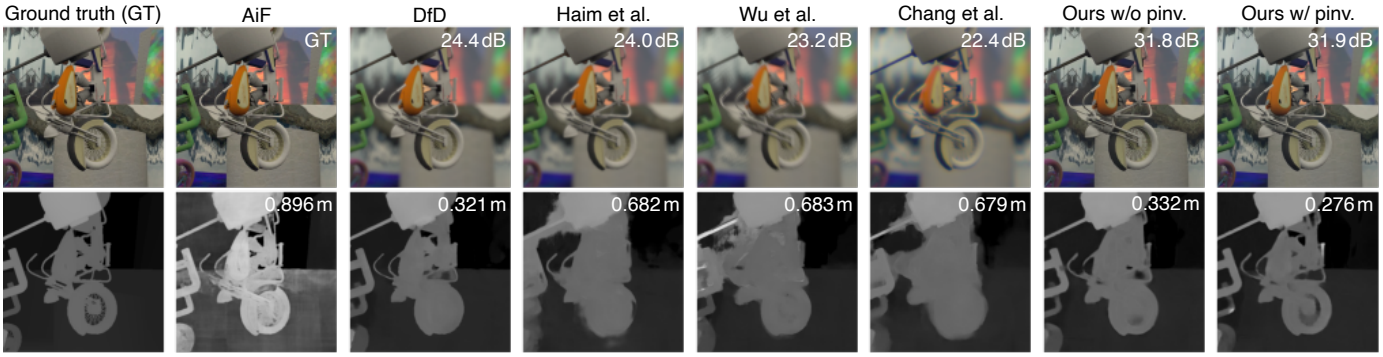


Fig. 3. *Top*: Ground truth (GT) RGB image (left) along with the simulated sensor images of all baseline approaches. These baselines (columns 3–6) do not attempt to estimate the GT RGB image. Estimated RGB images of our approach, without and with the proposed preconditioning. *Bottom*: GT and estimated depth maps of all approaches. The quality of RGB image and depth map estimated by our method is best for this scene. PSNR of the image and RMSE of the depth map are shown on the top right. See the Supplemental Material for their corresponding PSFs and captured images.

TABLE 1
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.

			Image			Depth					
	Model	Refinement	MAE↓	PSNR↑	SSIM↑	MAE↓	RMSE↓	$\log_{10}$↓	δ<1.25↑	δ<1.25^2↑	δ<1.25^3↑
Prior Work	All in focus (AiF)	—	GT	GT	GT	0.357	0.500	0.099	0.658	0.807	0.874
	Depth from def. (DfD)	—	3.24e-2	24.95	0.711	0.097	0.228	0.039	0.929	0.965	0.979
	Haim et al. [31]	—	3.28e-2	24.90	0.708	0.297	0.635	0.109	0.803	0.879	0.923
	Wu et al. [30]	—	3.49e-2	24.54	0.704	0.207	0.521	0.090	0.865	0.918	0.945
	Chang et al. [32]	—	3.62e-2	24.28	0.694	0.205	0.490	0.077	0.888	0.945	0.968
Rot. Symmetric	Linear w/o pinv	—	2.02e-2	30.01	0.870	0.268	0.598	0.108	0.845	0.898	0.925
	Linear w/ pinv	—	1.99e-2	30.86	0.891	0.258	0.554	0.103	0.856	0.899	0.927
	Linear w/o pinv	Nonlin. w/o pinv	1.89e-2	31.43	0.900	0.127	0.264	0.065	0.901	0.952	0.964
	Linear w/ pinv	Nonlin. w/ pinv	1.83e-2	31.58	0.902	0.095	0.203	0.038	0.931	0.969	0.979
	Nonlin. w/o pinv	—	1.82e-2	31.61	0.903	0.104	0.237	0.041	0.925	0.963	0.977
	Nonlin. w/ pinv	—	**1.76e-2**	**31.88**	**0.905**	**0.089**	**0.191**	**0.034**	**0.941**	**0.970**	**0.981**

- **Wu et al. [30]:** The PhaseCam3D approach implemented with a DOE size of 256×256 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 $\mathcal{L}_{\text{Depth}}$ with regularization of depth maps and the PSFs. For a fair comparison of optical models, we used only $\mathcal{L}_{\text{Depth}}$ 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

TABLE 2
Evaluating different weightings of the loss function.

Loss weights $(\psi_{\mathrm{RGB}}, \psi_{\mathrm{Depth}})$	Image PSNR	Depth RMSE
$(1.0, 1.0)$	31.88	0.191
$(1.0, 0.1)$	33.83	0.307
$(0.1, 1.0)$	29.91	0.184

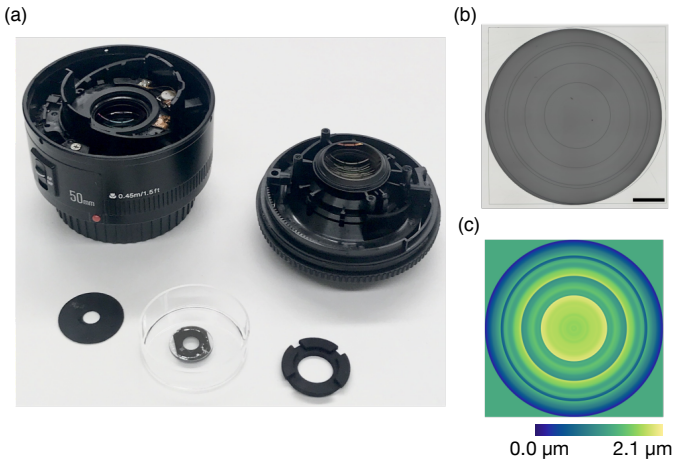


Fig. 4. (a) A disassembled camera lens next to our fabricated DOE with a 3D-printed mounting adapter. (b) A microscopic image of the fabricated DOE. The dark gray area is the DOE made of NOA61, and the light gray area is the light-blocking metal aperture made of chromium and gold. The black scale bar on the bottom right is $1\,\mathrm{mm}$. (c) The height profile of the designed DOE. The maximum height is $2.1\,\mu\mathrm{m}$.

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 *pinv* improves both image and depth quality compared to the results not using the *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 PSNR for depth RMSE with our choice of parameters (1.0,1.0) being a good tradeoff between the two.

5 Experimental Assessment

In this section, we discuss modifications to the training procedure that account for physical constraints as well as fabrication details and experimentally captured results.

5.1 Training for Camera Prototype

5.1.1 Additional Datasets

While the `FlyingThings3D` dataset provides complete depth maps aligned with RGB images, the images are synthetic and do not represent natural scenes. Therefore, we additionally used the `DualPixels` dataset [54] to learn the features of natural scenes. This dataset consists of a set of multi-view images and their depth maps captured by smartphone cameras. It has 2,506 captured images for training and 684 for validation. We used only the central view out of the five views for training. As the provided depth map is sparse, we inpainted the depth map to obtain a complete depth map [45], [81]. While the completed depth map is used for the simulated image formation during training, the loss function $\mathcal{L}_{\mathrm{Depth}}$ is evaluated only at valid (i.e., non-inpainted) depth values. Training images are drawn from `DualPixels` and `FlyingThings3D` with the same probability.

TABLE 3
Evaluating diffraction efficiency (DE) using RGB PSNR / depth RMSE metrics. We train DOEs from scratch for several different DEs (rows) and test them using the same and other DEs (columns).

Designed \ Tested	100%	75%	50%
100%	31.88 / 0.19	30.98 / 0.36	27.66 / 0.79
75%	29.79 / 0.27	31.74 / 0.19	29.50 / 0.35
50%	27.91 / 0.67	30.38 / 0.34	31.24 / 0.21

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 μ, we model the resulting PSF as

$$\mathrm{PSF} = \mu \cdot \mathrm{PSF}_{\mathrm{design}} + (1 - \mu) \cdot \mathrm{PSF}_{\mathrm{native}}. \tag{9}$$

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.

5.1.3 Robust Optimization of PSF

Our image formation model assumes shift invariance of the PSF on any one depth plane. In practice, however, the PSF slightly changes due to optical aberrations as visualized in the Supplemental Material. Moreover, discretizing the scene depths does not model the PSFs between the sampled depth planes. We empirically found that this discrepancy destabilizes the accuracy of our method when applied to experimentally captured data. To overcome this issue, we randomly shift the red and blue channels of the PSF with a maximum shift of 2 pixels during training, leaving the green channel fixed. Furthermore, each depth plane is randomly sampled in between equidistant depth planes for the PSF simulation per batch. The farthest plane is randomly sampled between $5\,\mathrm{m}$ and $100\,\mathrm{m}$.

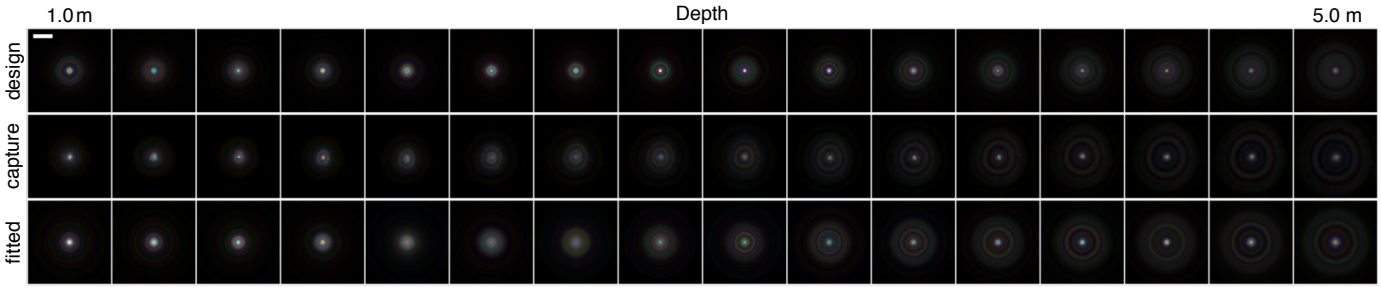


Fig. 5. Depth-dependent point spread functions (PSFs). The designed PSF (top row) is optimized with our end-to-end simulator. Optical imperfections result in the captured PSF (center row), slightly deviating from the design. Instead of working directly with the captured PSF, we fit a parametric model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

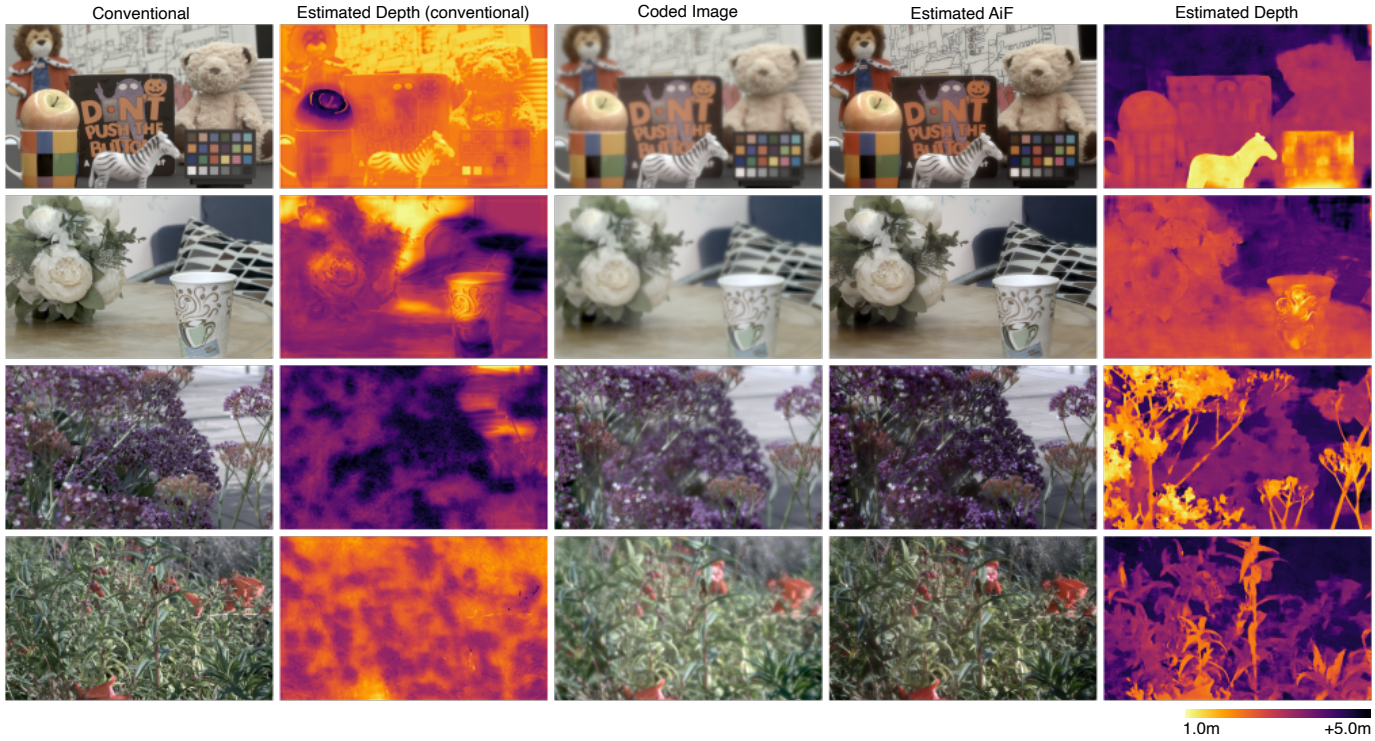


Fig. 6. Experimentally captured results of indoor and outdoor scenes. *From left:* Images of scenes captured with a conventional camera, depth maps estimated by a CNN from these conventional camera images, images captured by our phase-coded camera prototype with the optimized DOE, AiF images estimated by our algorithm from these coded sensor images, depth maps estimated by our algorithm from these coded sensor images. A top view of the indoor scene (top row) and the size of the receptive field of our neural network are visualized in the Supplemental Material.

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 $\psi_{\mathrm{RGB}} = \psi_{\mathrm{Depth}} = \psi_{\mathrm{PSF}} = 1$.

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 photoresist 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, $50\,\mathrm{mm}$, 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 $5.6\,\mathrm{mm}$ 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, party contributing to the undiffracted light. Specifically, we measured a misalignment of $\sim 140\,\mu\mathrm{m}$ between these two components.

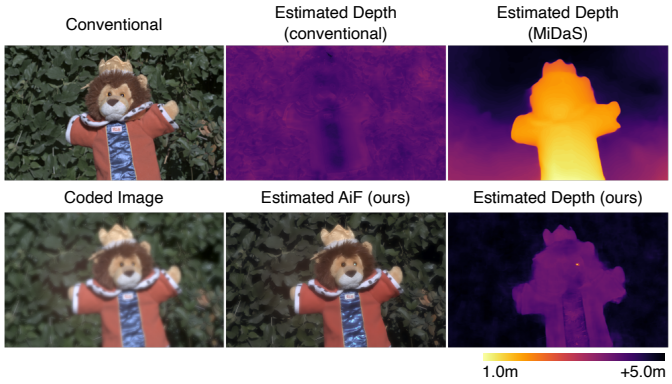


Fig. 7. Selected frames of experimentally captured dynamic scenes. The full dynamic scenes are available as supplemental movies. *From top-left to bottom-right:* an image of the scene captured with a conventional camera, a depth map estimated by a CNN comparable to ours from this conventional camera image, a depth map estimated from the conventional image by MiDaS [19], an image captured by our phase-coded camera prototype with the optimized DOE, an AiF image estimated by our CNN from this coded sensor image, and a depth map estimated by our CNN from the coded sensor image.

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 $15\,\mu\mathrm{m}$ 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 temporarily 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 $0.17\,\mathrm{m}$ 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

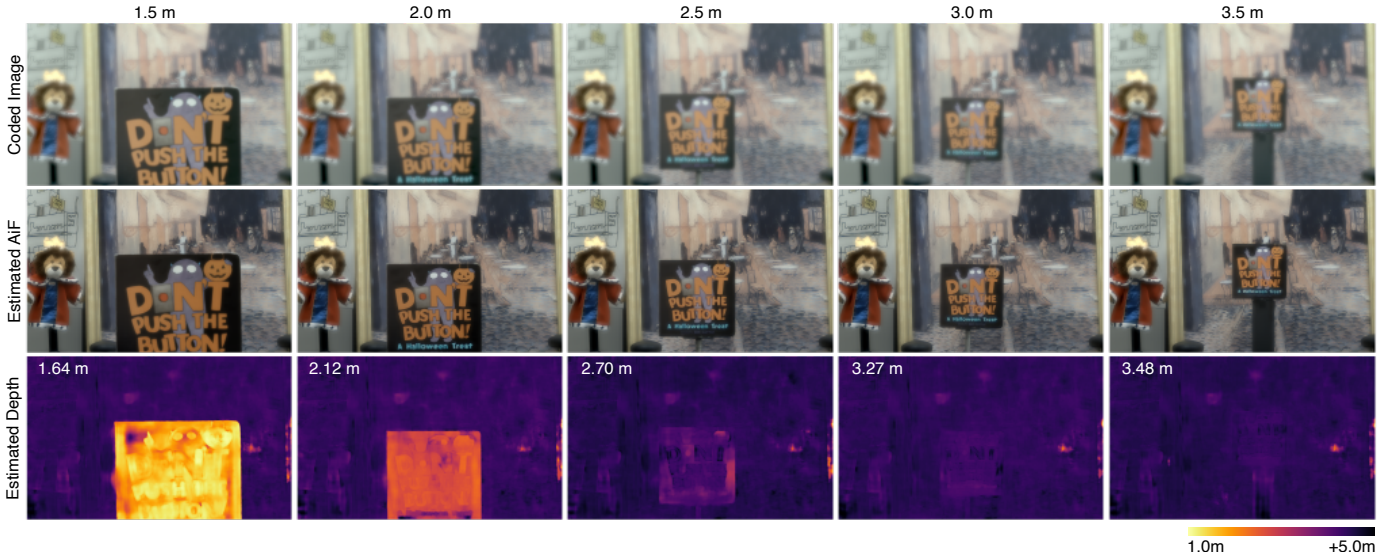


Fig. 8. Experimental quantitative analysis. A scene containing several objects, including a book, is photographed multiple times with the book positioned at different depths. The depth of this book is determined from the estimated depth maps. The root mean square error evaluated for all five depth planes is $0.17\,\mathrm{m}$.

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).

REFERENCES

[1] M. Ye, E. Johns, A. Handa, L. Zhang, P. Pratt, and G.-Z. Yang, "Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery," *arXiv:1705.08260*, 2017.

[2] A. Sabnis and L. Vachhani, "Single image based depth estimation for robotic applications," in *IEEE Recent Advances in Intelligent Computational Systems*, 2011, pp. 102–106.

[3] N. Metni, T. Hamel, and F. Derkx, "Visual tracking control of aerial robotic systems with adaptive depth estimation," in *IEEE Conference on Decision and Control*, 2005, pp. 6078–6084.

[4] J. Stowers, M. Hayes, and A. Bainbridge-Smith, "Altitude control of a quadrotor helicopter using depth map from Microsoft Kinect sensor," in *IEEE International Conference on Mechatronics*, 2011, pp. 358–362.

[5] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8445–8453.

[6] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1907–1915.

[7] W. Lee, N. Park, and W. Woo, "Depth-assisted real-time 3D object detection for augmented reality," in *International Conference on Artificial Reality and Telexistence (ICAT)*, vol. 11, no. 2, 2011, pp. 126–132.

[8] J. Kopf, K. Matzen, S. Alsisan, O. Quigley, F. Ge, Y. Chong, J. Patterson, J.-M. Frahm, S. Wu, M. Yu *et al.*, "One shot 3D photography," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 76–1, 2020.

[9] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Advances in Neural Information Processing Systems*, 2006, pp. 1161–1168.

[10] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems*, 2014, pp. 2366–2374.

[11] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1119–1127.

[12] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2024–2039, 2015.

[13] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *International Conference on 3D Vision (3DV)*, 2016, pp. 239–248.

[14] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5354–5362.

[15] J. Li, R. Klein, and A. Yao, "A two-streamed network for estimating fine-scaled depth maps from single rgb images," in *International Conference on Computer Vision (ICCV)*, 2017, pp. 3372–3380.

[16] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2002–2011.

[17] Y. Cao, T. Zhao, K. Xian, C. Shen, Z. Cao, and S. Xu, "Monocular depth estimation with augmented ordinal depth relationships," *arXiv:1806.00585*, 2018.

[18] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *arXiv:1812.11941*, 2018.

[19] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *arXiv:1907.01341*, 2020.

[20] A. P. Pentland, "A new sense for depth of field," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-9, no. 4, pp. 523–531, 1987.

[21] M. Watanabe and S. K. Nayar, "Rational filters for passive depth from defocus," *International Journal of Computer Vision*, vol. 27, pp. 203–225, 1998.

[22] P. Favaro, "Recovering thin structures via nonlocal-means regularization with application to depth from defocus," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1133–1140.

[23] P. Trouvé, F. Champagnat, G. Le Besnerais, J. Sabater, T. Avignon, and J. Idier, "Passive depth estimation using chromatic aberration and a depth from defocus approach," *Applied Optics*, vol. 52, no. 29, pp. 7152–7164, 2013.

[24] E. Alexander, Q. Guo, S. Koppal, S. Gortler, and T. Zickler, "Focal flow: Measuring distance and velocity with defocus and differential motion," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 667–682.

[25] M. Carvalho, B. Le Saux, P. Trouvé-Peloux, A. Almansa, and F. Champagnat, "Deep Depth from Defocus: how can defocus blur improve 3d estimation using dense neural networks?" in *European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[26] A. Levin, R. Fergus, F. Durand, and W. T. Freeman, "Image and depth from a conventional camera with a coded aperture," *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, pp. 70–es, 2007.

[27] A. Veeraraghavan, R. Raskar, A. Agrawal, A. Mohan, and J. Tumblin, "Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing," *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, p. 69, 2007.

[28] C. Zhou, S. Lin, and S. Nayar, "Coded aperture pairs for depth from defocus," in *International Conference on Computer Vision (ICCV)*, 2009, pp. 325–332.

[29] P. A. Shedligeri, S. Mohan, and K. Mitra, "Data driven coded aperture design for depth recovery," in *International Conference on Image Processing (ICIP)*, 2017, pp. 56–60.

[30] Y. Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghavan, "Phasecam3d—learning phase masks for passive single view depth estimation," in *International Conference on Computational Photography (ICCP)*. IEEE, 2019, pp. 1–12.

[31] H. Haim, S. Elmalem, R. Giryes, A. M. Bronstein, and E. Marom, "Depth estimation from a single image using deep learned phase coded mask," *IEEE Transactions on Computational Imaging*, vol. 4, no. 3, pp. 298–310, 2018.

[32] J. Chang and G. Wetzstein, "Deep optics for monocular depth estimation and 3d object detection," in *International Conference on Computer Vision (ICCV)*, 2019.

[33] J. A. Marshall, C. A. Burbeck, D. Ariely, J. P. Rolland, and K. E. Martin, "Occlusion edge blur: a cue to relative visual depth," *Journal of the Optical Society of America A*, vol. 13, no. 4, pp. 681–688, Apr 1996.

[34] S. E. Palmer and T. Ghose, "Extremal edge: A powerful cue to depth perception and figure-ground organization," *Psychological Science*, vol. 19, no. 1, pp. 77–83, 2008.

[35] M. Zannoli, G. D. Love, R. Narain, and M. S. Banks, "Blur and the perception of depth at occlusions," *Journal of Vision*, vol. 16, no. 6, pp. 17–17, 2016.

[36] L. He, M. Yu, and G. Wang, "Spindle-net: Cnns for monocular depth inference with dilation kernel method," in *International Conference on Pattern Recognition (ICPR)*, 2018, pp. 2504–2509.

[37] M. Heo, J. Lee, K.-R. Kim, H.-U. Kim, and C.-S. Kim, "Monocular depth estimation using whole strip masking and reliability-based refinement," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 36–51.

[38] F. Tosi, F. Aleotti, M. Poggi, and S. Mattoccia, "Learning monocular depth estimation infusing traditional stereo knowledge," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9799–9809.

[39] H. Zhan, R. Garg, C. Saroj Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 340–349.

[40] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 270–279.

[41] T.-C. Wang, A. A. Efros, and R. Ramamoorthi, "Depth estimation with occlusion modeling using light-field cameras," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 11, pp. 2170–2181, 2016.

[42] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "Geonet: Geometric neural network for joint depth and surface normal estimation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 283–291.

[43] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, "Lego: Learning edge with geometry all at once by watching videos," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 225–234.

[44] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[45] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 746–760.

[46] J.-H. Lee and C.-S. Kim, "Monocular depth estimation using relative depth maps," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9729–9738.

[47] J.-H. Lee, M. Heo, K.-R. Kim, and C.-S. Kim, "Single-image depth estimation based on fourier domain analysis," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 330–339.

[48] P. P. Srinivasan, R. Garg, N. Wadhwa, R. Ng, and J. T. Barron, "Aperture supervision for monocular depth estimation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6393–6401.

[49] L. He, G. Wang, and Z. Hu, "Learning depth from single images with deep neural network embedding focal length," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4676–4689, 2018.

[50] M. Nishimura, D. B. Lindell, C. Metzler, and G. Wetzstein, "Disambiguating monocular depth estimation with a single transient," in *European Conference on Computer Vision (ECCV)*, 2020.

[51] S. K. Nayar and Y. Nakagawa, "Shape from focus," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 8, pp. 824–831, 1994.

[52] M. Subbarao and G. Surya, "Depth from defocus: a spatial domain approach," *International Journal of Computer Vision*, vol. 13, no. 3, pp. 271–294, 1994.

[53] H. Tang, S. Cohen, B. Price, S. Schiller, and K. N. Kutulakos, "Depth from defocus in the wild," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2740–2748.

[54] R. Garg, N. Wadhwa, S. Ansari, and J. T. Barron, "Learning single camera depth estimation using dual-pixels," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 7628–7637.

[55] C. Zhou, S. Lin, and S. K. Nayar, "Coded aperture pairs for depth from defocus and defocus deblurring," *International Journal of Computer Vision*, vol. 93, no. 1, pp. 53–72, 2011.

[56] S. R. P. Pavani, M. A. Thompson, J. S. Biteen, S. J. Lord, N. Liu, R. J. Twieg, R. Piestun, and W. Moerner, "Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function," *Proceedings of the National Academy of Sciences*, vol. 106, no. 9, pp. 2995–2999, 2009.

[57] A. Levin, S. W. Hasinoff, P. Green, F. Durand, and W. T. Freeman, "4D frequency analysis of computational cameras for depth of field extension," *ACM Transactions on Graphics (TOG)*, vol. 28, no. 3, pp. 1–14, 2009.

[58] A. Chakrabarti, "Learning sensor multiplexing design through back-propagation," in *Advances in Neural Information Processing Systems*, 2016, pp. 3081–3089.

[59] L. Wang, T. Zhang, Y. Fu, and H. Huang, "HyperReconNet: Joint coded aperture optimization and image reconstruction for compressive hyperspectral imaging," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2257–2270, May 2019.

[60] E. Nehme, D. Freedman, R. Gordon, B. Ferdman, L. E. Weiss, O. Alalouf, R. Orange, T. Michaeli, and Y. Shechtman, "DeepSTORM3D: Dense three dimensional localization microscopy and point spread function design by deep learning," *Nature Methods*, vol. 17, pp. 734–740, 2020.

[61] Q. Sun, J. Zhang, X. Dun, B. Ghanem, Y. peng, and W. Heidrich, "End-to-end learned, optically coded super-resolution spad camera," *ACM Transactions on Graphics (TOG)*, vol. 39, 2020.

[62] V. Sitzmann, S. Diamond, Y. Peng, X. Dun, S. Boyd, W. Heidrich, F. Heide, and G. Wetzstein, "End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–13, 2018.

[63] X. Dun, H. Ikoma, G. Wetzstein, Z. Wang, X. Cheng, and Y. Peng, "Learned rotationally symmetric diffractive achromat for full-spectrum computational imaging," *Optica*, vol. 7, no. 8, pp. 913–922, 2020.

[64] C. A. Metzler, H. Ikoma, Y. Peng, and G. Wetzstein, "Deep optics for single-shot high-dynamic-range imaging," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1375–1385.

[65] Q. Sun, E. Tseng, Q. Fu, W. Heidrich, and F. Heide, "Learning rank-1 diffractive optics for single-shot high dynamic range imaging," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020.

[66] J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, "Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification," *Scientific Reports*, vol. 8, no. 1, p. 12324, 2018.

[67] J. Martel, L. Müller, S. Carey, P. Dudek, and G. Wetzstein, "Neural Sensors: Learning Pixel Exposures for HDR Imaging and Video Compressive Sensing with Programmable Sensors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, p. 1642–1653, 2020.

[68] Y. Li, M. Qi, R. Gulve, M. Wei, R. Genov, K. N. Kutulakos, and W. Heidrich, "End-to-end video compressive sensing using anderson-accelerated unrolled networks," in *International Conference on Computational Photography (ICCP)*, 2020, pp. 1–12.

[69] G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljacic, C. Denz, D. A. B. Miller, and D. Psaltis, "Inference in artificial intelligence with deep optics and photonics," *Nature*, vol. 588, 2020.

[70] S.-H. Baek, H. Ikoma, D. S. Jeon, Y. Li, W. Heidrich, G. Wetzstein, and M. H. Kim, "End-to-end hyperspectral-depth imaging with learned diffractive optics," *arXiv:2009.00436*, 2020.

[71] J. W. Goodman, *Introduction to Fourier optics*. Roberts and Company Publishers, 2005.

[72] S. W. Hasinoff and K. N. Kutulakos, "A layer-based restoration framework for variable-aperture photography," in *International Conference on Computer Vision (ICCV)*. IEEE, 2007, pp. 1–8.

[73] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo Magnification: learning view synthesis using multiplane images," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, 2018.

[74] X. Zhang, K. Matzen, V. Nguyen, D. Yao, Y. Zhang, and R. Ng, "Synthetic defocus and look-ahead autofocus for casual videography," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, Jul. 2019.

[75] S. J. Reeves, "Fast image restoration without boundary artifacts," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1448–1453, 2005.

[76] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention (MICCAI)*, 2015, pp. 234–241.

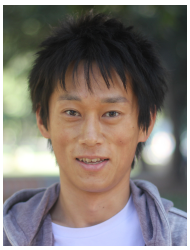
[77] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.

[78] J. Flynn, M. Broxton, P. Debevec, M. DuVall, G. Fyffe, R. Overbeck, N. Snavely, and R. Tucker, "DeepView: View synthesis with learned gradient descent," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 2367–2376.

[79] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4040–4048.

[80] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2462–2470.

[81] A. Levin, D. Lischinski, and Y. Weiss, "Colorization using optimization," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3, p. 689–694, Aug. 2004.



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.