# PhaseCam3D — Learning Phase Masks for Passive Single View Depth Estimation

Yicheng Wu[1], Vivek Boominathan[1], Huaijin Chen[1], Aswin Sankaranarayanan[2], and Ashok Veeraraghavan[1]
[1]Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005 USA
[2]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA

**There is an increasing need for passive 3D scanning in many applications that have stringent energy constraints. In this paper, we present an approach for single frame, single viewpoint, passive 3D imaging using a phase mask at the aperture plane of a camera. Our approach relies on an end-to-end optimization framework to jointly learn the optimal phase mask and the reconstruction algorithm that allows an accurate estimation of range image from captured data. Using our optimization framework, we design a new phase mask that performs significantly better than existing approaches. We build a prototype by inserting a phase mask fabricated using photolithography into the aperture plane of a conventional camera and show compelling performance in 3D imaging.**

***Index Terms*—computational photography, passive depth estimation, coded aperture, phase masks**

## I. Introduction

3D Imaging is critical for a myriad of applications such as autonomous driving, robotics, virtual reality, and surveillance. The current state of art relies on active illumination based techniques such as LIDAR, radar, structured illumination or continuous-wave time-of-flight. However, many emerging applications, especially on mobile platforms, are severely power and energy constrained. Active approaches are unlikely to scale well for these applications and hence, there is a pressing need for robust passive 3D imaging technologies.

Multi-camera systems provide state of the art performance for passive 3D imaging. In these systems, triangulation between corresponding points on multiple views of the scene allows for 3D estimation. Stereo and multi-view stereo approaches meet some of the needs mentioned above, and an increasing number of mobile platforms have been adopting such technology. Unfortunately, having multiple cameras within a single platform results in increased system cost as well as implementation complexity.

The principal goal of this paper is to develop a passive, single-viewpoint 3D imaging system. We exploit the emerging computational imaging paradigm, wherein the optics and the computational algorithm are co-designed to maximize performance within operational constraints.

### A. Key Idea

We rely on a bevy of existing literature on coded aperture [1]–[4]. It is well known that the the depth-dependent defocus 'bokeh' (point spread function) depends on the amplitude and phase of the aperture used. Is it possible to optimize a mask on the aperture plane with the exclusive goal of maximizing depth estimation performance?

We exploit recent advances in deep learning [5], [6] to develop an end-to-end optimization technique. Our proposed framework is shown in Figure 1, wherein the aperture mask and the reconstruction algorithm (in terms of the network parameters) for depth estimation are simultaneously optimized. To accomplish this, we model light propagation from the scene to the sensor, including the modulation by the mask as front-end layers of a deep neural network. Thus in our system, the first layer corresponds to physical optical elements. All subsequent layers of our network are digital layers and represent the computational algorithm that reconstructs depth images. We run the back-propagation algorithm to update this network, including the physical mask, end-to-end.

Once the network is trained, the parameters of the front-end provide us with the optimized phase mask. We fabricate this optimized phase mask and place it in the aperture plane of a conventional camera (Figure 2) to realize our 3D imaging system. The parameters of the back-end provide us with a highly accurate reconstruction algorithm, allowing us to recover the depth image from the captured data.

### B. Contributions

The main technical contributions of our work are as follows.

- We propose *PhaseCam3D*, a passive, single-viewpoint 3D imaging system that jointly optimizes the front-end optics (phase mask) and the back-end reconstruction algorithm.
- Using end-to-end optimization, we obtain a novel phase mask that provides superior depth estimation performance compared to existing approaches.
- We fabricated the optimized phase mask and build a coded aperture camera by integrated the phase mask into the aperture plane of the lens. We demonstrate compelling 3D imaging performance using our prototype.

Our current prototype system consists of a phase mask inserted into the aperture plane of a conventional imaging lens. In practice, it might be more efficient to fabricate a single optical element that accomplishes the task of both the main lens and the phase mask simultaneously. This would especially be the case for mobile platforms, where custom fabricated plastic lenses are the de-facto norm.

Manuscript received December 17, 2018; revised April 5, 2019. Corresponding author: Ashok Veeraraghavan (email: vashok@rice.edu).

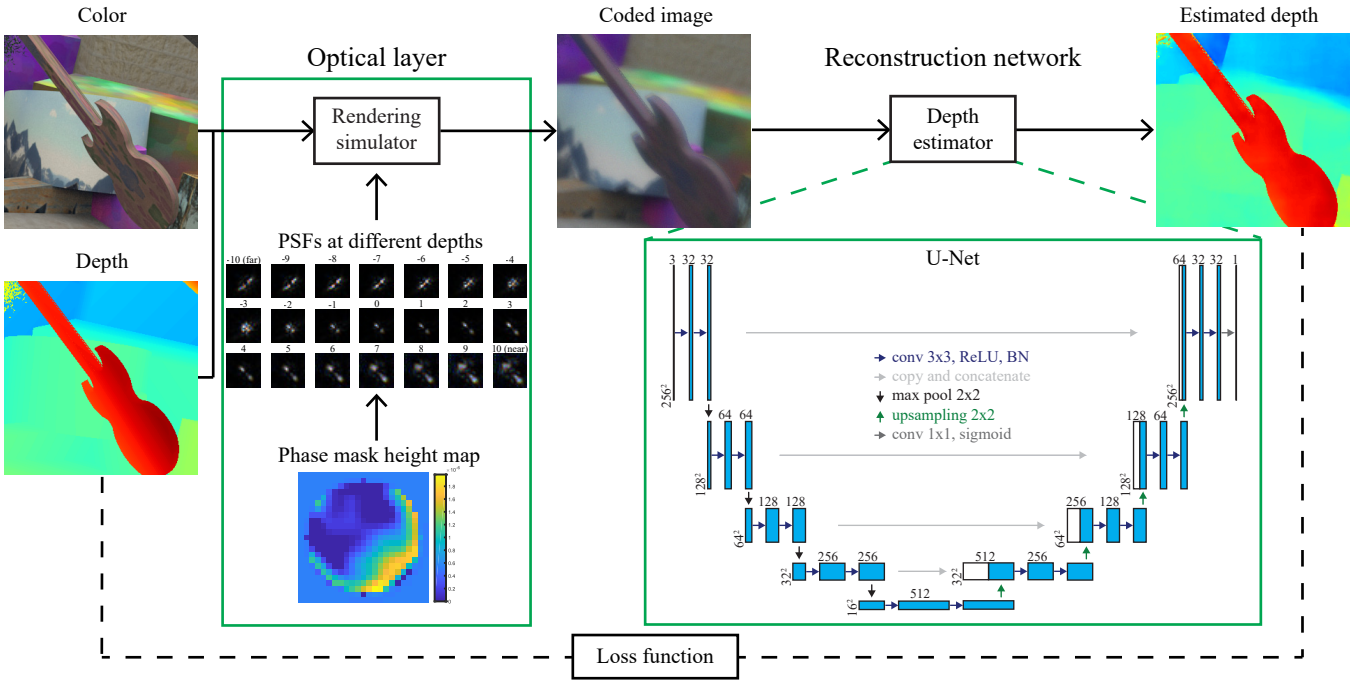


Fig. 1. **Framework overview.** Our proposed end-to-end architecture consists of two parts. In the optical layer, a physics-based model first simulates depth-dependent PSFs given a learnable phase mask, and then applies these PSFs to RGB-D input to formulate the coded image on the sensor. In the reconstruction network, a U-Net based network estimates the depth from the coded image. Both parameters in the optical layer, as well as the reconstruction network, are optimized based on the loss defined between the estimated depth and ground truth depth.

### C. Limitations

PhaseCam3D relies on the defocus cue which is not available in regions without texture. As a consequence, depth estimates obtained in texture-less regions are mainly through prior statistics and interpolation, both of which are implicitly learned by the deep neural network. Our results seem to indicate that the network has been able to successfully learn sufficient prior statistics to provide reasonable depth estimates even in texture-less regions. Nevertheless, large texture-less regions will certainly challenge our approach. Unlike most active approaches that provide per-pixel independent depth estimates, PhaseCam3D utilizes spatial blur to estimate depth and therefore will likely have a lower spatial resolution.

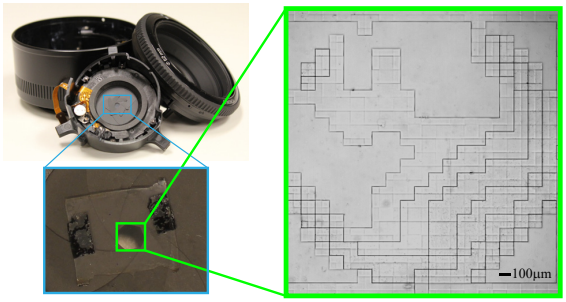


Fig. 2. **Fabricated phase mask.** A 2.835mm diameter phase mask is fabricated by photolithography and attached on the back side of the lens aperture. The image on the right shows a close-up image of the fabricated phase mask taken using a 2.5× microscope objective.

## II. Related Work

Image sensors capture 2D intensity information. Therefore, estimating the 3D geometry of the actual world from one or multiple 2D images is an essential problem in optics and computer vision. Over the last decades, numerous approaches were proposed for 3D imaging.

### A. Active Depth Estimation

When a coherent light source is available, holography is an ideal approach for 3D imaging. Holography [7] encodes the phase of the light in intensity based on the principle of wave interference. Once the interference image is recorded, the phase and therefore the 3D information can be derived [8], [9]. However, even though analog recording and reconstruction are straightforward (with even educational toy kits available now [10], [11]), the digital reconstruction process can be computationally expensive, and the requirement of the coherent light source and precise optical interference setup largely limited its usage in microscopy imaging [12]. With a more accessible incoherent light source, structured light [13] and time-of-flight (ToF) 3D imagers [14] became popular and made their ways to commercialized products, such as the Microsoft Kinect [15]. However, when lighting conditions are complex (i.e. outdoors under sunlight), given that both methods rely on active light sources, the performance of depth estimation can be poor. Therefore specialty hardware setup or additional computations are needed [16]–[18]. With a passive depth estimation method, such as the proposed PhaseCam3D, this problem can be avoided.

### B. Passive Depth Estimation

*a) Stereo vision:* One of the most widely used passive depth estimation methods is binocular or multi-view stereo (MVS). MVS is based on the principle that, if two or more cameras see the same point in the 3D scene from different viewpoints, granted the geometry and the location of the cameras, one can triangulate the location of the point in the 3D space [19]. Stereo vision can generate high-quality depth maps [20], and is deployed in many commercialized systems [21] and even the Mars Express Mission [22]. Similarly, structure from motion (SfM) use multiple images from a moving camera to reconstruct the 3D scene and estimate the trajectory and pose of the camera simultaneously [23]. However, both SfM and stereo 3D are fundamentally prone to occlusion [24]–[26] and texture-less areas [27], [28] in the scene; thus special handling of those cases have to be taken. Moreover, stereo vision requires multiple calibrated cameras in the setup, and SfM requires a sequence of input images, resulting in increased cost and power consumption and reduced robustness. In comparison, the proposed PhaseCam3D is single-view and single-shot, therefore, has much lower cost and energy consumption. Moreover, even though phase mask-based depth estimation relies on textures in the scene for depth estimation as well, PhaseCam3D's use of the data-driven reconstruction network can help to provide depth estimation with implicit prior statistics and interpolation from the deep neural networks.

*b) Coded aperture:* Previously, amplitude mask designs have demonstrated applications in depth estimation [1], [2] and light-field imaging [3]. PhaseCam3D uses novel phase mask to help with the depth estimation, and the phase mask-based approach provides several advantages compared to amplitude maks: First, unlike the amplitude masks that block the light, phase masks bend light, thus has much higher light throughput, consequently delivers lower noise level. Secondly, the goal of designing the mask-based imaging system for depth estimation is to make the point spread functions (PSFs) of different depth to have maximum variability. Even though the PSFs of amplitude mask-based system is depth dependent, the difference in PSFs across depth is only in scale. On the contrary, phase masks produce PSFs with much higher depth dependent variability. As a result, the phase mask should help distinguish the depth better in theory and the feature size can be made smaller. Lastly, the phase mask also preserves cross-channel color information, which could be useful for reconstruction algorithms. Recently, Haim *et al.* [4] demonstrate to use a phase mask for depth estimation. However, they only explore a two-ring structure, which constrains the design space with limited PSF shapes, whereas our PhaseCam3D has a degree of freedom (DoF) of 55 given the Zernike basis we choose to use, described in Section III-D(a).

### C. Semantics-based Single Image Depth Estimation

More recently, deep learning based single-image depth estimation methods demonstrated that high-level semantics itself can be useful enough for depth estimation without any physics-based models [29]–[35]. However, while those results sometimes appear visually pleasing, they might deviate from reality and usually have a low spatial resolution, thus getting the precise absolution depth is difficult. Some recent work suggested to add physics-based constraints elevated the problems [36]–[39], but extra inputs such as multiple viewpoints were required. In addition, many of those methods focus and work very well on certain benchmark datasets, such as NYU Depth [40], KITTI [41], but the generalization to scenes in the wild beyond the datasets is unknown.

### D. End-to-end Optimization of Optics and Algorithms

Deep learning has now been used as a tool for end-to-end optimization of the imaging system. The key idea is to model the optical imaging formation models as parametric neural network layers, connect those layers with the application layers (i.e., image recognition, reconstruction, etc.) and finally use back-propagation to train on a large dataset to update the parameters in optics design. An earlier example is designing the optimal Bayer color filter array pattern of the image sensor [5]. More recently, [6] shows that the learned diffractive optical element achieves a good result for achromatic extended depth of field. Haim *et al.* [4] learned the phase mask and reconstruction algorithm for depth estimation using Deep learning. However, their framework is not entirely end-to-end, since their phase mask is learned by a separate depth classification algorithm besides the reconstruction network, and the gradient back-propagation is performed individually for each network. Such a framework limits their ability to find the optimal mask for depth estimation.

## III. PhaseCam3D Framework

We consider a phase mask-based imaging system capable of reproducing the 3D scenes with single image capture. Our goal is to achieve state-of-the-art single image depth estimation results with jointly optimized front-end optics along with the back-end reconstruction algorithm. We achieve this via end-to-end training of a neural network for the joint optimization problem. As shown in Figure 1, our proposed solution network consists of two major components: 1) a differentiable optical layer, whose learnable parameter is the height map of the phase mask, that takes in as input an all-in-focus image and a corresponding depth map and outputs a physically-accurate coded intensity image; and 2) a U-Net based deep network to reconstruct the depth map from the coded image.

During the training, the RGB all-in-focus image and the corresponding ground truth depth are provided. The optical layer takes this RGB-D input and generates the simulated sensor image. This phase-modulated image is then provided as input to the reconstruction network, which outputs the estimated depth. Finally, the loss between the estimated depth and ground truth depth is calculated. From the calculated loss, we back-propagate the gradient to update both the reconstruction network and the optical networks. As a result, the parameters in the reconstruction network, as well as the phase mask design, are updated.

We next describe our proposed system components in detail.

### A. Optical Layer

To simulate the system accurately, we model our system based on Fourier optics theory [42], which takes account for diffraction and wavelength dependence. To keep the consistency with natural lighting conditions, we assume that the light source is incoherent.

The optical layer simulates the working of a camera with a phase mask in its aperture plane. Given the phase mask, describes as a height map, we can first define the pupil function induced by it, calculate the point spread function on the image plane and render the coded image produced by it given an RGBD image input.

*a) Pupil function:* Since the phase mask is placed on the aperture plane, the pupil function is the direct way to describe the forward model. The pupil function is a complex-valued function of the 2D coordinates $(x_1, y_1)$ describing the aperture plane.

$$P(x_1, y_1) = A(x_1, y_1) \exp[i\phi(x_1, y_1)] \tag{1}$$

The amplitude $A(\cdot, \cdot)$ is constant within the disk aperture and zero outside since there is no amplitude attenuation for phase masks. The phase $\phi$ has two components from the phase mask and defocus.

$$\phi(x_1, y_1) = \phi^M(x_1, y_1) + \phi^{DF}(x_1, y_1) \tag{2}$$

$\phi^M(x_1, y_1)$ is the phase modulation caused by height variation on the mask.

$$\phi^M(x_1, y_1) = k_\lambda \Delta n \; h(x_1, y_1) \tag{3}$$

$\lambda$ is the wavelength, $k_\lambda = \frac{2\pi}{\lambda}$ is the wave vector, and $\Delta n$ is the reflective index difference between air and the material of the phase mask. The material used for our phase mask has little refractive index variations in the visible spectrum [43]; so, we keep $\Delta n$ as a constant. $h$ denotes the height map of the mask, which is what we need to learn in the optical layer.

The term $\phi^{DF}(x_1, y_1)$ is the defocus aberration due to the mismatch between in-focus depth $z_0$ and the actual depth $z$ of a scene point. The analytical expression for $\phi^{DF}(x_1, y_1)$ is given as [42]

$$\phi^{DF}(x_1, y_1) = k_\lambda \frac{x_1^2 + y_1^2}{2} \left( \frac{1}{z} - \frac{1}{z_0} \right) = k_\lambda W_m r(x_1, y_1)^2, \tag{4}$$

where $r(x_1, y_1) = \sqrt{x_1^2 + y_1^2}/R$ is the relative displacement, $R$ is the radius of the lens aperture, and $W_m$ is defined as

$$W_m = \frac{R^2}{2} \left( \frac{1}{z} - \frac{1}{z_0} \right). \tag{5}$$

$W_m$ combines the effect from the aperture size and the depth range, which is a convenient indication of the severity of the focusing error. For depths that are closer to the camera than the focal plane, $W_m$ is positive. For depths that are further than the focal plane, $W_m$ is negative.

*b) PSF induced by the phase mask:* For an incoherent system, the PSF is the squared magnitude of the Fourier transform of the pupil function.

$$PSF_{\lambda, W_m}(x_2, y_2) = \left| \mathcal{F}\{P_{\lambda, W_m}(x_1, y_1)\} \right|^2 \tag{6}$$

The PSF is dependent on the wavelength of the light source and defocus. In the numerical simulations, the broadband color information in the training datasets — characterized as red (R), blue (B) and green (G) channels — are approximated by three discretized wavelengths, 610 nm (R), 530 nm (G) and 470 nm (B), respectively.

*c) Coded image formulation:* If the scene is comprised of a planar object at a constant depth from the camera, the PSF is uniform over the image, and the image rendering process is just a simple convolution for each of the color channels. However, most real-world scenes contain depth variations, and the ensuing PSF is spatially varying. While there are plenty of algorithms to simulate the depth-of-field effect [44]–[46], we require four fundamental properties to be satisfied. First, the rendering process has to be physically accurate and not just photo-realistic. Second, it should have the ability to model arbitrary phase masks and the PSF induced by them, rather than assuming a specific model on the PSF (e.g., Gaussian distribution). Third, since the blurring process will be one part of the end-to-end framework, it has to be differentiable. Fourth, this step should be computationally efficient because the rendering process needs to be done for each iteration with updated PSFs.

Our method is based on the layered depth of field model [45]. The continuous depth map is discretized based on $W_m$. Each layer is blurred by its corresponding PSF calculated from (6) with a convolution. Then, the blurred layers are composited together to form the image.

$$I_\lambda^B(x_2, y_2) = \sum_{W_m} I_{\lambda, W_m}^S(x_2, y_2) \otimes PSF_{\lambda, W_m}(x_2, y_2) \tag{7}$$

This approach does not model the occlusion and hence, the rendered image is not accurate near the depth boundaries due to intensity leakage; however, for the most part, it does capture the out-of-focus effect correctly. We will discuss fine-tuning of this model to reduce the error at boundaries in Section V-D.

To mimic noise during the capture, we apply Gaussian noise to the image. A smaller noise level will improve the performance during the reconstruction but also makes the model to be more sensitive to noise. In our simulation, we set the standard deviation $\sigma = 0.01$.

### B. Depth Reconstruction Network

There are a variety of networks to be applied for our depth estimation task. Here, we adopt the U-Net [47] since it is widely used for pixel-wise prediction.

The network is illustrated in Figure 1, which is an encoder-decoder architecture. The input to the network is the coded image with three color channels. The encoder part consists of the repeated application of two $3 \times 3$ convolutions, each followed by a rectified linear unit (ReLU) and a batch normalization (BN) [48]. At each downsampling step, we halve the resolution using a $2 \times 2$ max pooling operation with stride 2 and double the number of feature channels. The decoder part consists of an upsampling of the feature map followed by a $2 \times 2$ convolution that halves the number of feature channels and two $3 \times 3$ convolutions, each followed by a ReLU and a

BN. Concatenation is applied between the encoder and decoder to avoid the vanishing gradient problem. At the final layer, a 1x1 convolution is used with a sigmoid to map each pixel to the given depth range.

During the training, the input image size is $256 \times 256$. But the depth estimation network can be run fully-convolutionally for images size of any multiple of 16 at test time.

### C. Loss Function

Instead of optimizing depth $z$ directly, we optimize $W_m$ which is linear to the inverse of the depth. Intuitively, since defocus blur is proportional to the inverse of the depth, estimating depth directly would be highly unstable since even a small perturbation in defocus blur estimation could potentially lead to an arbitrarily large change in depth. Further, since $W_m$ is relative to the depth of the focus plane, it removes an additional degree of freedom that would otherwise need to be estimated. Once we estimate $W_m$, the depth map can be calculated using (5).

We use a combination of multiple loss functions

$$L_{\text{total}} = \lambda_{RMS} L_{\text{RMS}} + \lambda_{grad} L_{\text{grad}} + \lambda_{CRLB} L_{\text{CRLB}} \quad (8)$$

Empirically, we found that setting the weights of the respective loss functions (if included) as $\lambda_{RMS} = 1$, $\lambda_{grad} = 1$, and $\lambda_{CRLB} = 1e^{-4}$ generates good results. We describe each loss function in detail.

- *Root Mean Square (RMS).* In order to force the estimated $\widehat{W}_m$ to be similar to the ground truth $W_m$, we define a loss term using the RMS error.

$$L_{\text{RMS}} = \frac{1}{\sqrt{N}} \|W_m - \widehat{W}_m\|_2, \quad (9)$$

  where $N$ is the number of pixels.
- *Gradient.* In a natural scene, it is common to have multiple objects located at different depths, which creates sharp boundaries in the depth map. To emphasize the network to learn these boundaries, we introduce an RMS loss on the gradient along both $x$ and $y$ directions.

$$L_{\text{grad}} = \frac{1}{\sqrt{N}} \left( \left\| \frac{\partial W_m}{\partial x} - \frac{\partial \widehat{W}_m}{\partial x} \right\| + \left\| \frac{\partial W_m}{\partial y} - \frac{\partial \widehat{W}_m}{\partial y} \right\| \right) \quad (10)$$

- *Cramér-Rao Lower Bound (CRLB).* The effectiveness of depth-varying PSF to capture the depth information can be expressed using a statistical information theory measure called the Fisher information. Fisher information provides a measure of the sensitivity of the PSF to changes in the 3D location of the scene point [49]. Using the Fisher information function, we can compute CRLB, which provides the fundamental bound on how accurately a parameter (3D location) can be estimated given the noisy measurements. In our problem setting, the CRLB provides a scene-independent characterization of our ability to estimate the depth map. Prior work on 3D microscopy [49] has shown that optimizing a phase mask using CRLB as the loss function provides diverse PSFs for different depths.

  The Fisher information matrix, which is a $3 \times 3$ matrix in our application, is given as

$$I_{ij}(\theta) = \sum_{t=1}^{N_p} \frac{1}{PSF_\theta(t) + \beta} \left( \frac{\partial PSF_\theta(t)}{\partial \theta_i} \right) \left( \frac{\partial PSF_\theta(t)}{\partial \theta_j} \right), \quad (11)$$

  where $PSF_\theta(t)$ is the PSF intensity value at pixel $t$, $N_p$ is the number of pixels in the PSF, and $\theta = (x, y, z)$ corresponds to the 3D location.

  The diagonal of the inverse of the Fisher information matrix yields the CRLB vector, which bounds the variance of the 3D location.

$$\text{CRLB}_i \equiv {\sigma_i}^2 = E(\hat{\theta}_i - \theta_i)^2 \geq \left[ (I(\theta))^{-1} \right]_{ii} \quad (12)$$

  Finally, the loss is a summation of CRLB for different directions, different depths, and different colors.

$$L_{\text{CRLB}} = \sum_{i=\hat{x},\hat{y},\hat{z}} \sum_{z \in Z} \sum_{c=R,G,B} \sqrt{\text{CRLB}_i(z,c)} \quad (13)$$

  In theory, smaller $L_{\text{CRLB}}$ indicates better 3D localization.

### D. Training / Implementation Details

We describe key elements of the training procedure used to perform the end-to-end optimization of the phase mask and reconstruction algorithm.

*a) Basis for height maps:* Recall that the phase mask is described in terms of a height map. We describe the height map at a resolution of $23 \times 23$ pixels. To speed up the optimization convergence, we constrain the height map further by modeling it using the basis of Zernike polynomials [50]; this approach was used previously by [49]. Specifically, we constrain the height map to the of the form

$$h(x, y) = \sum_{j=1}^{55} a_j Z_j(x, y) \quad (14)$$

where $\{Z_j(x, y)\}$ is the set of Zernike polynomials. The goal now is to find the optimal coefficient vector $\mathbf{a}^{1 \times 55}$ that represents the height map of the phase mask.

*b) Depth range:* We choose the range of $k_G W_m$ to be $[-10.5, 10.5]$. The term $k_G$ is the wave vector for green wavelength ($k_G = \frac{2\pi}{\lambda_G}$; $\lambda_G = 530nm$) and we choose the range of $k_G W_m$ so that the defocus phase $\phi^{DF}$ is within a practical range, as calculated by (4). For the remainder of the paper, we will refer to $k_G W_m$ as the normalized $W_m$.

During the image rendering process, $W_m$ needs to be discretized so that the clean image is blurred layer by layer. There is a tradeoff between the rendering accuracy and speed. For the training, we discretize normalized $W_m$ to $[-10 : 1 : 10]$, so that it has 21 distinct values.

*c) Datasets:* As discussed in the framework, our input data requires both texture and depth information. The NYU Depth dataset [51] is a commonly used RGBD dataset for depth-related problems. However, since Kinect captures the ground-truth depth map, the dataset has issues in boundary mismatch and missing depth. Recently, synthetic data has been applied to geometric learning tasks because it is fast and

TABLE I
QUANTITATIVE EVALUATION OF ABLATION STUDIES

| Exp. | Learn mask | Initialization | Loss | Error (RMS) |
|---|---|---|---|---|
| A | No | No mask | RMS | 2.69 |
| B | Yes | Random | RMS | 1.07 |
| C | No | Fisher mask | RMS | 0.97 |
| D | Yes | Random | RMS+CRLB | 0.88 |
| E | Yes | Fisher mask | RMS | 0.74 |
| F | Yes | Fisher mask | RMS+CRLB | 0.85 |
| **G** | **Yes** | **Fisher mask** | **RMS+gradient** | **0.56** |

cheap to produce and contains precise texture and depth. We use FlyingThings3D from Scene Flow Datasets [40], which includes both all-in-focus RGB images and corresponding disparity map for 2247 training scenes. Each scene contains ten successive frames. We used the first and last frames in each sequence to avoid redundancies.

To accurately generate $256 \times 256$ coded images using PSFs of size $23 \times 23$ pixels, we need all-in-focus images at a resolution $278 \times 278$ pixels. We generate such data by cropping patches of appropriate size from the original images (whose resolution is $960 \times 540$) with a sliding window of 200 pixels. We only select the image whose disparity map ranges from 3 to 66 pixels and convert them to $W_m$ linearly.

With this pre-processing, we obtain 5077 training patches, 553 validation patches, and 419 test patches. The data is augmented with rotation and flip, as well as brightness scaling randomly between 0.8 to 1.1.

*d) Training process:* Given the forward model and the loss function, the back-propagation error can be derived using the chain rule. In our system, the back-propagation is obtained by the automatic differentiation implemented in TensorFlow [52]. For those who are interested in the derivation for the optical layer, please refer to our supplementary material. During the training, we use Adam [53] optimizer with parameters $\beta_1 = 0.99$ and $\beta_2 = 0.999$. Empirically, we found that using different learning rates for the phase mask and depth reconstruction improves the performance. We suspect this is due to the large influence that the phase mask has on the U-Net given that even small changes to the mask produces large changes in the coded image. In our simulation, the learning rates for phase mask and depth reconstruction were $10^{-8}$ and $10^{-4}$, respectively. A learning rate decay of $0.1$ was applied at $10K$ and $20K$ iterations. We observed that the training converges after about $30K$ iterations. We used a training mini-batch size to be 40. Finally, the training and testing were performed on NVIDIA Tesla K80 GPUs.

## IV. SIMULATION

The end-to-end framework learns the phase mask design and reconstruction algorithm in the simulation. In this section, We perform ablation studies to identify elements that contribute most to the overall performance as well as identify the best operating point. Finally, we provide comparisons with other depth estimation methods using simulations.

### A. Ablation Studies

To clearly understand our end-to-end system as well as choosing the correct parameters in our design space, we carry out several ablation experiments. We discuss our findings below, provide quantitative results in Table I and the qualitative visualizations in Figure 3. For convenience, we use the numbering in the first column of Table I when referring to the experiment performed and the corresponding models acquired in the ablation study. For all the experiments here, we use the same U-Net architecture as discussed in Section III-B for depth reconstruction. The baseline for all comparison is model (A), a depth-reconstruction-only network trained with a fixed open aperture and RMS loss.

*a) Learned vs. fixed mask:* In this first experiment, we use our end-to-end framework to learn both the phase mask and the reconstruction layer parameters from *randomly initialized* values (Exp. B). For comparison, we have Exp. C where the phase mask is fixed to the Fisher mask, which is designed by minimizing $L_{\mathrm{CRLB}}$ in our depth range, and we learn only the reconstruction layer from random initialization.

To our surprise, shown in Table I and Figure 3 (Exp. B vs. C), when learning from scratch (random phase mask parameters), our end-to-end learned masks (B) underperforms the Fisher mask that was designed using a model-based approach (C). We believe that there are two insights to be gained from this observation. First, the CRLB cost is very powerful by itself and leads to a phase mask that is well suited for depth estimation; this is expected given the performance of prior work that exploits the CRLB cost. Second, a random initialization fails to converge to the desired solution in part due to the highly non-convex nature of the optimization problem and the undue influence of the initialization. We visualize the corresponding phase mask height map is visualized in Figure 4, where 4(a) is the mask learned from scratch in Exp. B, and 4(b) is the fixed Fisher in Exp. C.

*b) Effect of initialization conditions:* With our hypothesis drawn from the previous experiment, we explore if careful initialization would help in improving overall performance. Instead of initializing with random values in Exp. B, we initialize the mask as a Fisher mask in Exp. E, and perform end-to-end optimization of both the mask design and the reconstruction network (there is no constraint forcing the optical network to generate masks that are close to the Fisher mask). Interestingly, under such an initialization, the end-to-end optimization improves the performance compared to the randomly initialized mask (B) by a significant margin (1.07 vs. 0.74 in RMS), and it also out-performs the fixed Fisher mask (Exp. C) noticeably (0.97 vs. 0.74 in RMS), suggesting the CRLB-model-based mask design can be further improved by data-driven fine-tuning. This is reasonable given that the model-based mask design does not optimize directly on the end objective – namely, a high-quality precise depth map that can capture both depth discontinuities and smooth depth variations accurately. Fisher mask is the optimal solution for 3D localization when the scene is sparse [49]. However, most real-world scenes are not sparse and hence optimizing for the actual depth map allows us to beat the performance of the

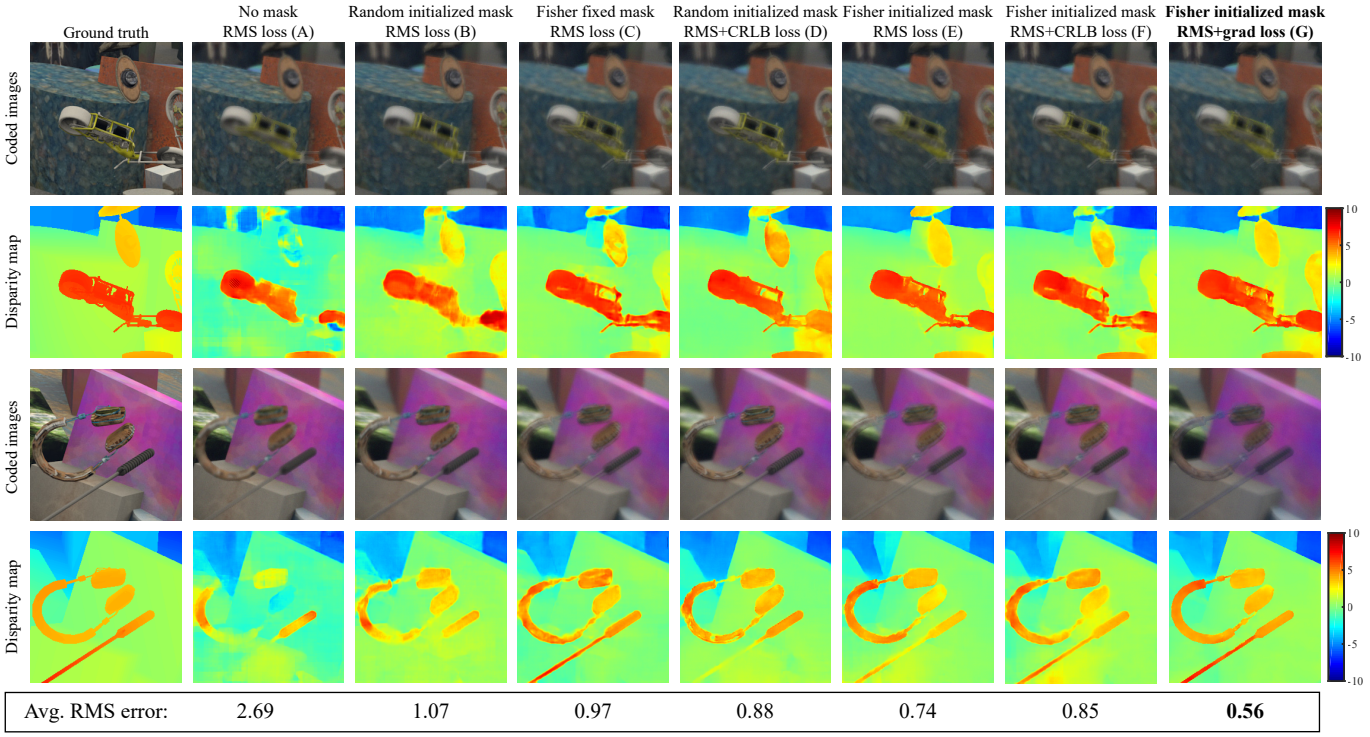


Fig. 3. **Qualitative results from our ablation studies.** Across the columns, we show the inputs to the reconstruction network and the depth estimation results from the network. The numbering A-G here correspond to the experiment setup A-G in Table I. The best result is achieved when we initialize the optical layer with the phase mask derived using Fisher information and then letting the CNN further optimize the phase mask. The last column (G) shows the results from our best phase mask.

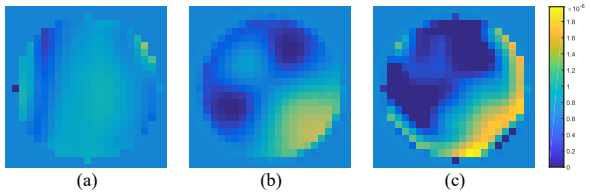


Fig. 4. **Phase mask height maps from ablation studies.** (a) Trained from random initialization with RMS loss. (b) Fisher initialized mask. (c) Trained from Fisher initialization with RMS and gradient loss.

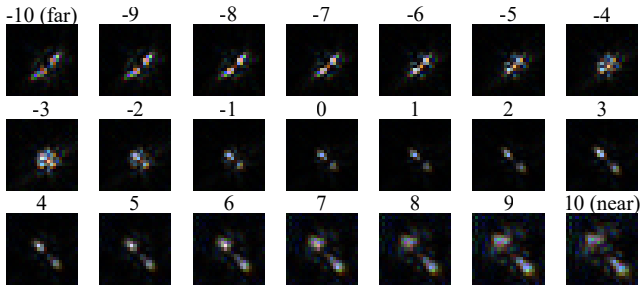


Fig. 5. **Simulated PSFs of our optimal phase mask.** The PSFs are labeled in terms of $W_m$. Range $-10$ to 10 corresponds to the depth plane from far to near.

Fisher mask.

The use of Fisher mask to initialize the network might raise the concern whether the proposed approach is still end-to-end. We believe the answer is positive, because initializing a network from designed weights instead of from scratch is a common practice in deep learning (i.e., the Xavier approach [54] and the He approach [55]). Likewise, here we incorporate our domain knowledge and use a model-based approach in designing the initialization condition of our optical layers.

*c) Effect of loss functions:* Finally, we also test different combinations of Losses discussed in Section III-C with the Fisher mask as the initialization (E, F, and G). We found that RMS with gradient loss (G) gives the best results. For completeness, we also show the performance of randomly initialized mask with RMS and CRLB loss in D.

### *B. Operating Point with Best Performance*

Figure 4(c) shows the best phase mask design based on our ablation study. It shares some similarity with the Fisher mask since we take the Fisher mask as our initialization. But our mask is further optimized based on the depth map from our data. Figure 5 displays depth-dependent PSFs in the range $[-10 : 1 : 10]$ of normalized $W_m$. These PSFs have large variability across different depths for improving the performance of depth estimation. More simulation results are shown in Figure 6.

### *C. Comparisons with the State-of-the-Art*

We compare our result with state-of-the-art passive, single viewpoint depth estimation methods.

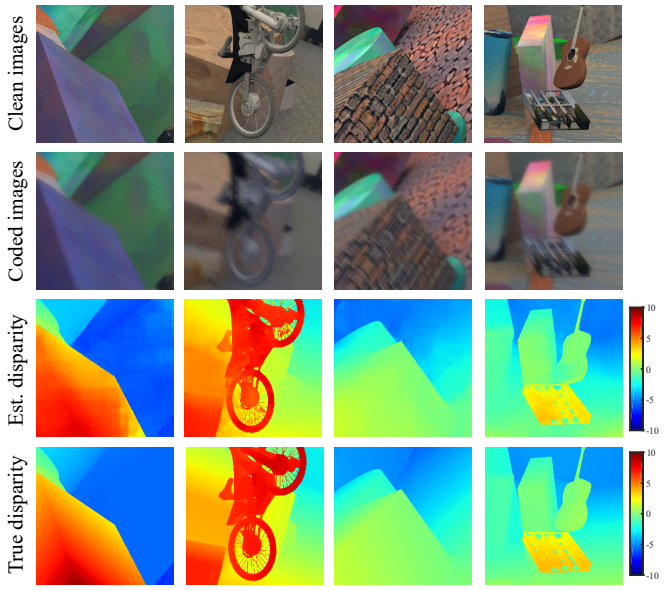


Fig. 6. **Simulation results with our best phase mask.** The reconstructed disparity maps closely match the ground truth disparity maps. The scaled disparity map have units in terms of normalized $W_m$.

TABLE II
COMPARISON WITH AMPLITUDE MASK DESIGN

| Mask design | $L_{RMS}$ |
|---|---|
| Levin *et al.* [1] | 1.04 |
| Veeraraghavan *et al.* [3] | 1.08 |
| **Ours** | **0.56** |

*a) Coded amplitude masks:* There are two well-known amplitude masks for depth estimation. Levin *et al.* [1] design a mask by maximizing the blurry image distributions from different depths using Kullback-Leibler divergence. Veeraraghavan *et al.* [3] select the best mask by maximizing the minimum of the discrete Fourier transformation magnitudes of the zero padded code. To make a fair comparison between their masks and our proposed mask, we render blurry image datasets based on each mask with the same noise level ($\sigma = 0.01$). Since U-Net is a general pixel-wise estimation network, we use it with same architecture introduced in III-B for depth reconstruction. Parameters in the U-Net are learned for each dataset using RMS and gradient loss.

The quantitative results are shown in Table II and qualitative results are shown in Figure 7. Our proposed mask offers the best result with the smallest RMS error. One key reason is that these amplitude masks only change the scaling factor of PSF at different depths, while our mask creates a more dramatic difference in PSF at different depths.

*b) Two-ring phase mask:* Recently, Haim *et al.* [4] propose a two-ring phase mask for depth estimation. To compare the performance, we use their dataset "TAU-Agent" and the same parameters described in their paper. Performance is evaluated by the $L_1$ loss of $W_m$. As shown in Table III, both our reconstruction network and our phase mask contribute to achieving smallest estimation error.

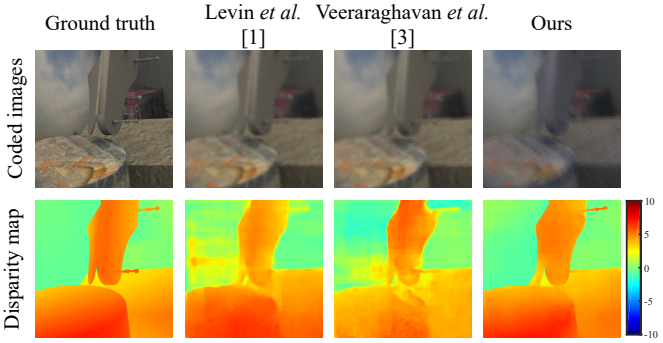


Fig. 7. **Depth estimation comparing with coded amplitude masks.** Our reconstructed disparity map achieves the best performance. Also, our system has higher light efficiency by using the phase mask. The scaled disparity map have units in terms of normalized $W_m$.

TABLE III
COMPARISON WITH THE TWO-RING PHASE MASK [4]

| Method | $\lvert W_m - \hat{W}_m \rvert$ |
|---|---|
| Two-ring mask + Haim's network | 0.6 |
| Two-ring mask + U-Net | 0.51 |
| **Our Optimized Mask + U-Net** | **0.42** |

*c) Semantics-based single image depth estimation:* To compare the performance of our proposed methods with other deep-learning-based depth estimation methods using a single all-focus image, we run evaluation experiments on standard NYU Depth V2 datasets [51]. We used the default training/testing splits provided by the datasets. The size of training and testing images are re-sized from $640 \times 480$ to $320 \times 240$ following the data augmentations the common practice [29]. We show the comparison of our proposed methods with other state-of-the-art passive single image depth estimation results [29]–[35] in Table IV. We use the standard performance metrics used by all the aforementioned works for comparison, including linear root mean square error (RMS), absolution relative error (REL), logarithm-scale root mean square error (Log10) and depth estimation accuracy within a threshold margin ($\delta$ within $1.25, 1.25^2$ and $1.25^3$ away from the ground truth). We refer the readers to [29] for the detailed definitions of the metrics. As one can see, we achieve better performance in every metrics category for depth estimation error and accuracy, which suggests that the added end-to-end optimized phase mask does help improve the depth estimation. Moreover, we don't have the issue of scaling ambiguity in depth like those semantics based single-image depth estimation methods since our PSFs are based on absolute depth values.

## V. EXPERIMENTS ON REAL HARDWARE

We fabricate the phase masks learned through our end-to-end optimization, and evaluated its performance on a range of real-world scenes. The experiment details are discussed below, and the qualitative results are shown in Figure 11.

### A. Experiment Setup

In the experiment, we use a Yongnuo 50mm $f/1.8$ standard prime lens, which is easy to access the aperture plane. The

TABLE IV
COMPARISON WITH SEMANTICS-BASED SINGLE IMAGE DEPTH ESTIMATION METHODS ON NYU DEPTH V2 DATASETS.

| Method | Error | | | Accuracy, $\delta <$ | | |
|---|---|---|---|---|---|---|
| | RMS | REL | Log10 | 1.25 | $1.25^2$ | $1.25^3$ |
| Make3D [29] | 1.214 | 0.349 | | 0.447 | 0.745 | 0.897 |
| Eigen [29] | 0.907 | 0.215 | - | 0.611 | 0.887 | 0.971 |
| Liu [30] | 0.824 | 0.23 | 0.095 | 0.614 | 0.883 | 0.971 |
| Cao [32] | 0.819 | 0.232 | 0.091 | 0.646 | 0.892 | 0.968 |
| Chakrabarti [31] | 0.620 | 0.149 | - | 0.806 | 0.958 | 0.987 |
| Qi [33] | 0.569 | 0.128 | 0.057 | 0.834 | 0.96 | 0.99 |
| Laina [34] | 0.573 | 0.127 | 0.055 | 0.811 | 0.953 | 0.988 |
| Hu [35] | 0.530 | 0.115 | **0.050** | 0.866 | 0.975 | 0.993 |
| **Ours** | **0.382** | **0.093** | **0.050** | **0.932** | **0.989** | **0.997** |

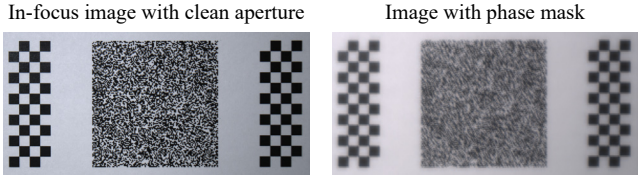


Fig. 8. **Calibration target for PSF estimation.** An example of a sharp image (left) taken using a camera lens without the phase mask and a coded image (right) taken through the phase mask. The checkerboard pattern around the calibration target is used for the alignment of the image pairs.

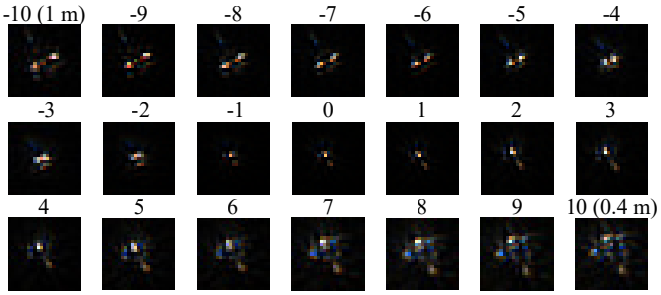


Fig. 9. **Calibrated PSFs of the fabricated phase mask.** The camera lens with the phase mask in its aperture is calibrated for depths 0.4 m to 1 m, which corresponds to the normalized $W_m$ range for an aperture size of 2.835 mm.

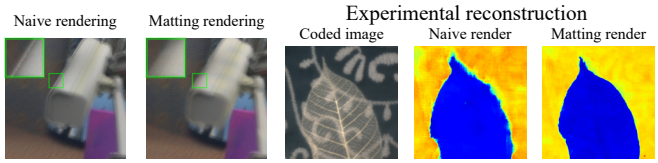


Fig. 10. **Fine-tune digital network with matting-based rendering.** (Left) Example comparison between naive rendering and matting-based rendering. Without blending between the depth layers, the naive rendering show artifacts on depth boundaries as shown in the insets. The matting-based rendering is more realistic throughout the image. (Right) Improvement in depth estimation of real experimental data is observed when the digital network is fine-tuned with matting-based rendered training data. The improvement is visible along the edges of the leaf.

sensor is a $5472 \times 3648$ machine vision color camera (BFS-PGE-200S6C-C) with $2.4$ $\mu$m pixel size. We set the diameter of the mask phase to be 2.835 mm. Thus, the simulated pixel size is about $9.4$ $\mu$m for the green channel, which corresponds to 4 pixels in our actual camera. For each $4 \times 4$ region, we group it to be one pixel with RGB channels by averaging each color channel based on the Bayer pattern, therefore the final output resolution of our system is $1344 \times 894$.

### *B. Phase Mask Fabrication*

The size of the designed phase mask is $21 \times 21$, with each grid corresponding to a size of $135$ $\mu$m $\times$ $135$ $\mu$m. The full size of the phase mask is 2.835 mm $\times$ 2.835 mm.

The phase mask was fabricated using two-photon lithography 3D printer (Photonic Professional *GT*, Nanoscribe GmbH [56]). For a reliable print, the height map of the designed phase mask was discretized into steps of 200 nm. The phase mask was printed on a 170 $\mu$m thick, 30 mm diameter glass substrate using Nanoscribe's IP-L 780 photoresist in a direct laser writing configuration with a 63$\times$ microscope objective lens. The glass substrate was then cut to a smaller size to fit into the camera lens' aperture. Close-up of the phase mask in the camera lens aperture is shown in Figure 2.

### *C. PSF Calibration*

Although the depth-dependant PSF response of the phase mask is known from simulation, we calibrate our prototype camera to account for any mismatch born out of physical implementation such as aberrations in fabricated phase mask and phase mask aperture alignment. We adopted an optimization-based approach where we estimate the PSFs from a set of sharp and coded image pairs [57], [58] of a calibration pattern.

Estimating the PSF can be posed as a deconvolution problem, where both a sharp image and a coded image of the same calibration target are given. The calibration target we used is a random binary pattern that was laser-printed on paper. We used two identical camera lenses, one without the phase mask to capture the sharp image and the other with the phase mask in the aperture to capture the coded image. Image pairs are then obtained for each depth plane of interest. The lens focus was adjusted at every depth plane to capture sharp images while the focus of the camera lens with the phase mask was kept fixed. Checkerboard pattern was used around the calibration pattern to assist in correcting for any misalignment between the sharp and the coded image.

For a particular depth plane, let $\mathbf{I}$ be the sharp image and $\mathbf{J}$ be the coded image taken using the phase mask. We can estimate the PSF $\mathbf{p}_{opt}$ by solving the following convex optimization problem

$$\mathbf{p}_{opt} = \underset{\mathbf{p}}{\operatorname{argmin}} \left\| \mathbf{I} * \mathbf{p} - s \cdot \mathbf{J} \right\|_2^2 + \lambda \left\| \nabla \mathbf{p} \right\|_1 + \mu \left\| \mathbf{1}^T \mathbf{p} - 1 \right\|_2^2 \tag{15}$$

where the first term is a least-squares data fitting term ('$*$' denotes convolution), and the scalar $s = \sum_{m,n} \mathbf{I}(m,n) / \sum_{m,n} \mathbf{J}(m,n)$ normalizes the difference in exposure between the image pairs. The second term constraints the gradients of the PSF to be sparse and the third

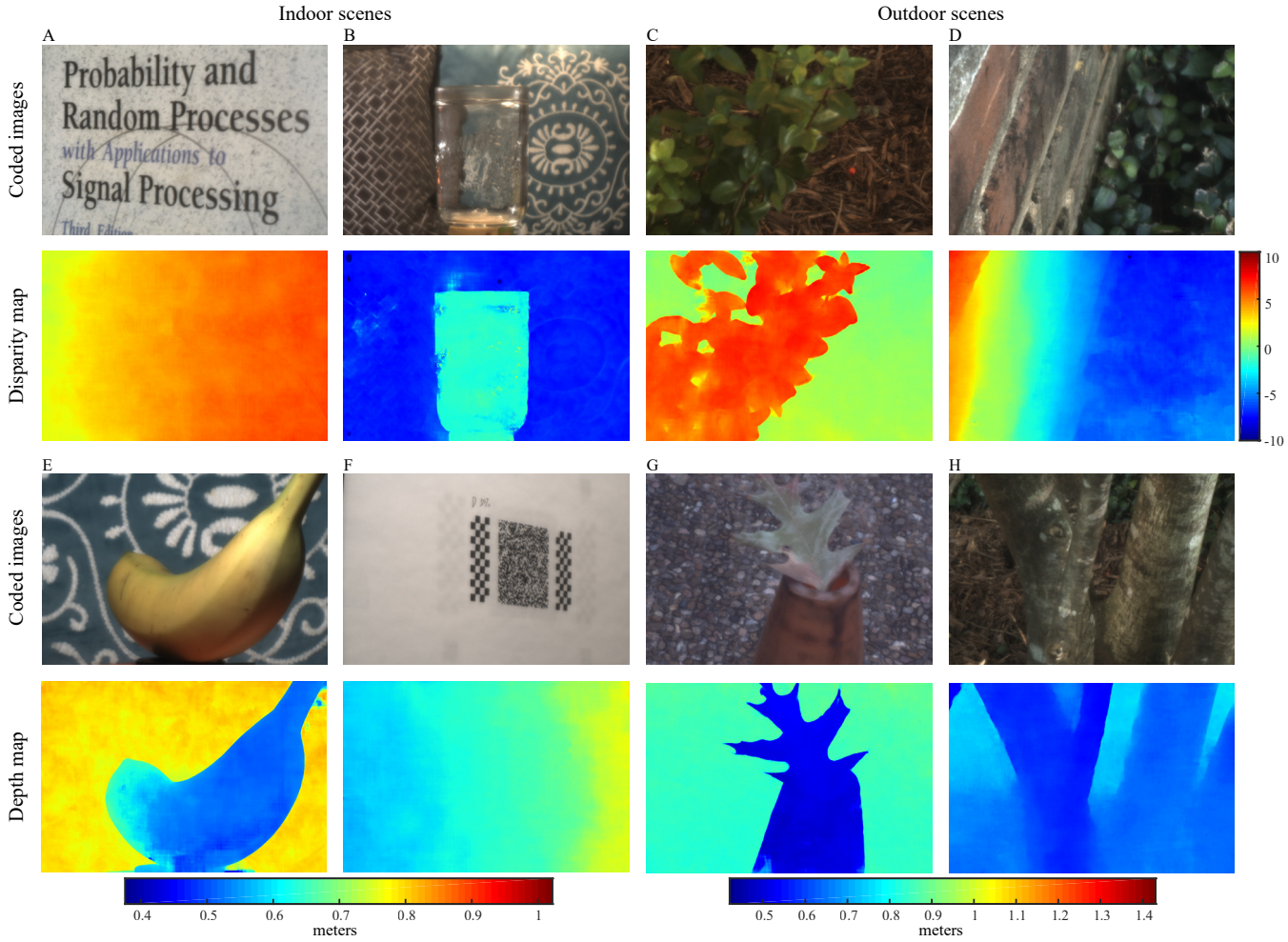


Fig. 11. **Real-world results.** Results of various scenario are shown and compared: Indoor scenes (A, B, E, and F) are shown on the left and outdoor scenes (C, D, G, and H) are on the right; Smoothly changing surfaces are presented in (A, D and F) and sharp object boundaries in (B, C, E, G, and H); Special cases of a transparent object (B) and texture-less areas (E and F) are also included.

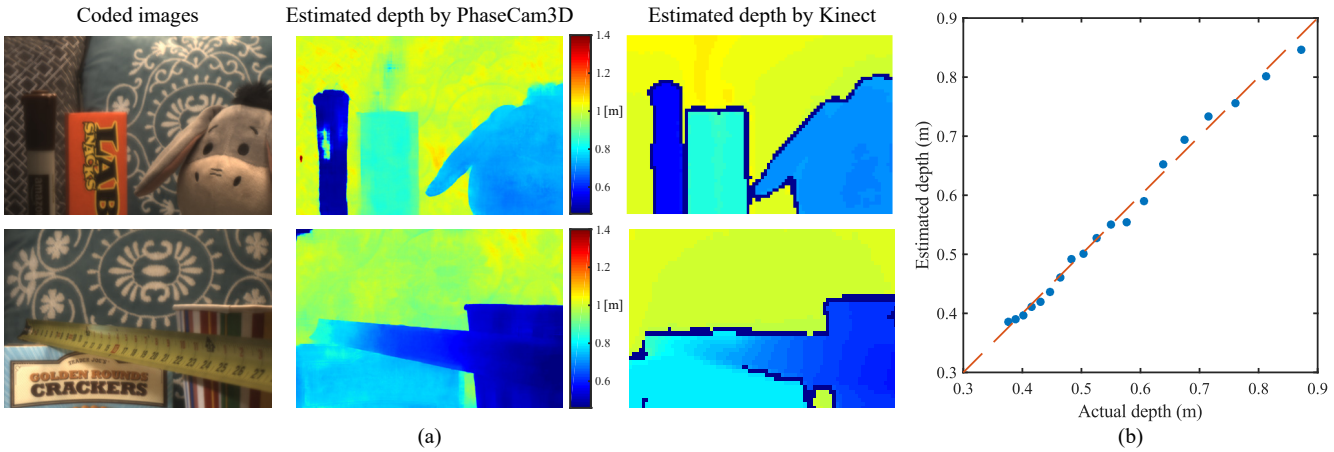


Fig. 12. **Validation experiments. (a)** Comparison with the Microsoft Kinect V2. **(b)** Depth accuracy evaluation of PhaseCam3D by capturing targets at known depths. The actual depth is measured by a tape measure.

term enforces an energy conservation constraint. The above optimization problem can be solved using first-order primal-dual algorithm presented in [58], [59]. The PSF estimation is performed for each color channel and each depth plane independently.

### D. Fine-tuning the Digital Network

When training for phase mask profile using our framework, we used naive rendering to simulate the coded image as described in Section III-A(c). Such a rendering process is fast, allowing for multiple cycles of rendering and sufficient to explain most out-of-focus regions of the scene. However, without blending between the depth layers, the naive rendering is not realistic at depth boundaries. Hence, the digital reconstruction network trained using naive rendering shows artifacts at object boundaries as shown in Figure 10.

To improve the performance of the depth reconstruction network, we fix the optimized phase mask and retrain the digital network with a matting-based rendering technique [60]. Matting for each depth layer was computed by convolving the corresponding PSF with the depth layer mask. The coded image was then composited, ordered from farther blurred layers to nearer blurred layers. The layers were linearly blended using the normalized matting weights [61]. Since the PSFs are fixed, rendering of all the coded imaged can be created apriori and fed into the training of the depth reconstruction network. The use of closer-to-reality matting-based rendering improved our experimental reconstructions significantly at the object boundaries, as shown in Figure 10.

### E. Real-world Results

Using the hardware prototype we built, we acquire the depth of the real world scenes. We show the results in Figure 11. As one can observe, our system is robust to lighting condition as reasonable depth estimation for both indoor scenes (A, B, E, and F) and outdoor scene (C, D, G, and H) are produced. Both smoothly changing surface (A, D and F) and sharp object boundaries (B, C, E, G, and H) are nicely portrayed. Special cases of a transparent object (B) and texture-less areas (E and F) are also nicely handled.

In addition, given the Microsoft Kinect V2 [15] is the one of the best ToF-based depth camera available on the mainstream market, we show our depth estimation results against the Kinect results in Figure 12(a). As one can see, the Kinect indeed output smoother depth on flat surfaces than our system, however, our method handles the depth near the object boundary better than Kinect.

To validate the depth-reconstruction accuracy of our prototype, we captured a planar target placed at various known depths. We compute the depth of the target and then compare against the known depths. As shown in Figure 12(b), we reliably estimate the depth throughout the entire range.

For comparison, we also tested the Fisher mask in experiments. The results show that our proposed mask provides better depth estimation. Detailed description can be found in the supplementary material.

## VI. Conclusion

In this work, we apply phase mask to the aperture plane of a camera to help estimate the depth of the scene and use a novel end-to-end approach to design the phase mask and the reconstruction algorithm jointly. In our end-to-end framework, we model the optics as learnable neural network layers and connected them to the consequent reconstruction layers for depth estimation. As a result, we are able to use back-propagation to optimize the reconstruction layers and the optics layers end-to-end. Compared to existing depth estimation methods, such as stereo vision and ToF sensors, our phase mask-based approach uses only single-shot, single-viewpoint and requires no specialty light source, making it easy to set up, suitable for dynamic scenes, consumes less energy and robust to any lighting condition. Following our proposed framework, we build a prototype depth estimation camera using the end-to-end optimized phase mask and reconstruction network. The fabrication of the phase mask is low cost and can be easily scaled up for mass production. Looking into the future, we hope to extend our framework to more applications, such as microscopy. We also are interested in modeling other components in the imaging system (i.e. ISP pipeline, lenses, and spectral filters) in our end-to-end framework, so as to aim for a more completely optimized the camera for higher-level computer vision tasks.

## Acknowledgment

This work was supported in part by NSF grants IIS-1652633, IIS-1618823, CCF-1527501, CCF-1730574, CCF-1652569 and DARPA NESD program HR0011-17-C-0026. Y. W. was partially supported by Information Technology Oil & Gas HPC Conference Graduate Fellowship from the Ken Kennedy Institute.

## References

[1] 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, p. 70, 2007.

[2] 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.

[3] 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.

[4] H. Haim, S. Elmalem, R. Giryes, A. 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.

[5] A. Chakrabarti, "Learning sensor multiplexing design through back-propagation," in *Advances in Neural Information Processing Systems*, 2016.

[6] 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 Fraphics (TOG)*, vol. 37, no. 4, pp. 1–13, 2018.

[7] D. Gabor, "A new microscopic principle," 1948.

[8] Y. N. Denisyuk, "On the reflection of optical properties of an object in a wave field of light scattered by it," *Doklady Akademii Nauk SSSR*, vol. 144, no. 6, pp. 1275–1278, 1962.

[9] E. N. Leith and J. Upatnieks, "Reconstructed wavefronts and communication theory," *Journal of the Optical Society of America A (JOSA)*, vol. 52, no. 10, pp. 1123–1130, 1962.

[10] "Holokit hologram kits," https://www.integraf.com/shop/hologram-kits.
[11] "Liti holographics litiholo kits," https://www.litiholo.com/.
[12] T. Tahara, X. Quan, R. Otani, Y. Takaki, and O. Matoba, "Digital holography and its multidimensional imaging applications: a review," *Microscopy*, vol. 67, no. 2, pp. 55–67, 2018.
[13] J. Geng, "Structured-light 3d surface imaging: a tutorial," *Advances in Optics and Photonics*, vol. 3, no. 2, pp. 128–160, 2011.
[14] S. Foix, G. Alenya, and C. Torras, "Lock-in time-of-flight (tof) cameras: A survey," *IEEE Sensors Journal*, vol. 11, no. 9, pp. 1917–1926, 2011.
[15] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE Multimedia*, vol. 19, no. 2, pp. 4–10, 2012.
[16] M. Gupta, A. Agrawal, A. Veeraraghavan, and S. G. Narasimhan, "Structured light 3d scanning in the presence of global illumination," in *CVPR*, 2011.
[17] N. Matsuda, O. Cossairt, and M. Gupta, "Mc3d: Motion contrast 3d scanning," in *IEEE International Conference on Computational Photography (ICCP)*, 2015.
[18] S. Achar, J. R. Bartels, W. L. Whittaker, K. N. Kutulakos, and S. G. Narasimhan, "Epipolar time-of-flight imaging," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, p. 37, 2017.
[19] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge university press, 2003.
[20] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *German Conference on Pattern Recognition*, 2014.
[21] "Light l16 camera," https://www.light.co/camera.
[22] G. Neukum, R. Jaumann, H. Hoffmann, E. Hauber, J. Head, A. Basilevsky, B. Ivanov, S. Werner, S. Van Gasselt, J. Murray *et al.*, "Recent and episodic volcanic and glacial activity on mars revealed by the high resolution stereo camera," *Nature*, vol. 432, no. 7020, p. 971, 2004.
[23] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *International Journal of Computer Vision*, vol. 9, no. 2, pp. 137–154, 1992.
[24] J. Sun, Y. Li, S. B. Kang, and H.-Y. Shum, "Symmetric stereo matching for occlusion handling," in *CVPR*, 2005.
[25] C. L. Zitnick and T. Kanade, "A cooperative algorithm for stereo matching and occlusion detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 7, pp. 675–684, 2000.
[26] A. F. Bobick and S. S. Intille, "Large occlusion stereo," *International Journal of Computer Vision*, vol. 33, no. 3, pp. 181–200, 1999.
[27] Q. Yang, C. Engels, and A. Akbarzadeh, "Near real-time stereo for weakly-textured scenes." in *British Machine Vision Conference*, 2008.
[28] K. Konolige, "Projected texture stereo," in *Robotics and Automation*, 2010.
[29] 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.
[30] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *CVPR*, 2015.
[31] A. Chakrabarti, J. Shao, and G. Shakhnarovich, "Depth from a single image by harmonizing overcomplete local network predictions," in *Advances in Neural Information Processing Systems*, 2016.
[32] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3174–3182, 2018.
[33] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "Geonet: Geometric neural network for joint depth and surface normal estimation," in *CVPR*, 2018.
[34] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *3D Vision (3DV)*, 2016.
[35] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," *arXiv:1803.08673*, 2018.
[36] R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*, 2016.
[37] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, 2017.
[38] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *CVPR*, 2017.
[39] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "Demon: Depth and motion network for learning monocular stereo," in *CVPR*, 2017.
[40] 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 *CVPR*, 2016.
[41] 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.
[42] J. W. Goodman, *Introduction to Fourier Optics*. Roberts and Company Publishers, 2005.
[43] T. Gissibl, S. Wagner, J. Sykora, M. Schmid, and H. Giessen, "Refractive index measurements of photo-resists for three-dimensional direct laser writing," *Optical Materials Express*, vol. 7, no. 7, pp. 2293–2298, 2017.
[44] B. Barsky and T. J. Kosloff, "Algorithms for Rendering Depth of Field Effects in Computer Graphics," *World Scientific and Engineering Academy and Society (WSEAS)*, pp. 999–1010, 2008.
[45] C. Scofield, "212-d depth-of-field simulation for computer animation," in *Graphics Gems III (IBM Version)*, 1992.
[46] M. Kraus and M. Strengert, "Depth-of-field rendering by pyramidal image processing," *Computer Graphics Forum*, vol. 26, no. 3, pp. 645–654, 2007.
[47] 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*, 2015.
[48] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv:1502.03167*, 2015.
[49] Y. Shechtman, S. J. Sahl, A. S. Backer, and W. E. Moerner, "Optimal point spread function design for 3D imaging," *Physical Review Letters*, vol. 113, no. 3, pp. 1–5, 2014.
[50] M. Born and E. Wolf, *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*. Elsevier, 2013.
[51] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European Conference on Computer Vision*, 2012.
[52] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *Symposium on Operating Systems Design and Implementation (OSDI)*, 2016.
[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.
[54] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *the International Conference on Artificial Intelligence and Statistics*, 2010.
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *International Conference on Computer Vision*, 2015.
[56] "Nanoscribe gmbh," https://www.nanoscribe.de/.
[57] L. Yuan, J. Sun, L. Quan, and H.-Y. Shum, "Image deblurring with blurred/noisy image pairs," *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3, p. 1, 2007.
[58] F. Heide, M. Rouf, M. B. Hullin, B. Labitzke, W. Heidrich, and A. Kolb, "High-quality computational imaging through simple lenses," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 5, pp. 1–14, 2013.
[59] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.
[60] M. Kraus and M. Strengert, "Depth-of-field rendering by pyramidal image processing," in *Computer Graphics Forum*, 2007.
[61] S. Lee, G. J. Kim, and S. Choi, "Real-time depth-of-field rendering using point splatting on per-pixel layers," in *Computer Graphics Forum*, 2008.