TiDy-PSFs: Computational Imaging with Time-Averaged Dynamic Point-Spread-Functions

Sachin Shah[*]
University of Maryland, College Park
shah2022@umd.edu

Sakshum Kulshrestha[*]
University of Maryland, College Park
sakshumk@umd.edu

Christopher A. Metzler
University of Maryland, College Park
metzler@umd.edu

Abstract

Point-spread-function (PSF) engineering is a powerful computational imaging technique wherein a custom phase mask is integrated into an optical system to encode additional information into captured images. Used in combination with deep learning, such systems now offer state-of-the-art performance at monocular depth estimation, extended depth-of-field imaging, lensless imaging, and other tasks. Inspired by recent advances in spatial light modulator (SLM) technology, this paper answers a natural question: Can one encode additional information and achieve superior performance by changing a phase mask dynamically over time? We first prove that the set of PSFs described by static phase masks is non-convex and that, as a result, time-averaged PSFs generated by dynamic phase masks are fundamentally more expressive. We then demonstrate, in simulation, that time-averaged dynamic (TiDy) phase masks can leverage this increased expressiveness to offer substantially improved monocular depth estimation and extended depth-of-field imaging performance.

1. Introduction

Extracting depth information from an image is a critical task across a range of applications including autonomous driving [29, 33], robotics [23, 34], microscopy [7, 19], and augmented reality [31, 14]. To this end, researchers have developed engineered phase masks and apertures which serve to encode depth information into an image [12, 25]. To optimize these phase masks, recent works have exploited deep learning: By simultaneously optimizing a phase mask and a reconstruction algorithm "end-to-end learning" is able to dramatically improve system performance [32, 26].

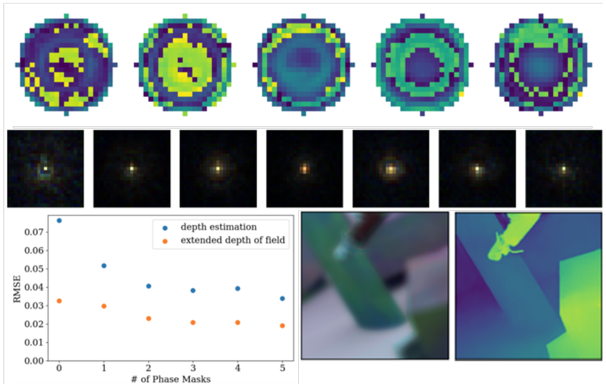

Figure 1. **Time-averaged Dynamic PSFs** Top: Phase mask sequence that was optimized to perform simultaneous extended depth-of-field imaging and monocular depth estimation. Middle: Proposed TiDy PSFs at specific depths. Bottom left: Depth estimation and all-in-focus imaging performance improve as one averages over more phase masks. Bottom right: Depth-encoded image and reconstructed depth map.

Most existing works have focused on learning or optimizing a single phase mask for passive depth perception. We conjecture that this restriction leaves much room for improvement. Perhaps by using an SLM to introduce a sequence of phase masks over time, one could do much better.

Supporting this idea is the fact, which we prove in Theorem 3, that the set of PSFs described by a single phase mask is non-convex. This implies that time-averaged PSFs, which span the convex hull of this set, can be significantly more expressive. In this work, we exploit the PSF non-convexity by developing a multi-phase mask end-to-end optimization approach for learning a sequence of phase masks whose PSFs are averaged over time.

This work's central contributions are as follows:

- We prove the set of PSFs generated by a single phase

[*] These authors contributed equally to this work

mask is non-convex. Thus, dynamic phase-masks offer a fundamentally larger design space.

- We extend the end-to-end learning optics and algorithm design framework to design a dynamic set of phase masks.
- We demonstrate, in simulation, that time-averaged PSFs can achieve superior monocular depth estimation and extended depth-of-field imaging performance.

2. Background

Image Formation Model. One can simulate the formation of an image in a camera by discretizing an RGB image by depth, convolving each depth with it's the corresponding PSF, and compositing the outputs to form the signal on the sensor. This process can be represented by the equation

$$I = \sum_{d=1}^{D} O_d \left(L * h_d \right), \tag{1}$$

where L represents all-in-focus image, $\{1, \cdots, D\}$ represent a set of discrete depth layers, O_d is the occlusion mask at depth d, and the set $\{h_1, \cdots, h_D\}$ represent the depth-dependent PSF, i.e., the cameras response to point sources at various depths [9]. Other works assume no depth discontinuities [26] or add additional computation to improve blurring at depth boundaries [10]. Our model is similar to those used in [32, 3].

PSF Formation Model. A PSF h_d can be formed as a function of distance d and phase modulation ϕ^M caused by height variation on a phase mask.

$$h_d = |\mathcal{F}[A \exp(i\phi^{DF}(d) + i\phi^M)]|^2 \tag{2}$$

where $\phi^{DF}(d)$ is the defocus aberration due to the distance d between the focus point and the depth plane. Note that because this PSF depends on depth, it can be used to encode depth information into I [8].

The key idea behind PSF-engineering and end-to-end learning is that one can use the aforementioned relationships to encode additional information into a captured image I by selecting a particularly effective mask ϕ^M.

3. Related Work

3.1. Computational Optics for Depth Tasks

Optics based approaches for depth estimation use sensors and optical setups to encode and recover depth information. Many methods have used the depth-dependent blur induced by the imaging system to estimate the depth of pixels in an image [20]. These approaches compare the blur at different ranges to the expected blur caused by an aperture focused at a fixed distance. Groups improved on this idea by implementing coded apertures, retaining more high frequency information about the scene to disambiguate depths [12]. Similar to depth estimation tasks, static phase masks have been used to produce tailored PSFs more *invariant* to depth, allowing for extended depth-of-field imaging [6]. However, these optically driven approaches with numerical analysis have been passed in performance by modern deep neural networks, allowing for joint optimization of optical elements and neural reconstruction networks.

3.2. Deep Optics for Depth Tasks

Many methods have engineered phase masks with specific depth qualities. By maximizing Fisher information for depth, the coded image theoretically will have the most amount of depth cues as possible [24] and by minimizing Fisher information, one may achieve an extended depth-of-field image [6]. Deep learning techniques can be used to jointly train the optical parameters and neural network based estimation methods. The idea is that one can "code" an image to retain additional information about a scene, and then use a deep neural network to produce reconstructions. By using a differentiable model for light propagation, back-propagation can be used to update phase mask values simultaneously with neural network parameters. This approach was demonstrated for extended depth-of-field imaging [28, 26, 10, 13], depth estimation [32, 3, 10], high-dynamic-range imaging [16, 27], and holography [5, 4]. While these previous approaches successfully improved performance, they focused on enhancing a single phase mask. We build on these works by simultaneously optimizing multiple phase masks, which allows us to search over a larger space of PSFs.

4. Theory

Micro-electromechanical system (MEMS) based SLMs offer high framerates but have limited phase precision due to heavy quantization [1]. As [4] noted, intensity averaging of multiple frames can improve quality by increasing effective precision to overcome quantization. Our key insight is that even as SLM technology improves, intensity averaging yields a more expressive design space than a single phase mask. This is supported by the claim that the set of PSFs that can be generated by a single phase mask is non-convex. We provide a rigorous proof for the claim as follows.

Theorem 1. *The set of PSFs that can be generated by a phase mask with an infinite aperture ($A(x) = 1$ for all $x \in \mathbb{R}^2$) is non-convex.*

Proof. Consider two tilt shift phase masks, $M_1(x) = ax$ and $M_2(x) = -ax$, with respect to a coordinate x with non-zero a. The corresponding averaged PSF is $\frac{1}{2}\delta(a) + \frac{1}{2}\delta(-a)$, where δ denotes a Dirac delta function. To realize this PSF with a single phase mask we would need the field at the aperture to satisfy

$$|\mathcal{F}(E)|^2 = \frac{1}{2}\delta(a) + \frac{1}{2}\delta(-a). \tag{3}$$

This implies

$$\begin{aligned} E(x) &= \frac{1}{\sqrt{2}}e^{-iax}e^{i\gamma_1} + \frac{1}{\sqrt{2}}e^{iax}e^{i\gamma_2}, \qquad (4)\\ &= \frac{2}{\sqrt{2}}e^{i\frac{\gamma_1+\gamma_2}{2}}\cos\left(ax + \frac{\gamma_1-\gamma_2}{2}\right) \qquad (5)\end{aligned}$$

for some $\gamma_1, \gamma_2 \in \mathbb{R}$. $E(x)$'s amplitude varies according to a cosine and thus cannot be realized by a phase-only mask. □

We now demonstrate the non-convexity claim holds for finite apertures that lie on a discrete grid.

Definition 1. *$A \in \{0,1\}^{N\times N}$ is some valid aperture with a non-zero region S such that there exists lines L_1 and L_2 where S can be contained between them, and $L_1 \parallel L_2$ and $u = S \cap L_1$ and $v = S \cap L_2$ are single points (Figure 2).*

This definition of A supports most commonly used apertures including but not limited to circles, squares, and n-sided regular polygons. See supplement for proof for all shapes.

Definition 2. *Let $T_A(N)$ be the set of $N \times N$ matrices in $\mathbb{T}^{N\times N}$ with non-zero support A, i.e. the matrix is supported only where $A = 1$, where $\mathbb{T}$ is the complex unit circle.*

The PSF induced by a phase mask M can be modeled as the squared magnitude of the Fourier transform of the pupil function f [32].

Definition 3. *Let $f : \mathbb{R}^{N\times N} \to T_A(N)$ be defined by*

$$f(M) = A \odot \exp(iD + icM) \tag{6}$$

where $\odot$ denotes entry-wise multiplication, and $D \in \mathbb{R}^{N\times N}$ and $c \in \mathbb{R} - \{0\}$ (the reals except for 0) are fixed constants.

Definition 4. *Let $g : T_A(N) \to \mathbb{R}^{N\times N}$ be defined by*

$$g(X) = \frac{|\mathcal{F}(X)| \odot |\mathcal{F}(X)|}{\|\mathcal{F}(X)\|_F^2} \tag{7}$$

where $\mathcal{F}$ denotes the discrete Fourier Transform with sufficient zero-padding, $|\cdot|$ denotes entry-wise absolute value, and $\|\cdot\|_F$ denotes the Frobenius norm.

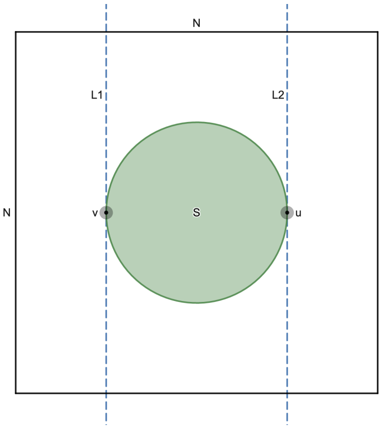

Figure 2. **Example aperture that satisfies constraints on A.** The aperture is fitted between parallel lines $L1$ and $L2$, which only intersect the aperture at one point each. Common aperture shapes fit into these constraints.

Lemma 2. *From fourier optics theory [8], any single phase mask's PSF at a specific depth can be written as*

$$PSF = g \circ f.$$

Theorem 3. *The range of PSF is not a convex set.*

Proof. f is clearly surjective, so it suffices to argue the range of g is not convex. Assume by way of contradiction that the range of g is convex. Then, for all $X^{(1)}, \ldots, X^{(k)} \in T_A(N)$ there exists $Y \in T_A(N)$ such that $g(Y) = \frac{1}{k}\sum_{i=1}^{k} g(X^{(i)})$. By Parseval's Theorem,

$$\|\mathcal{F}(X)\|_F^2 = N^2\|X\|_F^2 = N^2\sum_{i=0}^{N}\sum_{j=0}^{N} A_{i,j} \tag{8}$$

so the condition is

$$|\mathcal{F}(Y)| \odot |\mathcal{F}(Y)| = \frac{1}{k}\sum_{i=1}^{k} |\mathcal{F}(X^{(i)})| \odot |\mathcal{F}(X^{(i)})| \tag{9}$$

or equivalently

$$\mathcal{F}(Y) \odot \overline{\mathcal{F}(Y)} = \frac{1}{k}\sum_{i=1}^{k} \mathcal{F}(X^{(i)}) \odot \overline{\mathcal{F}(X^{(i)})}. \tag{10}$$

Then the cross-correlation theorem reduces it to

$$\mathcal{F}(Y \star Y) = \frac{1}{k}\sum_{i=1}^{k} \mathcal{F}(X^{(i)} \star X^{(i)}) \tag{11}$$

where $\star$ denotes cross-correlation. Because the Fourier Transform is linear we finally have

$$Y \star Y = \frac{1}{k}\sum_{i=1}^{k} X^{(i)} \star X^{(i)}. \tag{12}$$

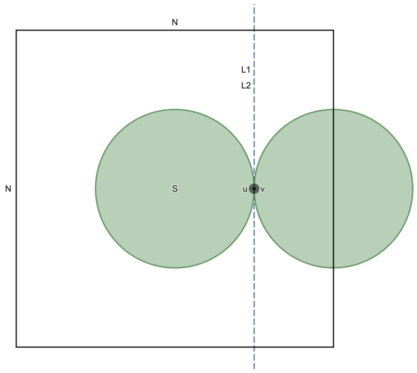

Figure 3. **Geometric interpretation of correlation $(\mathbf{X} \star \mathbf{X})_{\mathbf{v}-\mathbf{u}}$.** The figure represents the correlation step when the shift is $v - u$. Notice that only u and v overlap once the shift is applied.

Therefore, the convexity of the range of g is equivalent to the convexity of the set $\{X \star X : X \in T_A(N)\}$. We will show the set's projection onto a particular coordinate is not convex.

$$(X \star X)_{s,r} = \sum_{i=0}^{N} \sum_{j=0}^{N} X_{i,j} \overline{X_{i+s,j+r}} \tag{13}$$

where we adopt the convention that $X_{s,r} = 0$ when $s, r > N$ or $s, r < 0$. Take the points u and v from the definition of A (1). Also observe that correlation can be represented geometrically as shifting $\overline{X}$ over X. In this representation, notice that as the shift (s, r) approaches $v - u$, the non-zero overlap between X and $\overline{X}$ shifted by (s, r) approaches 1 by construction. That is, when L_1 is shifted to overlap L_2, u and v will be the only non-zero overlaps between the shifted and original non-zero points (Figure 3). No other non-zero points can overlap above or below L_2 by definition of S. Therefore, $(X \star X)_{v-u}$ becomes

$$X_u \overline{X_v} + \sum_{i=1}^{N^2-1} 0. \tag{14}$$

Because $X_u \overline{X_v} \in \mathbb{T}$, $(X \star X)_{v-u} \in \mathbb{T}$ which is a non-convex set. Therefore, the set of correlation's of values on the complex unit circle masked by A is also not convex, and so is PSF. □

Time-averaged PSFs span the convex hull of the set of static-mask PSFs, meaning there exists some PSFs achievable only through intensity averaging PSFs from a sequence of phase masks.

5. Multi-Phase Mask Optimization

5.1. Optical Forward Model

Similar to PhaseCam3D [32], we model light propagation using Fourier optics theory [8]. In contrast to previous work, we compute the forward model (1) for multiple phase masks, producing a stack of output images, which form our coded image when averaged. This coded image simulates the recorded signal from imaging a scene using a sequence of phase masks in a single exposure (Figure 4).

5.2. Specialized Networks

For the monocular depth estimation task, we use the MiDaS Small network [22]. This is a well known convolutional monocular depth estimation network designed to take in natural images and output relative depth maps. The network is trained end-to-end with the phase masks. A mean-squared error (MSE) loss term is defined in terms of the depth reconstruction prediction, $\hat{D}$ and the ground truth depth map D,

$$L_{Depth} = \frac{1}{N} \|D - \hat{D}\|_2^2 \tag{15}$$

where N is the number of pixels. This process allows for the simultaneous optimization of the phase masks as well as fine tuning MiDaS to reconstruct from our coded images.

For the extended depth-of-field task, we use an Attention U-Net [18] to reconstruct all-in-focus images. The network is optimized jointly with the phase mask sequence. To learn a reconstruction $\hat{I}$ to be similar to the all-in-focus ground truth image I, we define the loss term using MSE error

$$L_{AiF} = \frac{1}{N} \|I - \hat{I}\|_2^2 \tag{16}$$

where N is the number of pixels.

5.3. Joint Task Optimization

We also present an alternative to the specialized networks: a single network jointly trained for monocular depth estimation and extended depth-of-field using a sequence of phase masks. This network has a basic Attention U-Net architecture outputting 4 channels representing depth maps as well as all-in-focus images. Similar to prior works, we use a combined loss function, adding a coefficient to weight the losses for each individual task:

$$L_{total} = \lambda_{Depth} L_{Depth} + \lambda_{AiF} L_{AiF}. \tag{17}$$

6. Experimental Details

6.1. Training Details

We use the FlyingThings3D from Scene Flow Datasets [15], which uses synthetic data generation to obtain all-in-focus RGB images and disparity maps. We use the cropped 278 × 278 all-in-focus images from [32]. In total, we use 5077 training patches and 419 test patches.

Both the optical layer and reconstruction networks are differentiable, so the phase mask sequence and neural network can be optimized through back-propagation. Each

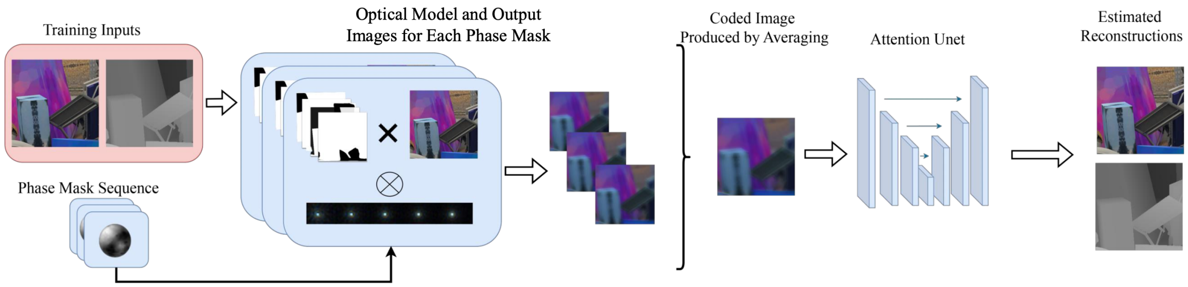
Training Inputs
Optical Model and Output Images for Each Phase Mask
Coded Image Produced by Averaging
Attention Unet
Estimated Reconstructions
Phase Mask Sequence
×
⊗

Figure 4. **Multi-phase mask forward model overview.** A sequence of phase masks are used to generate a sequence of depth-dependent PSFs. These PSFs are convolved with depth masked clean images to simulate depth dependent convolution. The images produced by each phase mask are averaged to create a coded image which is fed into an attention U-Net. The reconstruction loss is back-propagated end-to-end through the network and the optical model to design phase masks and algorithms capable of performing monocular depth estimation and extended depth-of-field simultaneously.

part is implemented in PyTorch. During training, we use the Adam [11] optimizer with parameters $\beta_1 = 0.99$ and $\beta_2 = 0.999$. The learning rate for the phase masks is 10^{-8} and for the reconstruction network it is 10^{-4}, and the batch size was 32. Finally, training and testing were performed on NVIDIA Quadro P6000 GPUs.

We parameterize 23×23 phase masks pixel-wise as [13] found pixel-wise parameterization to produce the best overall performance. The monocular depth estimation task uses the MiDaS Small architecture pretrained weights for monocular depth estimation downloadable from PyTorch [22]. The extended depth-of-field task pretrains an Attention U-Net with a fixed Fresnel lens for 300 epochs. For the joint task, we set $\lambda_{Depth} = \lambda_{AiF} = 1$ to balance overall performance, and we pretrain the Attention U-Net for 300 epochs with a fixed Fresnel lens. In simulation, the red, blue, and green channels are approximated by discretized wavelengths, 610 nm, 530 nm, and 470 nm respectively. Additionally, the depth range is discretized into 21 bins on the interval $[-20, 20]$, which is larger than previous works.

6.2. Evaluation Details

For ablation studies on our method, we used the testing split of the FlyingThings3D set for both monocular depth estimation and extended depth-of-field imaging [15]. For comparisons to existing work, we also tested our monocular depth estimation network on the labeled NYU Depth v2 set [17]. The ground truth depth maps were translated to layered masks for the clean images by bucketing the depth values into 21 bins, allowing us to convolve each depth in an image with the required PSF. We use root mean squared error (RMSE) between ground truth and estimated depth maps for depth estimation and RMSE between ground truth and reconstructed all-in-focus images for extended depth-of-field imaging. We also use peak signal-to-noise ratio (PSNR) and structural similarity index [30] (SSIM) for extended depth-of-field imaging.

7. Results

We compare our time averaged dynamic PSF method to the state-of-the-art methods for both extended depth-of-field imaging and monocular depth estimation. The relevant works we compare to are as follows:

1. PhaseCam3D [32] used a 23×23 phase mask based on 55 Zernike coefficients. The phase mask parameters were then end-to-end optimized with a U-Net reconstruction network to perform depth estimation.

2. Chang et al. [3] used a singlet lens introducing chromatic aberrations with radially symmetric PSFs. Similar to [32], the lens parameters were also then end-to-end optimized.

3. Ikoma et al. [10] used a radially symmetric diffractive optical element (DOE). The blurred image was preconditioned with an approximate inverse of the PSF depth dependent blur. The RGB image stack was fed into a U-Net to produce both an all-in-focus image and a depth map. The DOE and U-Net parameters were optimized in an end-to-end fashion.

4. Liu et al. [13] used various phase mask parameterizations with the same U-Net architecture as [10]. One method used pixel-wise height maps (PW) and the other introduced orbital angular momentum (OAM).

5. Sitzmann et al. [26] implements a single DOE based on Zernike coefficients, and solves the Tikhonov-regularized least-squares problem to reconstruct an all-in-focus image.

6. MiDaS [21] and ZoeDepth [2] are state of the art single shot monocular depth estimation methods with all-in-focus images as inputs.

Because both [10] and [13] simultaneously learn all-in-focus images and depth maps, when comparing against our

Method	FlyingThings3D	NYUv2
PhaseCam3D [32]	0.521	0.382
Chang et al. [3]	0.490	0.433
Ikoma et al. [10]	0.184	-
MiDaS [21]	-	0.357
ZoeDepth [2]	-	0.277
TiDy (1)	0.026	0.259
TiDy (5)	**0.019**	**0.175**

Table 1. **RMSE comparison of monocular depth estimation methods.** We present quantitative results on two datasets to compare to state of the art optical and single shot monocular depth estimation methods. Our method performs best with our 5 phase mask system achieving the lowest error on both datasets.

Method	RMSE↓	PSNR↑	SSIM↑
Liu et al. [13]	-	29.80	-
Ikoma et al. [10]	0.1327	31.88	0.905
Sitzmann et al. [26]	-	32.44	-
TiDy (1)	0.0148	37.33	0.968
TiDy (5)	**0.0092**	**41.11**	**0.989**

Table 2. **Comparison of extended depth-of-field imaging methods.** We present quantitative results on FlyingThings3D to compare to state-of-the-art. Our method performs best with our 5 phase mask system achieving the best PSNR.

specialized methods, we take their best performing weighting of each task.

Individual Tasks. For monocular depth estimation, our specialized method using a sequence of 5 phase masks trained for 300 epochs outperforms prior work on FlyingThings3D (Table 1). Additionally, our approach performs significantly better and achieves lower error than previous methods on NYUv2 without any additional fine tuning. For extended depth-of-field, our specialized method using a sequence of 5 phase masks outperforms prior work on FlyingThings3D (Table 2). This demonstrates the benefit of multi-phase mask learning on computational imaging tasks.

Multi-Objective Optimization. We also evaluate our method against other joint all-in-focus and depth map learning approaches. This problem is challenging because good depth cues to produce depth maps is antithetical to producing an all-in-focus image. Our combined 5 phase mask trained for 300 epochs approach outperforms prior jointly trained approaches (Table 3).

Method	All-in-focus PSNR↑	Depth RMSE↓
Ikoma et al. [10]	31.88	0.191
Liu et al. [13] - PW	29.80	0.056
Liu et al. [13] - OAM$_t$	25.86	0.053
TiDy (1)	31.20	0.052
TiDy (5)	**34.79**	**0.034**

Table 3. **Comparison of multi-objective optimization of extended depth-of-field imaging and depth estimation methods.** We compare quantitative results on FlyingThings3D to the state-of-the-art. Our method performs best with our 5 phase mask system achieving the best balance between objectives.

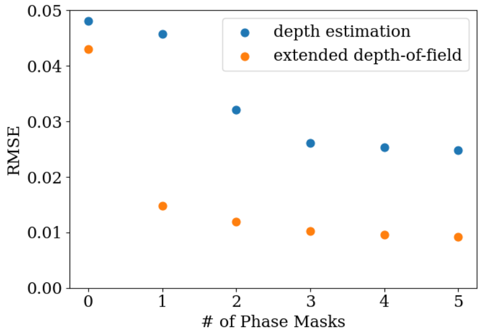

Figure 5. **RMSE for specialized tasks** for each phase mask sequence length. RMSE decreases with respect to phase mask sequence length for both specialized extended depth-of-field imaging and monocular depth estimation tasks. 0 phase masks refers to a reconstruction neural network with a fixed Fresnel lens.

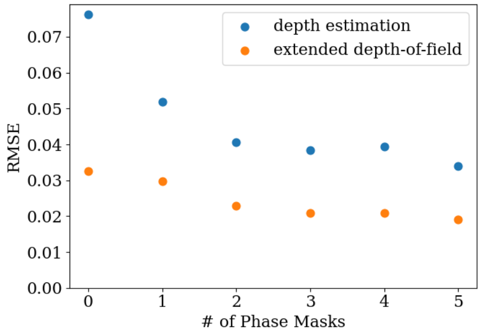

Figure 6. **RMSE for joint optimization of monocular depth estimation and extended depth-of-field imaging** for each phase mask sequence length. RMSE decreases with respect to phase mask sequence length for this complex joint task, demonstrating the benefit of multi-phase mask learning. 0 phase masks refers to a reconstruction neural network with a fixed Fresnel lens.

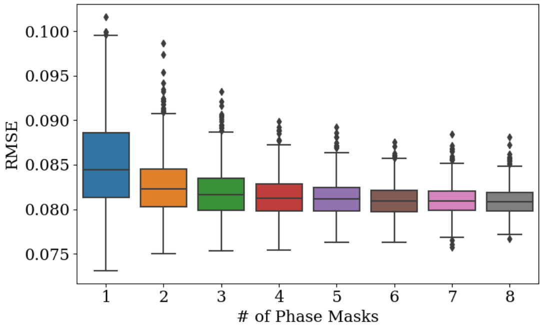

Figure 7. **All-in-focus imaging RMSE distribution** for each phase mask length without a reconstruction network. The best RMSE for each phase mask count has low correlation with respect to phase mask sequence length, but the variance of RMSE decreases.

8. Ablation Studies

8.1. Effect of Phase Mask Sequence Length

For both all-in-focus imaging and depth estimation, we vary the phase mask count that the end-to-end system is trained with to gauge the benefits of using multiple phase masks. The forward model and initial phase masks were held standard while the phase mask count was varied. The resulting networks were evaluated at convergence. For the extended depth-of-field task, the masks were all initialized with random noise uniform from 0 to 1.2×10^{-6}. For the depth estimation task, the masks were initialized with the Fisher mask with added Gaussian noise parameterized by a 5.35×10^{-7} mean and 3.05×10^{-7} standard deviation.

End-to-end optimization on each task with a specialized network yielded improved performance as the phase mask count increased, visualized in Figure 5. This result implies that sequences of phase masks are successful in making the PSF space more expressive. Additionally, even for the more complex joint task, learning a system that can produce both all-in-focus images and depth maps, error decreases with phase mask count until a plateau is reached (Figure 6).

8.2. All-in-Focus without Reconstruction Networks

A phase mask generating a PSF of the unit impulse function at every depth would be ideal for extended depth-of-field as each depth would be in focus. If possible, this phase mask would not require any digital processing. We optimize phase mask sequences of varying lengths to produce an averaged PSF close to the unit impulse function for all depths. For each sequence length, phase masks are optimized using MSE loss between the unit impulse function and the averaged PSF at each depth until convergence. We ran 1000 trials of random phase mask initialization for each length. Observe that a side-effect of longer phase masks is training stability. The range of RMSE between the simulated capture image and ground truth all-in-focus image decreases as the sequence length increases (Figure 7). This indicates training longer sequences is more resilient to initialization.

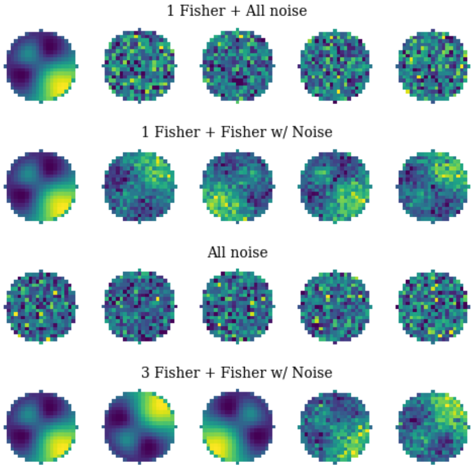

Figure 8. **Visualization of phase mask initializations.** Each row represents a different initial phase mask sequence.

8.3. Phase Mask Initialization for Depth Perception

Deep optics for depth perception can be very dependent on the initialization of optical parameters before training [32]. To find the extent of the effect of mask initialization on performance, we varied the the initial phase masks while keeping number of masks, the optical model, and duration of training fixed. We trained for 200 epochs. We tested four initializations of sequences of 5 phase masks as shown in Figure 8. The first was uniformly distributed noise from 0 to 1.2×10^{-6}. The second was the first mask in the sequence set to a Fisher mask while the rest are uniform noise. The third is setting each mask to a rotation of the Fisher mask and adding Gaussian noise parameterized by a 5.35×10^{-7} mean and 3.05×10^{-7} standard deviation to 4 masks. Lastly, we set each mask to a rotation of the Fisher mask and added noise to only the last two masks in the sequence. Of the four initializations, it is clear that the 3 Fisher masks and 2 Fisher masks with noise performed the best (Table 4).

8.4. Modeling SLM Imperfections

State Switching. Our optical forward model assumes an SLM can swap between two phase patterns instantly. In practice, however, some light will be captured during the intermediate states between phase patterns. These phase patterns, in the worst case, could be random phase patterns, effectively adding noise to our coded images. We model these intermediate states by averaging output images pro-

Initialization	RMSE↓
1 Fisher + All noise	0.0329
1 Fisher + Fisher w/ Noise	0.0271
All noise	0.0254
3 Fisher + Fisher w/ Noise	**0.0207**

Table 4. **Quantitative evaluation of phase mask initializations.** Four sequence initializations are evaluated on the monocular depth estimation task. Ultimately, 3 Fisher masks and 2 noisy Fisher masks have the best performance after training.

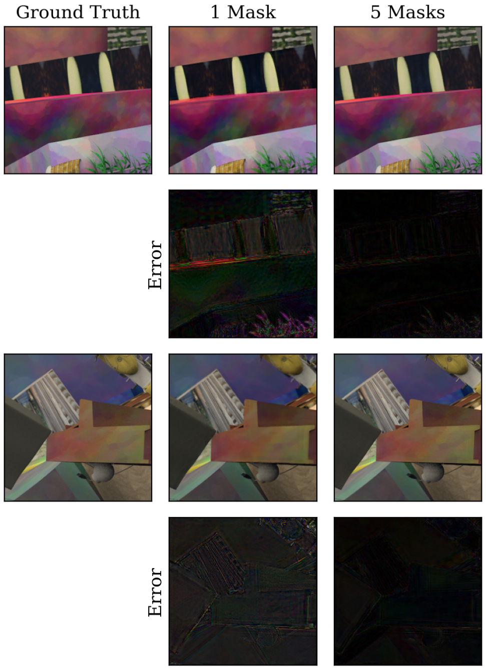

Figure 9. **Qualitative results of a specialized network on extended depth-of-field imaging.** Both 1 and 5 phase mask systems are evaluated on FlyingThings3D. Error is computed pixel wise between the ground truth all-in-focus image and the reconstructed output and is boosted by a factor of 3. Notice that the 5 phase mask system introduces minimal error.

duced by phase masks and the randomized phase patterns weighted by the time that they are displayed for. We model the total exposure time as 100ms, with various durations of switching times from 1 to 16ms per swap. We evaluate our joint optimized network on these new, more noisy, coded images without any additional training (Figure 12). Observe that because the 5 phase mask system includes more swaps, performance degrades faster than fewer phase mask systems. However, for short switching times, 5 phase masks still outperform the others without needing any fine tuning.

Quantization. Current high-speed MEMS based SLM technology experience heavy quantization [1]. We model the effects of quantization error by adding varying amounts of noise during inference. As demonstrated in Figure 13,

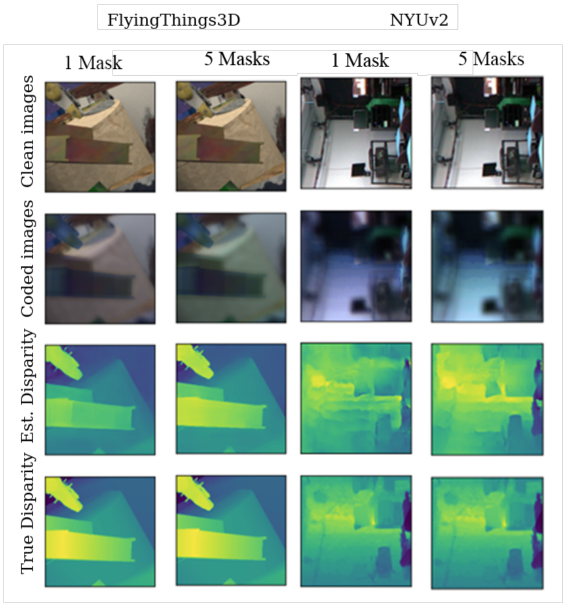

Figure 10. **Qualitative results of a specialized networks on monocular depth estimation.** Performance using the five phase mask method outperforms one phase mask on both datasets.

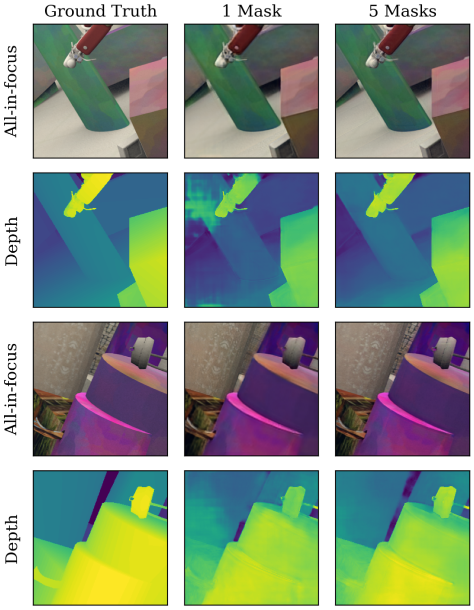

Figure 11. **Qualitative results of a joint optimized system for extended depth-of-field imagining and monocular depth estimation.** Both one and five phase mask networks are evaluated on the FlyingThings3D datasets. Notice that five masks has fewer artifacts than a single mask.

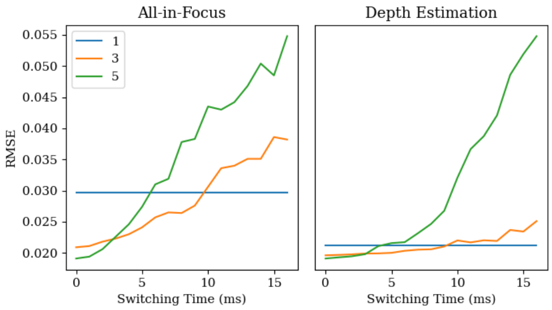

Figure 12. **Effect of switching time on joint system performance.** Reconstruction error across phase mask counts as a function of switching time with 100ms overall exposure. Performance of the jointly optimized system degrades as the switching time between phase masks increases, as expected. Our system still performs well when the time spent switching is less than 25% of the overall exposure.

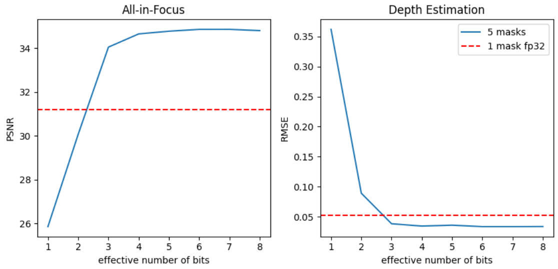

Figure 13. **Effect of quantization on joint system performance.** Performance of the jointly optimized system degrades as the number of effective bits decrease; however, even under heavy quantization, our system outperforms single mask setups.

time-averaged PSFs are robust to quantization noise: 5 masks quantized to 4-bits outperform one mask quantized to 32-bits. This is intuitive as multiple masks can leverage averaging to achieve better precision and remove noise [4].

8.5. A Path Towards Better Single-Mask Designs

The performance of end-to-end designed phase masks is highly sensitive to how they are initialized [32]. Can we use our 5-mask design to provide a better initialization for a 1-mask design?

Starting from random initializations, we optimized 200 single phase masks to minimize the mean squared error between each of their PSFs and the time-averaged PSF formed by our optimized 5-mask system. We then took the single mask whose PSF was closest to the 5-mask design and fine-tuned our reconstruction network using its PSF. This static design improved performance over a randomly initialized end-to-end trained 1-mask system; $+2.67$ PSNR for all-in-focus imaging and -0.003 RMSE for depth estimation. However, its performance was still strictly worse than a 5-mask system.

8.6. Modulating Amplitude and Phase

As illustrated in our proof of Theorem 1, "complex" masks which modulate both amplitude and phase are more expressive than phase-only masks. Using end-to-end learning, we optimized a pixel wise complex mask (initialized with a random phase mask and all ones amplitude mask) on the joint depth estimation and all-in-focus imaging task. The complex mask offered substationally improved performance; $+2.49$ PSNR for all-in-focus imaging and -0.015 RMSE for depth estimation. However, its performance too was still strictly worse than a 5-mask system. We conjecture this is due to improved training stability of the 5-mask system.

9. Limitations

While we were successful in learning dynamic phase masks to improve state-of-the-art performance on imaging tasks, our method carries some limitations. First, our optical model assumes perfect switching between phase masks during training. While evaluation with non-zero switching times showed little degradation of performance, accounting for state switching while training could produce phase masks that are more performant. Our optical model also simulates depths as layered masks over an image, which does not account for blending at depth boundaries. Additionally, our method assumes that scenes are static for the duration of a single exposure. Lastly, though their prices are falling, SLMs are still quite expensive and bulky.

10. Conclusion

This work is founded upon the insight that the set of PSFs that can be described by a single phase mask is nonconvex and that, as a result, time-averaged PSFs are fundamentally more expressive. We demonstrate that one can learn a sequence of phase masks that, when one dynamically switches among them over time, can substantially improve computational imaging performance across a range of tasks, including depth estimation and all-in-focus imaging. Our work unlocks an exciting new direction for PSF engineering and computational imaging system design.

Acknowledgements

This work was partially supported by AFOSR Young Investigator Program Award FA9550-22-1-0208.

References

[1] Terry A. Bartlett, William C. McDonald, and James N. Hall. Adapting Texas Instruments DLP technology to demonstrate a phase spatial light modulator. In Michael R. Douglass, John

Ehmke, and Benjamin L. Lee, editors, *Emerging Digital Micromirror Device Based Systems and Applications XI*, volume 10932, page 109320S. International Society for Optics and Photonics, SPIE, 2019. 2, 8

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. 5, 6

[3] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *Proc. IEEE ICCV*, 2019. 2, 5, 6

[4] Suyeon Choi, Manu Gopakumar, Yifan Peng, Jonghyun Kim, Matthew O'Toole, and Gordon Wetzstein. Time-multiplexed neural holography: A flexible framework for holographic near-eye displays with fast heavily-quantized spatial light modulators. In *Proceedings of the ACM SIGGRAPH*, page 1–9, 2022. 2, 9

[5] Suyeon Choi, Manu Gopakumar, Yifan Peng, Jonghyun Kim, and Gordon Wetzstein. Neural 3d holography: Learning accurate wave propagation models for 3d holographic virtual and augmented reality displays. *ACM Trans. Graph.*, 40(6), December 2021. 2

[6] Edward R. Dowski and W. Thomas Cathey. Extended depth of field through wave-front coding. *Appl. Opt.*, 34(11):1859–1866, April 1995. 2

[7] Robert Fischer, Yicong Wu, Pakorn Kanchanawong, Hari Shroff, and Clare Waterman-Storer. Microscopy in 3d: A biologist's toolbox. *Trends in cell biology*, 21:682–91, October 2011. 1

[8] Joseph W. Goodman. *Introduction to fourier optics*. Freeman, 2017. 2, 3, 4

[9] Samuel W. Hasinoff and Kiriakos N. Kutulakos. A layer-based restoration framework for variable-aperture photography. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007. 2

[10] Hayato Ikoma, Cindy M. Nguyen, Christopher A. Metzler, Yifan Peng, and Gordon Wetzstein. Depth from defocus with learned optics for imaging and occlusion-aware depth estimation. *IEEE International Conference on Computational Photography (ICCP)*, 2021. 2, 5, 6

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5

[12] Anat Levin, Rob Fergus, Frédo Durand, and William T. Freeman. Image and depth from a conventional camera with a coded aperture. In *ACM SIGGRAPH 2007 Papers*, SIGGRAPH '07, page 70–es, New York, NY, USA, 2007. Association for Computing Machinery. 1, 2

[13] Xin Liu, Linpei Li, Xu Liu, Xiang Hao, and Yifan Peng. Investigating deep optics model representation in affecting resolved all-in-focus image quality and depth estimation fidelity. *Opt. Express*, 30(20):36973–36984, September 2022. 2, 5, 6

[14] Yawen Lu, Sophia Kourian, Carl Salvaggio, Chenliang Xu, and Guoyu Lu. Single image 3d vehicle pose estimation for augmented reality. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5, 2019. 1

[15] Nikolaus Mayer, Eddy Ilg, Philip Häusser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. 4, 5

[16] Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1375–1385, 2020. 2

[17] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5

[18] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. In *Medical Imaging with Deep Learning*, 2018. 4

[19] Luca Palmieri, Gabriele Scrofani, Nicolò Incardona, Genaro Saavedra, Manuel Martínez-Corral, and Reinhard Koch. Robust depth estimation for light field microscopy. *Sensors*, 19(3), 2019. 1

[20] Alex Paul Pentland. A new sense for depth of field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(4):523–531, 1987. 2

[21] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 5, 6

[22] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2022. 4, 5

[23] Anupa Sabnis and Leena Vachhani. Single image based depth estimation for robotic applications. In *2011 IEEE Recent Advances in Intelligent Computational Systems*, pages 102–106, 2011. 1

[24] Yoav Shechtman, Steffen J. Sahl, Adam S. Backer, and W. E. Moerner. Optimal point spread function design for 3d imaging. *Phys. Rev. Lett.*, 113:133902, September 2014. 2

[25] Yoav Shechtman, Lucien Weiss, Adam Backer, Steffen Sahl, and William Moerner. Precise three-dimensional scan-free multiple-particle tracking over large axial ranges with tetrapod point spread functions. *Nano letters*, 15, May 2015. 1

[26] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Trans. Graph.*, 37(4), July 2018. 1, 2, 5, 6

[27] Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1386–1396, 2020. 2

[28] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *ACM Trans. Graph.*, 26(3):69–es, jul 2007. 2

[29] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[30] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[31] Woontack Woo, Wonwoo Lee, and Nohyoung Park. Depth-assisted real-time 3d object detection for augmented reality. In *International Conference on Artificial Reality and Telexistence (ICAT)*, 2011. 1

[32] Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. Phasecam3d — learning phase masks for passive single view depth estimation. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12, 2019. 1, 2, 3, 4, 5, 6, 7, 9

[33] Feng Xue, Guirong Zhuo, Ziyuan Huang, Wufei Fu, Zhuoyue Wu, and Marcelo H Ang. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2330–2337. IEEE, 2020. 1

[34] Menglong Ye, Edward Johns, Ankur Handa, Lin Zhang, Philip Pratt, and Guang-Zhong Yang. Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery, 2017. 1