Deep Optics for Monocular Depth Estimation and 3D Object Detection

Julie Chang
Stanford University
jchang10@stanford.edu

Gordon Wetzstein
Stanford University
gordon.wetzstein@stanford.edu

Abstract

Depth estimation and 3D object detection are critical for scene understanding but remain challenging to perform with a single image due to the loss of 3D information during image capture. Recent models using deep neural networks have improved monocular depth estimation performance, but there is still difficulty in predicting absolute depth and generalizing outside a standard dataset. Here we introduce the paradigm of deep optics, i.e. *end-to-end design of optics and image processing, to the monocular depth estimation problem, using coded defocus blur as an additional depth cue to be decoded by a neural network. We evaluate several optical coding strategies along with an end-to-end optimization scheme for depth estimation on three datasets, including NYU Depth v2 and KITTI. We find an optimized freeform lens design yields the best results, but chromatic aberration from a singlet lens offers significantly improved performance as well. We build a physical prototype and validate that chromatic aberrations improve depth estimation on real-world results. In addition, we train object detection networks on the KITTI dataset and show that the lens optimized for depth estimation also results in improved 3D object detection performance.*

1. Introduction

Depth awareness is crucial for many 3D computer vision tasks, including semantic segmentation [33, 38, 10], 3D object detection [37, 22, 11, 40, 41], 3D object classification [45, 24, 30], and scene layout estimation [49]. The required depth information is usually obtained with specialized camera systems, for example using time-of-flight, structured illumination, pulsed LiDAR, or stereo camera technology. However, the need for custom sensors, high-power illumination, complex electronics, or bulky device form factors often makes it difficult or costly to employ these specialized devices in practice.

Single-image depth estimation with conventional cameras has been an active area of research. Traditional approaches make use of pre-defined image features that are statistically correlated with depth, *e.g.* shading, perspective distortions, occlusions, texture gradients, and haze [17, 35, 16, 48, 36, 18]. Recently, significant improvements have been achieved by replacing hand-crafted features with learned features via convolutional neural networks (CNNs) [5, 19, 8, 6]. While these methods tend to perform decently within consistent datasets, they do not generalize well to scenes that were not part of the training set. In essence, the problem of estimating a depth map from pictorial cues alone is ill-posed. Optically encoding depth-dependent scene information has the potential to remove some of the ambiguities inherent in all-in-focus images, for example using (coded) defocus blur [28, 26, 20, 43, 1] or chromatic aberrations [42]. However, it is largely unclear how different optical coding strategies compare to one another and what the best strategy for a specific task may be.

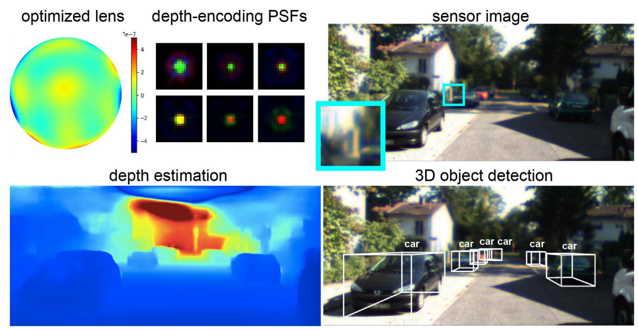

Figure 1. We apply deep optics, *i.e.* end-to-end design of optics and image processing, to build an optical-encoder, CNN-decoder system for improved monocular depth estimation and 3D object detection.

Inspired by recent work on deep optics [2, 39, 12], we interpret the depth estimation problem with coded defocus blur as an optical-encoder, electronic-decoder system that can be trained in an end-to-end manner. Although co-designing optics and image processing is a core idea in computational photography, only differentiable estimation algorithms, such as neural networks, allow for true end-to-end computational camera designs. Here, error backpropagation during training not only optimizes network weights but also physical lens parameters. With the proposed deep optics approach, we evaluate several optical coding strategies for two important 3D scene understanding problems: monocular depth estimation and 3D object detection.

In a series of experiments, we demonstrate that the deep optics approach optimizes the accuracy of depth estimation across several datasets. Consistent with previous work, we show that optical aberrations that are typically considered undesirable for image quality are highly beneficial for encoding depth cues. Our results corroborate that defocus blur provides useful information, and we additionally find that adding astigmatism and chromatic aberrations further improves accuracy. We achieve the best results by jointly optimizing a freeform lens, *i.e*. the spatially varying surface height of a lens, together with the CNN's weights. Surprisingly though, we find that the accuracy of the optimized lenses is only slightly better than standard defocus with chromatic aberrations. This insight motivates the use of simple cameras with only a single lens over complex lens systems when prioritizing depth estimation quality, which we validate with an experimental prototype.

We also evaluate the benefits of deep optics for higher-level 3D scene understanding tasks. To this end, we train a PointNet [29] 3D object detection network on the KITTI dataset. We find that, compared to all-in-focus monocular images, images captured through the optimized lenses also perform better in 3D object detection, a task which requires semantic understanding on top of depth estimation to predict 3D bounding boxes on object instances.

In sum, our experiments demonstrate that an optimized lens paired with a concurrently trained neural network can improve depth estimation without sacrificing higher-level image understanding. Specifically, we make the following contributions:

- We build a differentiable optical image formation model that accounts for either fixed (defocus, astigmatism, chromatic aberration) or optimizable (freeform or annular) lens designs, which we integrate with a differentiable reconstruction algorithm, i.e. a CNN.
- We evaluate the joint optical-electronic model with the various lens settings on three datasets (Rectangles, NYU Depth-v2, KITTI). The optimized freeform phase mask yields the best results, with chromatic aberrations coming in a close second.
- We build a physical prototype and validate that captured images with chromatic aberrations achieve better depth estimation than their all-in-focus counterparts.
- We train a 3D object detection network with the optimized lens and demonstrate that the benefits of improved depth estimation carry through to higher level 3D vision.

Note that the objective of our work is not to develop the state-of-the-art network architecture for depth estimation, but to understand the relative benefits of deep optics over fixed lenses. Yet, our experiments show that deep optics achieves lower root-mean-square errors on depth estimation tasks with a very simple U-Net [34] compared to more complex networks taking all-in-focus images as input.

2. Related Work

Deep Monocular Depth Estimation Humans are able to infer depth from a single image, provided enough contextual hints that allow the viewer to draw from past experiences. Deep monocular depth estimation algorithms aim to mimic this capability by training neural networks to perform this task [5, 19, 8, 6]. Using various network architectures, loss functions, and supervision techniques, monocular depth estimation can be fairly successful on consistent datasets such as KITTI [7] and NYU Depth [38]. However, performance is highly dependent on the training dataset. To address this issue, several recent approaches have incorporated physical camera parameters into their image formation model, including focal length [14] and defocus blur [1], to implicitly encode 3D information into a 2D image. We build on these previous insights and perform a significantly more extensive study that evaluates several types of fixed lenses as well as fully optimizable camera lenses for monocular depth estimation and 3D object detection tasks.

Computational Photography for Depth Estimation Modifying camera parameters for improved depth estimation is a common approach in computational photography. For example, coding the amplitude [20, 43, 50] or phase [21] of a camera aperture has been shown to improve depth reconstruction. Chromatic aberrations have also been shown to be useful for estimating the depth of a scene [42]. Whereas conventional defocus blur is symmetric around the focal plane, *i.e*. there is one distance in front of the focal plane that has the same PSF as another distance behind the focal plane, defocus blur with chromatic aberrations is unambiguous. In all these approaches, depth information is encoded into the image to help an algorithm succeed at a certain task, such as depth estimation. In this paper, we combine related optical coding techniques with more contemporary deep-learning methods. The primary benefit of a deep learning approach over previous work is that it allows a loss function to be applied to a high-level vision task, *e.g*. object detection, which can then directly influence physical camera parameters in a principled manner.

Deep Optics Deep learning can be used for jointly training camera optics and CNN-based estimation methods. This approach was recently demonstrated for extended depth of field and superresolution imaging [39], image classification [2], and multicolor localization microscopy [25]. For the application of monocular depth estimation, Haim *et al*. designed a phase mask consisting of concentric rings

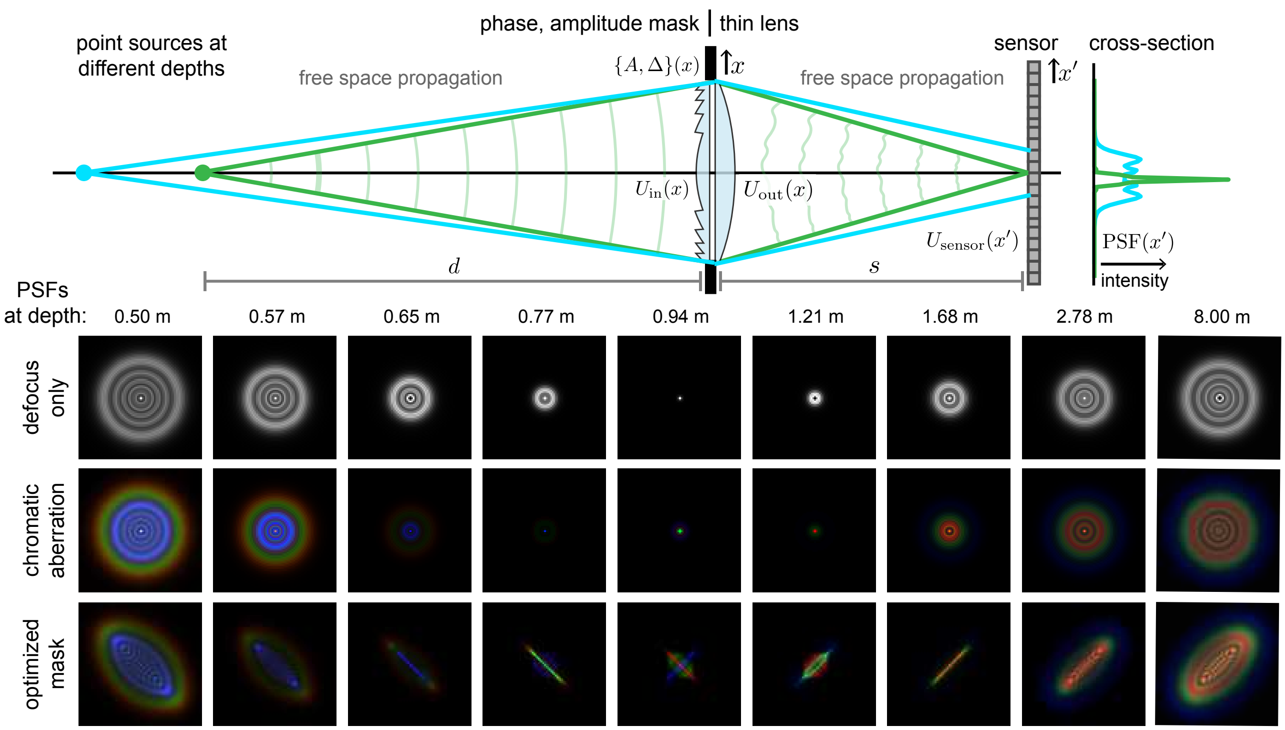

Figure 2. **PSF simulation model.** (Top) Optical propagation model of point sources through a phase mask placed in front of a thin lens. PSFs are simulated by calculating intensity of the electric field at the sensor plane. (Bottom) Sample PSFs from thin lens defocus only, with chromatic aberrations, and using an optimized mask initialized with astigmatism.

to induce chromatic aberrations that could serve as depth cues [12]. The training process optimized the ring radii and phase shifts within two or three annular rings but did not allow for deviation from this ring-based template. Concurrently and independently of our work, Wu *et al*. also developed a jointly optimized phase mask for CNN-based depth estimation [44]. However, unique to our paper, we evaluate the comparative performances of non-optimized aberrated lenses as well as fully optimizable freeform lenses, allowing us to fairly compare the optimized optics to these types of typically undesirable aberrations. Moreover, our work provides results for additional commonly used datasets as well as an evaluation of the benefits of our depth-optimized lens for a higher-level vision task like 3D object detection.

3. Differentiable Image Formation Model

To optimize optical lens elements that best encode depth-dependent scene information, we model light transport in the camera using wave optics. This is not only physically accurate but also allows for both refractive and diffractive optical elements to be optimized. Due to the fact that the light in most natural scenes is incoherent, we only rely on a coherent light transport model to simulate the depth- and wavelenth-dependent point spread function (PSF) of the system, which we then use to simulate sensor images.

3.1. Modeling Conventional Cameras

We begin by building a camera model consisting of a single convex thin lens with focal length f at a distance s from the sensor (see Fig. 2). The relationship between the in-focus distance and the sensor distance is given by the thin-lens equation:

$$1/f = 1/d + 1/s \tag{1}$$

Hence an object at a distance d in front of the lens appears in focus at a distance s behind the lens.

When imaging a real-world scene, there are likely to be objects at multiple depths that are imaged with different PSFs. To simulate the PSF at a depth z, we consider a point emitter of wavelength λ centered on the optical axis located a distance z away from the center of the thin lens. Our general approach is to propagate the wave of light through the optical system to the sensor. To begin, we first propagate the light emitted by the point, represented as a spherical wave, to the lens. The complex-valued electric field immediately before the lens is given by:

$$U_{\text{in}}(x, y) = \exp(ik\sqrt{x^2 + y^2 + z^2}) \tag{2}$$

where $k = 2\pi/\lambda$ is the wavenumber.

The next step is to propagate this wave field through the lens by multiplying the input by a phase delay, $t(x, y)$, induced by the thickness and index of refraction at each location on the lens. The thickness profile, $\Delta(x, y)$, of a convex thin lens with focal length f and index of refraction $n(\lambda)$ in a paraxial regime [9] is

$$\Delta(x, y) = \Delta_0 - \frac{x^2 + y^2}{2f(n(\lambda) - 1)} \tag{3}$$

where Δ_0 is the center thickness. Note that the refractive index is wavelength-dependent, which results in chromatic aberrations when using a fixed singlet lens with multiple wavelengths. Converting thickness to the corresponding phase shift, $\phi = k(n - 1)\Delta$, and neglecting the constant phase offset from Δ_0, the phase transformation is

$$t(x, y) = e^{i\phi(x,y)} = \exp\left[-i\frac{k}{2f}(x^2 + y^2)\right] \tag{4}$$

Additionally, since a lens has some finite aperture size, we insert an amplitude function $A(x, y)$ that blocks all light in regions outside the open aperture. To find the electric field immediately after the lens, we multiply the amplitude and phase modulation of the lens with the input electric field:

$$U_{\text{out}}(x, y) = A(x, y)\, t(x, y)\, U_{\text{in}}(x, y) \tag{5}$$

Finally, the field propagates a distance s to the sensor with the exact transfer function [9]:

$$H_{\text{s}}(f_x, f_y) = \exp\left[iks\sqrt{1 - (\lambda f_x)^2 - (\lambda f_y)^2}\right] \tag{6}$$

where (f_x, f_y) are spatial frequencies. This transfer function is applied in the Fourier domain as:

$$U_{\text{sensor}}(x', y') = \mathcal{F}^{-1}\left\{\mathcal{F}\left\{U_{\text{out}}(x, y)\right\} \cdot H_{\text{s}}(f_x, f_y)\right\} \tag{7}$$

where $\mathcal{F}$ denotes the 2D Fourier transform. Since the sensor measures light intensity, we take the magnitude-squared to find the final PSF:

$$\text{PSF}_{\lambda,z}(x', y') = |U_{\text{sensor}}(x', y')|^2 \tag{8}$$

By following this sequence of forward calculations, we can generate a 2D PSF for each depth and wavelength of interest. For chromatic aberrations, we calculate $t(x, y)$ for each color channel (Eq. 4), which results in three slightly different PSFs. To approximate an achromatic lens, we use the central wavelength PSF for all color channels (Fig. 2).

3.2. Modeling Freeform Lenses

Several variables such as focal length, focus distance, and aperture size are modeled by the above formulation. For maximum degrees of freedom to shape the PSF, we can also treat the optical element as a freeform lens by assuming that is has an additional arbitrary thickness profile $\Delta_{\text{ff}}(x, y)$. The corresponding phase delay is

$$t_{\text{ff}}(x, y) = \exp\left[jk(n_{\text{ff}}(\lambda) - 1)\Delta_{\text{ff}}(x, y)\right] \tag{9}$$

where $n_{\text{ff}}(\lambda)$ is the wavelength-dependent index of refraction of the lens material. We parametrize Δ_{ff} with the Zernike basis (indices 1-36, [27]), which leads to smoother surfaces. The intensity PSF of a freeform lens is then

$$\text{PSF}_{\lambda,z}(x, y; \lambda) = |\mathcal{F}^{-1}\{\mathcal{F}\{A \cdot t_{\text{lens}} \cdot t_{\text{ff}} \cdot U_{\text{in}}\} \cdot H_{\text{s}}\}|^2(x, y) \tag{10}$$

3.3. Depth-Dependent Image Formation

We can use these simulated PSFs to approximate a captured image of a 3D scene on an RGB sensor. We use a layered representation that models the scene as a set of surfaces on discrete depth planes [13]. This allows for precomputation of a fixed number of PSFs corresponding to each depth plane. We make a few modifications here to suit our datasets consisting of pairs of all-in-focus RGB images and their discretized depth maps. For an all-in-focus image $\mathbf{L}$, a set of $j = 1 \ldots J$ discrete depth layers, and occlusion masks $\{\mathbf{M}_j\}$, we calculate our final image by:

$$\mathbf{I}_\lambda = \sum_{j=1}^{J} (\mathbf{L}_\lambda * \text{PSF}_{\lambda,j}) \circ \mathbf{M}_j \tag{11}$$

where $*$ denotes 2D convolution for each color channel centered on λ, and $\circ$ denotes element-wise multiplication. The occlusion masks $\{\mathbf{M}_j\}$ represent the individual layers of the quantized depth map, with blurring and normalization such that $\sum_j \mathbf{M}_j = 1$ at each pixel, to ensure smooth transitions between depths (see Supplement).

4. Depth Estimation

In this section, we describe our experiments using deep optics for monocular depth estimation with encoded blur.

4.1. Network and Training

For depth estimation, we connect our differentiable image formation model to a U-Net [34] that takes as input either the simulated sensor images or the original all-in-focus dataset images. The network consists of 5 downsampling layers ({Conv-BN-ReLU}×2→MaxPool2×2) followed by 5 upsampling layers with skip connections (Conv^T+Concat→{Conv-BN-ReLU}×2). The output is the predicted depth map, at the same resolution as the input image. We use the ADAM optimizer with a mean-square-error (MSE) loss on the logarithmic depth. We train for 40,000 iterations at a learning rate of .001 (decayed to 1e-4 for the Rectangles dataset) and batch size of 3.

We evaluate on (1) a custom Rectangles dataset, which consists of white rectangles against a black background

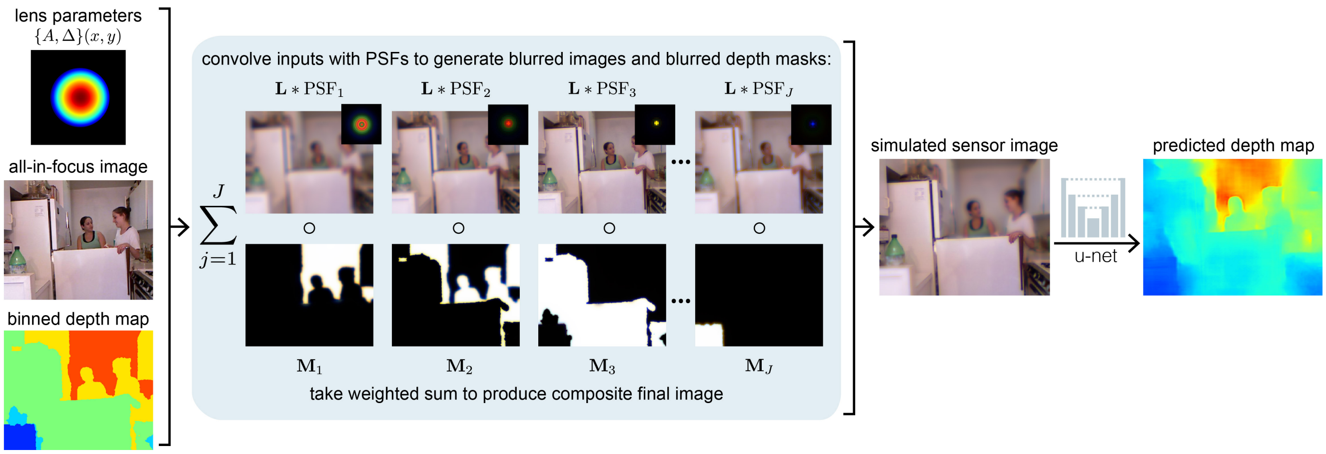

Figure 3. **Depth-dependent image formation.** Given a set of lens parameters, an all-in-focus image, and its binned depth map, the image formation model generates the appropriate PSFs and applies depth-dependent convolution with masking to simulate the corresponding sensor image, which is then passed into a U-Net for depth estimation.

placed at random depths (see Supplement), (2) the NYU Depth v2 dataset with standard splits, and (3) a subset of the KITTI depth dataset (5500 train, 749 val) that overlaps with the object detection dataset for which we obtained dense "ground truth" depth maps from Ma *et al*. [23]. We train on full-size images. We calculate loss for NYU Depth on the standard crop size, and for KITTI only on the official sparse ground truth depth.

For the Rectangles and NYU Depth datasets, we initialize the phase mask as an f/8, 50 mm focal length lens, focused to 1 m. For the KITTI dataset, we initialize an f/8, 80 mm focal length lens, focused to 7.6 m. When the lens is being optimized, we also initialize the U-Net with the optimized weights for the fixed lens, and each training step adjusts the parameters of the lens (Zernike coefficients for freeform, ring heights for annular) and the U-Net. We use 12 depth bins in our simulations, spaced linearly in inverse depth. When optimizing a freeform lens for the KITTI dataset, we reduce this to 6 intervals due to GPU memory constraints and train for 30,000 iterations; then we freeze the lens and increase back to 12 intervals to fine-tune the U-Net for an additional 30,000 iterations.

4.2. Analysis and Evaluation

Table 1 shows a summary of results for all datasets. Examples of simulated sensor images and predicted depth maps from NYU Depth and KITTI are shown in Fig. 4 (see Supplement for Rectangles).

We observe common trends across all datasets. When using the all-in-focus images, errors are highest. This is most intuitive to understand with the Rectangles dataset. If there is a randomly-sized white rectangle floating in space that is always in focus, there are no depth cues for the network to recognize, and the network predicts the mean depth for every rectangle. Depth from defocus-only improves performance, but there is still ambiguity due to symmetric blur along inverse depth in both directions from the focal plane. Astigmatism (see Supplement for details) helps resolve this ambiguity, and the inherent chromatic aberration of a singlet lens further improves results.

We optimize two freeform lenses for each dataset. The annular lens consists of three concentric layers of different heights, inspired by [12]. While these optimized lenses outperformed all-in-focus experiments, they did not yield higher accuracy than chromatic aberration from a fixed lens. In contrast, the optimized freeform lens showed the best results, demonstrating the ability of the end-to-end optimization to learn a new freeform lens that better encodes depth information. For NYU Depth, we found that additionally initializing Δ_{ff} with astigmatism yielded better results.

Table 2 compares default metrics on the NYU Depth test set with reported results from previous works. These comparisons suggest that adding this optical portion of the model can yield results on par with state-of-the-art methods with more heavyweight and carefully designed networks.

4.3. Experimental Results

We build a prototype for monocular depth estimation using chromatic aberration on real-world scenes. Although the fully optimized lens performed best in simulations, chromatic aberrations yield surprisingly good results almost on par with optimized optics. Unlike custom-manufactured optimized lenses, simple lenses with such aberrations are readily available, inexpensive, and provide a small form factor. Thus we chose to utilize off-the-shelf lenses with aberrations for our physical experiments. Our camera consisted of a Canon EOS Rebel T5 camera and a biconvex singlet lens (f = 35mm, Thorlabs) with a circular aperture (D = 0.8 mm). We captured a series of images of a point white light source to calibrate the modeled PSFs, primarily by tun-

Optical model	Rectangles RMSE$_{\text{lin}}$	Rectangles RMSE$_{\text{log}}$	NYU Depth v2 RMSE$_{\text{lin}}$	NYU Depth v2 RMSE$_{\text{log10}}$	KITTI* RMSE$_{\text{lin}}$	KITTI* RMSE$_{\text{log}}$
All-in-focus	0.4626	0.3588	0.9556	0.1452	2.9100	0.1083
Defocus, achromatic	0.2268	0.1805	0.4814	0.0620	2.5400	0.0776
Astigmatism, achromatic	0.1348	0.0771	0.4561	0.0559	2.3634	0.0752
Chromatic aberration	*0.0984*	*0.0563*	*0.4496*	*0.0556*	*2.2566*	*0.0702*
Optimized, annular	0.1687	0.1260	0.4817	0.0623	2.7998	0.0892
Optimized, freeform	**0.0902**	**0.0523**	**0.4325**	**0.0520**	**1.9288**	**0.0621**

Table 1. Depth estimation error with different optical models for various datasets. RMSEs are reported for linear and log (base e or 10) scaling of depth (m or log(m)). Lowest errors are bolded, and second-lowest are italicized. The KITTI* dataset is our KITTI dataset subset.

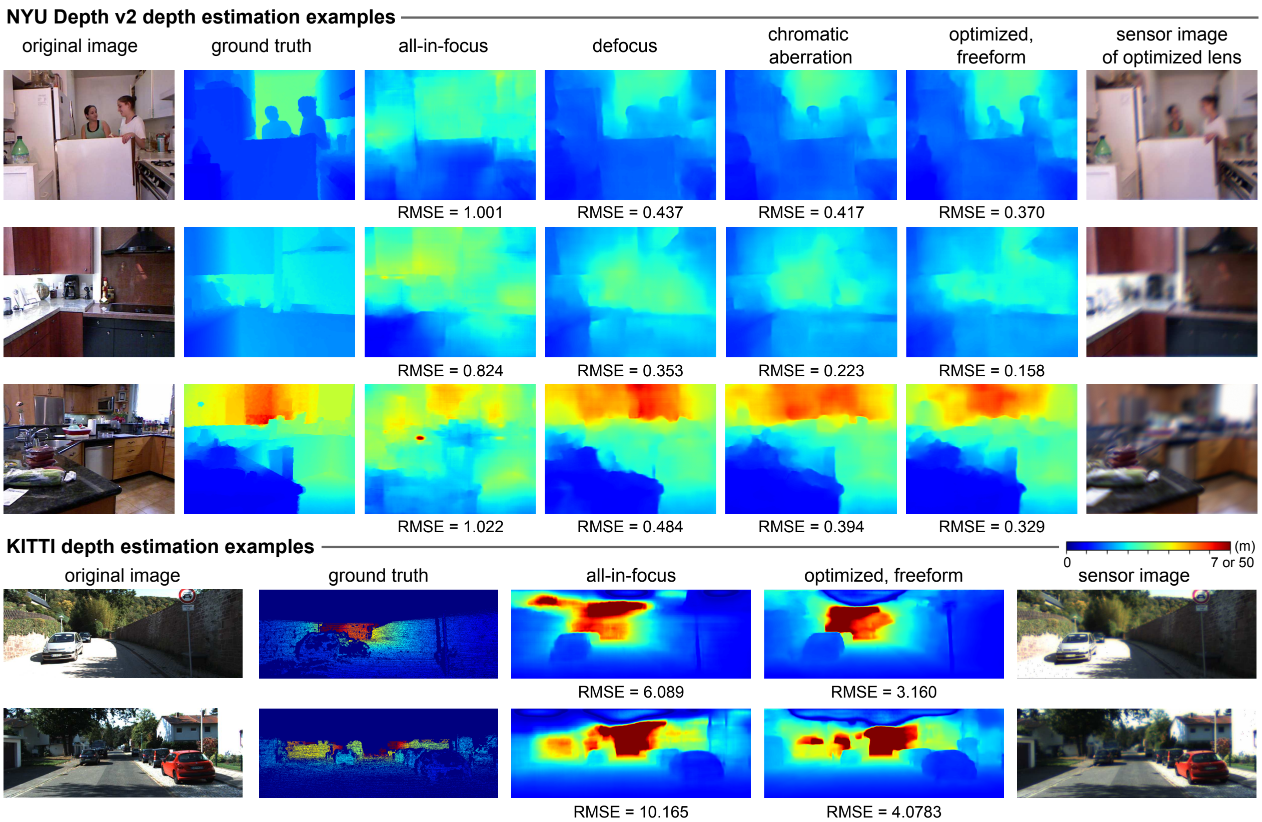

Figure 4. **Depth estimation.** (Top) Examples with RMSE (m) from the NYU Depth v2 dataset with all-in-focus, defocus, chromatic aberration, and optimized models. The simulated sensor image from the optimized system is also shown. (Bottom) Examples with RMSE (m) from the KITTI dataset (cropped to fit) with all-in-focus and optimized models; the sensor image from the optimized model is also shown. All depth maps use the same colormap, but the maximum value is 7 m for NYU Depth and 50 m for KITTI.

ing a spherical aberration parameter. We retrain a depth estimation network for the calibrated PSFs with the NYU Depth dataset, including a downsampling factor of four due to the smaller image size of dataset compared to the camera sensor. For this network, we apply sRGB conversion to produce the simulated sensor image, which allows us to directly input sRGB camera images during evaluation.

We capture pairs of images with the prototype as described along with an all-in-focus image obtained by adding a 1 mm pinhole (see Supplement). We use our retrained depth estimation network to predict a depth map from the blurry images, and we use the all-in-focus network to predict the corresponding depth map from the all-in-focus images. Fig. 5 shows a few examples; more are included in the supplement. Depth estimation with the optical model performs significantly better on the captured images, as physical depth information is encoded into the images, allowing the network to rely not just on dataset priors for prediction.

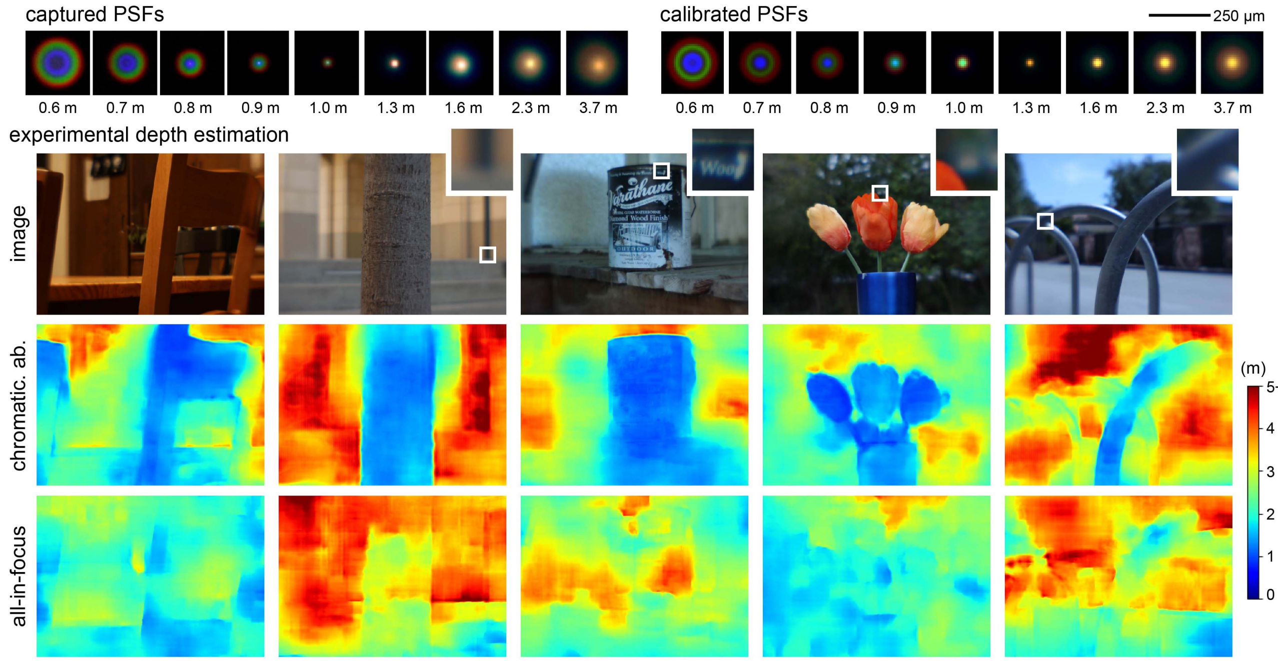

Figure 5. **Real-world capture and depth estimation.** (Top) Captured and calibrated depth-dependent PSFs, displayed at the same scale. (Bottom) Examples of images captured using our prototype with a zoomed region inset, depth estimation with chromatic aberration, and depth estimation from the corresponding all-in-focus image (not shown). Depth map colorscale is the same for all depth maps.

Method	rel	log10	rms	δ_1	δ_2	δ_3
Laina *et al.* [19]	0.127	0.055	0.573	0.811	0.953	0.988
MS-CRF [47]	0.121	0.052	0.586	0.811	0.954	0.987
DORN [6]	0.115	**0.051**	0.509	0.828	0.965	0.992
All-in-focus	0.293	0.145	0.956	0.493	0.803	0.936
Defocus	0.108	0.062	0.481	0.893	0.981	0.996
Astigmatism	0.095	0.056	0.456	0.916	0.986	0.998
Chromatic	0.095	0.056	0.450	0.916	0.987	0.998
Freeform	**0.087**	0.052	**0.433**	**0.930**	**0.990**	**0.999**

Table 2. Comparative performance on NYU Depth v2 test set, as calculated in [5]. Units are in meters or log10(m). Thresholds are denoted $\delta_i : \delta > 1.25^i$. Lowest errors and highest δs are bolded.

A limitation of our prototype is its smaller field of view, mainly due to the spatially varying nature of the real PSF, which prevented processing of full indoor room scenes. This could be improved by adding another lens to correct for off-axis aberrations [4] or by including these variations in the image formation model [15]. Modeling spatially varying PSFs is challenging because the image formation model becomes a much more computationally intensive simulation, and our U-Net-based network that works best for shift invariance may not be as well-suited. For these and other reasons, no existing deep optics-like approach actually models off-axis aberrations, yet this would be a very valuable direction of future work.

Object detection metric	**All-in-focus**	**Optimized**
2D mAP	78.01	**78.96**
2D AP, Car	**95.50**	95.15
2D AP, Pedestrian	80.06	**80.22**
2D AP, Cyclist	**89.77**	88.11
3D AP, Ped., Easy	9.74	**13.86**
3D AP, Ped., Moderate	7.10	**11.74**
3D AP, Ped., Hard	6.21	**11.90**
3D AP, Cyc., Easy	2.27	**7.18**
3D AP, Cyc., Moderate	2.36	**4.89**
3D AP, Cyc., Hard	1.98	**4.95**

Table 3. Object detection performance measured by 2D AP % (IoU = 0.5) and 3D AP % (IoU = 0.5) on our validation split of the KITTI object detection dataset using the all-in-focus and optimized mask models. Higher values are bolded.

5. 3D Object Detection

To assess whether an optical system optimized for improved depth estimation is beneficial for higher-level 3D scene understanding as well, we evaluate 3D object detection performance on the KITTI dataset using the earlier depth-optimized lens. 3D object detection requires recognizing different instances of objects and regressing an oriented 3D bounding box around each object instance. Depth information, whether implicitly contained in an image or

Method	Input	3D object localization			3D object detection		
		Easy	Moderate	Hard	Easy	Moderate	Hard
Mono3D [3]	RGB	5.22	5.19	4.13	2.53	2.31	2.31
MF3D [46]	RGB	22.03	13.63	11.6	10.53	5.69	5.39
MonoGRNet [31]	RGB	-	-	-	13.88	10.19	7.62
VoxelNet [51]	RGB+LIDAR	89.6	84.81	78.57	81.97	65.46	62.85
FPointNet [29]	RGB+LIDAR	88.16	84.02	76.44	83.76	70.92	63.65
(Ours) All-in-focus	RGB	26.71	19.87	19.11	16.86	13.82	13.26
(Ours) Optimized, freeform	RGB	37.51	25.83	21.05	25.20	17.07	13.43

Table 4. 3D object localization AP % (bird's eye view) and 3D object detection AP % (IoU= 0.7) for the car class. The listed numbers from literature are reported on the KITTI validation set; results from our methods are reported on our KITTI* validation split (Sec. 4.1).

explicitly provided from a depth sensor, is critical for this task, as is evidenced in the large gap in performance between the RGB and RGB+LIDAR methods in Table 4.

We train a 3D object detection network specific to the freeform lens optimized for KITTI depth estimation. In particular, we use a Frustum PointNet v1 (FPointNet, [29]), which was demonstrated to work with both sparse LIDAR point clouds and dense depth images. FPointNet uses 2D bounding box predictions on the RGB image to generate frustum proposals that bound a 3D search space; then 3D segmentation and box estimation occur on the 3D point cloud contained within each frustum. In our modified network, we substitute the LIDAR point clouds with our estimated depth projected into a 3D point cloud. As in the original method, ground truth 2D boxes augmented with random translation and scaling are used during training, but estimated 2D bounding boxes from a separately trained 2D object detection network (Faster R-CNN, [32]) are used during validation. For comparison, we train the same networks with all-in-focus images and their estimated depth maps. More details and videos are included in the Supplement.

Results of our object detection experiments are shown in Tables 3 and 4. Average precision (AP) values are computed by the standard PASCAL protocol, as described in the KITTI development kit. 2D object detection performance is similar between the all-in-focus and optimized systems, which implies that even though the sensor images from the optimized optical element appear blurrier than the all-in-focus images, the networks are able to extract comparable information from the two sets of images. More notably, 3D object detection improves with the optimized optical system, indicating that the FPointNet benefits from the improved depth maps enabled with the optimized lens.

6. Discussion

Throughout our experiments, we demonstrate that a joint optical-encoder, electronic-decoder model outperforms the corresponding optics-agnostic model using all-in-focus images. We build a differentiable optical image formation layer that we join with a depth estimation network to allow for end-to-end optimization from camera lens to network weights. The fully optimized system yields the most accurate depth estimation results, but we find that native chromatic aberrations can also encode valuable depth information. Additionally, to verify that improved depth encoding does not need to sacrifice other important visual content, we show that the lens optimized for depth estimation maintains 2D object detection performance while further improving 3D object detection from a single image.

As mentioned, our conclusions are drawn from the relative performances between our results. We do not claim to conclusively surpass existing methods, as we use the ground truth or pseudo-truth depth map in simulating our sensor images, and we are limited to a layer-based image formation model. These simulation approximations are not straightforward to disentangle unless the entire dataset could be recaptured through the different lenses. Nonetheless, our real-world experimental results are promising in supporting the advantage of optical depth encoding, though more experiments, especially with a larger field-of-view, would be valuable. We are interested in future work to see how an optical layer can further improve leading methods, whether for monocular depth estimation [19, 47, 6] or other tasks.

More broadly, our results consistently support the idea that incorporating the camera as an optimizable part of the network offers significant benefits over considering the image processing completely separately from image capture. We have only considered the camera as a single static optical layer in this paper, but there may be potential in more complex designs as research in both optical computing and computer vision continues to advance.

Acknowledgments

We thank Vincent Sitzmann and Mark Nishimura for insightful advice. This project was supported by an NSF CAREER Award (IIS 1553333), an Okawa Research Grant, a Sloan Fellowship, a Visual Computing Center CCF Grant of KAUST Office of Sponsored Research, and a PECASE Award (W911NF-19-1-0120).

References

[1] Marcela Carvalho, Bertrand Le Saux, Pauline Trouvé-Peloux, Andrés Almansa, and Frédéric Champagnat. Deep depth from defocus: how can defocus blur improve 3d estimation using dense neural networks? In *European Conference on Computer Vision*, pages 307–323. Springer, 2018. 1, 2

[2] Julie Chang, Vincent Sitzmann, Xiong Dun, Wolfgang Heidrich, and Gordon Wetzstein. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Scientific reports*, 8(1):12324, 2018. 1, 2

[3] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2147–2156, 2016. 8

[4] Oliver Cossairt and Shree Nayar. Spectral focal sweep: Extended depth of field from chromatic aberrations. In *2010 IEEE International Conference on Computational Photography (ICCP)*, pages 1–8. IEEE, 2010. 7

[5] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014. 1, 2, 7

[6] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018. 1, 2, 7, 8

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 2

[8] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017. 1, 2

[9] Joseph W Goodman. *Introduction to Fourier optics*. Macmillan Learnng, 4 edition, 2017. 4

[10] Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 564–571, 2013. 1

[11] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *European Conference on Computer Vision*, pages 345–360. Springer, 2014. 1

[12] Harel Haim, Shay Elmalem, Raja Giryes, Alex M Bronstein, and Emanuel Marom. Depth estimation from a single image using deep learned phase coded mask. *IEEE Transactions on Computational Imaging*, 4(3):298–310, 2018. 1, 3, 5

[13] Samuel W Hasinoff and Kiriakos N Kutulakos. A layer-based restoration framework for variable-aperture photography. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007. 4

[14] Lei He, Guanghui Wang, and Zhanyi Hu. Learning depth from single images with deep neural network embedding focal length. *IEEE Transactions on Image Processing*, 27(9):4676–4689, 2018. 2

[15] Felix Heide, Mushfiqur Rouf, Matthias B Hullin, Björn Labitzke, Wolfgang Heidrich, and Andreas Kolb. High-quality computational imaging through simple lenses. *ACM Trans. Graph.*, 32(5):149–1, 2013. 7

[16] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172, 2007. 1

[17] Berthold KP Horn. Obtaining shape from shading information. *The psychology of computer vision*, pages 115–155, 1975. 1

[18] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 89–96, 2014. 1

[19] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016. 1, 2, 7, 8

[20] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. *ACM transactions on graphics (TOG)*, 26(3):70, 2007. 1, 2

[21] Anat Levin, Samuel W Hasinoff, Paul Green, Frédo Durand, and William T Freeman. 4d frequency analysis of computational cameras for depth of field extension. In *ACM Transactions on Graphics (TOG)*, volume 28, page 97. ACM, 2009. 2

[22] Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1417–1424, 2013. 1

[23] Fangchang Mal and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018. 5

[24] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 922–928. IEEE, 2015. 1

[25] Tomer Michaeli, Yoav Shechtman, et al. Multicolor localization microscopy by deep learning. *arXiv preprint arXiv:1807.01637*, 2018. 2

[26] Shree K Nayar and H Murase. Illumination planning for object recognition in structured environments. In *1994 IEEE International Conference on Computer Vision and Pattern Recognition*, pages 31–38, 1994. 1

[27] Robert J Noll. Zernike polynomials and atmospheric turbulence. *JOsA*, 66(3):207–211, 1976. 4

[28] Alex Paul Pentland. A new sense for depth of field. *IEEE transactions on pattern analysis and machine intelligence*, (4):523–531, 1987. 1

[29] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 918–927, 2018. 2, 8

[30] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2016. 1

[31] Zengyi Qin, Jinglu Wang, and Yan Lu. Monogrnet: A geometric reasoning network for monocular 3d object localization. *arXiv preprint arXiv:1811.10247*, 2018. 8

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. 8

[33] Xiaofeng Ren, Liefeng Bo, and Dieter Fox. Rgb-(d) scene labeling: Features and algorithms. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2759–2766. IEEE, 2012. 1

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 4

[35] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng. Learning depth from single monocular images. In *Advances in neural information processing systems*, pages 1161–1168, 2006. 1

[36] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2009. 1

[37] Abhinav Shrivastava and Abhinav Gupta. Building part-based object detectors via 3d geometry. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1745–1752, 2013. 1

[38] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012. 1, 2

[39] 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 Transactions on Graphics (TOG)*, 37(4):114, 2018. 1, 2

[40] Shuran Song and Jianxiong Xiao. Sliding shapes for 3d object detection in depth images. In *European conference on computer vision*, pages 634–651. Springer, 2014. 1

[41] Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 808–816, 2016. 1

[42] Pauline Trouvé, Frédéric Champagnat, Guy Le Besnerais, Jacques Sabater, Thierry Avignon, and Jérôme Idier. Passive depth estimation using chromatic aberration and a depth from defocus approach. *Applied optics*, 52(29):7152–7164, 2013. 1, 2

[43] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. In *ACM transactions on graphics (TOG)*, volume 26, page 69. ACM, 2007. 1, 2

[44] 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. IEEE, 2019. 3

[45] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 1

[46] Bin Xu and Zhenzhong Chen. Multi-level fusion based 3d object detection from monocular images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2345–2353, 2018. 8

[47] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5354–5362, 2017. 7, 8

[48] Stella X Yu, Hao Zhang, and Jitendra Malik. Inferring spatial layout from a single image via depth-ordered grouping. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–7. IEEE, 2008. 1

[49] Jian Zhang, Chen Kan, Alexander G Schwing, and Raquel Urtasun. Estimating the 3d layout of indoor scenes and its clutter from depth sensors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1273–1280, 2013. 1

[50] Changyin Zhou, Stephen Lin, and Shree K Nayar. Coded aperture pairs for depth from defocus and defocus deblurring. *International journal of computer vision*, 93(1):53–72, 2011. 2

[51] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018. 8