Polka Lines: Learning Structured Illumination and Reconstruction for Active Stereo

Seung-Hwan Baek Felix Heide
Princeton University

Abstract

Active stereo cameras that recover depth from structured light captures have become a cornerstone sensor modality for 3D scene reconstruction and understanding tasks across application domains. Active stereo cameras project a pseudo-random dot pattern on object surfaces to extract disparity independently of object texture. Such hand-crafted patterns are designed in isolation from the scene statistics, ambient illumination conditions, and the reconstruction method. In this work, we propose a method to jointly learn structured illumination and reconstruction, parameterized by a diffractive optical element and a neural network, in an end-to-end fashion. To this end, we introduce a differentiable image formation model for active stereo, relying on both wave and geometric optics, and a trinocular reconstruction network. The jointly optimized pattern, which we dub "Polka Lines," together with the reconstruction network, makes accurate active-stereo depth estimates across imaging conditions. We validate the proposed method in simulation and using with an experimental prototype, and we demonstrate several variants of the Polka Lines patterns specialized to the illumination conditions.

1. Introduction

Active depth cameras have become essential for three-dimensional scene reconstruction and scene understanding, with established and emerging applications across disciplines, including robotics, autonomous drones, navigation, driver monitoring, human-computer interaction, virtual and mixed reality, and remote conferencing. When combined with RGB cameras, depth-sensing methods have made it possible to recover high-fidelity scene reconstructions [23]. Such RGB-D cameras also allowed researchers to collect large-scale RGB-D data sets that propelled work on fundamental computer vision problems, including scene understanding [44, 21] and action recognition [36]. However, while depth cameras under controlled conditions with low ambient light and little object motion are becoming reliable [1, 42], depth imaging in strong ambient light, at long ranges, and for fine detail and highly dynamic scenes remains an open challenge.

A large body of work has explored active depth sensing approaches to tackle this challenge [18, 27, 4, 41], with structure light and time-of-flight cameras being the most successful methods. Pulsed time-of-flight sensors emit pulses of light into the scene and measure the travel time of the returned photons directly by employing sensitive silicon avalanche photo-diodes [51] or single-photon avalanche diodes [5]. Although these detectors are sensitive to a single photon, their low fill factor restricts existing LiDAR sensors to point-by-point scanning with individual diodes, which prohibits the acquisition of dense depth maps. Correlation time-of-flight sensors [18, 25, 27] overcome this challenge by indirectly estimating round-trip time from the phase of temporally modulated illumination. Although these cameras provide accurate depth for indoor scenes, they suffer from strong ambient illumination and multi-path interference [45, 29], are limited to VGA resolution, and they require multiple captures, which makes dynamic scenes a challenge. Active stereo [55, 1, 2] has emerged as the only low-cost depth sensing modality that has the potential to overcome these limitations of existing methods for room-sized scenes. Active stereo cameras equip a stereo camera pair with an illumination module that projects a fixed pattern onto a scene so that, independently of surface texture, stereo correspondence can be reliably estimated. As such, active stereo methods allow for single-shot depth estimates at high resolutions using low-cost diffractive laser dot modules [1] and conventional CMOS sensor deployed in mass-market products including Intel RealSense cameras [1] and the Google Pixel 4 Phones [2]. However, although active stereo has become a rapidly emerging depth-sensing technology, existing approaches struggle with extreme ambient illumination and complex scenes, prohibiting reliable depth estimates in uncontrolled in-the-wild scenarios.

These limitations are direct consequences of the pipeline design of existing active stereo systems, which hand-engineer the illumination patterns and the reconstruction algorithms in isolation. Typically, the illumination pattern is designed in a first step using a diffractive optical element (DOE) placed in front of a laser diode. Existing dot patterns resulting from known diffractive gratings, such as the Dammann grating [10], are employed with the assumption that generating uniform textures ensures robust disparity estimation for the average scene. Given a fixed illumination

pattern, the reconstruction algorithm is then designed with the goal of estimating correspondence using cost-volume methods [7, 22] or learning-based methods [39, 12, 55, 38]. In this conventional design paradigm, the illumination pattern does not receive feedback from the reconstruction algorithm or the dataset of scenes, prohibiting end-to-end learning of optimal patterns, reconstruction algorithms, and capture configurations tailored to the scene.

In this work, we propose a method that jointly learns illumination patterns and a reconstruction algorithm, parameterized by a DOE and a neural network, in an end-to-end manner. The resulting optimal illumination patterns, which we dub "Polka Lines", together with the reconstruction network, allow for high-quality scene reconstructions. Moreover, our method allows us, for the first time, to learn environment-specific illumination patterns for active stereo systems. The proposed method hinges on a differentiable image formation model that relies on wave and geometric optics to make the illumination and capture simulation accurate and, at the same time, efficient enough for joint optimization. We then propose a trinocular active stereo network that estimates an accurate depth map from the sensor inputs. Unlike previous methods that only use binocular inputs from the stereo cameras, our network exploits the known illumination pattern, resulting in a trinocular stereo setup which reduces reconstruction errors near occlusion boundaries. We train the fully differentiable illumination and reconstruction model in a supervised manner and fine-tune the reconstruction for an experimental prototype in a self-supervised manner. The proposed Polka Lines patterns, together with the reconstruction network, allows us to achieve state-of-the-art active stereo depth estimates for a wide variety of imaging conditions.

Specifically, We make the following contributions:

- We introduce a novel differentiable image formation model for active stereo systems based on geometric and wave optics.
- We devise a novel trinocular active stereo network that uses the known illumination pattern in addition to the stereo inputs.
- We jointly learn optimal "Polka Lines" illumination patterns via differentiable end-to-end optimization, which can be specialized to specific illumination conditions.
- We validate the proposed method in simulation and with an experimental prototype. We demonstrate robust depth acquisition across diverse scene scenarios from low light to strong illumination.

2. Related Work

Depth Imaging. Depth cameras can be broadly categorized into two families, passive and active cameras. Passive methods exploit depth cues such as parallax [40, 13], defocus [28], and double refraction [6, 33] that do not require illumination control. Passive methods often fail on challenging scene parts, such as textureless surfaces, where they can produce catastrophic depth estimation errors. Active systems employ specialized illumination modules to tackle textureless surfaces. Major directions include pulsed and continuous-wave time-of-flight sensors [20, 19], gated imaging [15], structured-light sensor [16, 52], and active stereo systems [55]. Among these, active stereo is particularly attractive as it promises robust single-shot depth imaging at low system cost and small form factor. As such, active stereo systems have successfully been deployed in mass-market [1, 2]. However, existing active-stereo systems also struggle in challenging environments with strong ambient light and noisy inputs with varying scene reflectance. This reduced accuracy partly originates from the blind, compartmentalized design process of the illumination pattern, which often does not consider the reconstruction method, scene statistics, and illumination conditions. In this work, we close this gap by proposing to jointly optimize the illumination patterns and the reconstruction method for active stereo.

Illumination Patterns for Active Stereo. Designing an illumination pattern is crucial for the accuracy of correspondence matching in active stereo systems. Existing methods commonly employ Dammann gratings [10] and Vertical Cavity Surface Emitting Lasers that result in locally-distinct, but globally repetitive illumination patterns [30, 26, 1]. This heuristic design is blind to scene statistics, noise levels, and the reconstruction method. Existing methods have attempted to improve depth estimation by employing alternative hand-crafted DOE designs [11, 49, 34] that rely on alternative experts and heuristic metrics on the illumination patterns. We depart from these heuristic designs and instead directly optimize the illumination pattern with the depth reconstruction accuracy as a loss via end-to-end optimization.

Active Stereo Depth Estimation. Depth reconstruction for active-stereo systems aims to estimate accurate correspondence between stereo images with the aid of projected illumination patterns for feature matching. The corresponding large body of work can be categorized into methods relying on classic patch-based correspondence matching [22, 7] and recent learning-based methods [39, 12, 55, 38]. Zhang et al. [55] proposed an active stereo network with self-supervision, removing the cumbersome process of acquiring training data, and improving depth estimation accuracy.

All of these existing reconstruction methods are limited by the fixed illumination pattern. As such, these methods have to adapt to a given pattern and cannot vary the pattern to suit different imaging conditions. We jointly optimize the illumination and reconstruction module, allowing us to tailor the pattern to the reconstruction method and scene statistics. Moreover, departing from existing approaches, the proposed trinocular reconstruction is the first that exploits knowing illumination pattern itself.

Differentiable Optics. With the advent of auto-differentiation frameworks [3, 37], jointly optimizing imaging optics and reconstruction methods has shaped the design process of diverse vision systems [8, 50, 35, 47, 17, 53, 9, 43, 32, 46]. While existing methods have focused on the imaging optics and primarily assume near-field propagation, we instead optimize illumination optics, specifically a DOE in front of a collimated laser, using far-field wave propagation from a laser to the scene. At the same time, we rely on ray optics to simulate stereo imaging via epipolar geometry. This hybrid image formation, which exploits both wave and geometric optics, allows us to efficiently simulate light transport in active stereo systems while being efficient enough for gradient-based end-to-end optimization. We note that Wu et al. [54] proposed a depth-from-defocus method with a learned aperture mask for structured-light systems. However, this blur-based structured-light projection suffers from frequency-limited features. As such, it is orthogonal to the proposed method, which optimizes a diffraction pattern at the far field for active stereo. Related optimization principles for illumination design can also be found in reflectance imaging [24].

3. Differentiable Hybrid Image Formation

To jointly learn structured illumination patterns and reconstruction methods, we introduce a differentiable image formation model for active stereo sensing. Active stereo systems consist of stereo cameras and an illumination module that codes light with a laser-illuminated DOE as shown in Figure 1. The light transport of an active stereo system can be divided into two parts: one describing the propagation of the laser light into the scene with the output of the illumination pattern cast onto the scene, and the other describing the illumination returned from the scene to the stereo cameras. We rely on wave optics for the former part and geometric optics for the latter part, comprising the proposed hybrid image formation model.

3.1. Modeling the Projected Illumination Pattern

Simulating light transport from an active stereo illumination module to a scene amounts to computing the illumination pattern projected onto the scene from the laser (Figure 1). Relying on wave optics, we represent the light emitted by the laser as amplitude A and phase ϕ at each discrete spatial location x, y sampled with pitch u and with $N \times N$ resolution[1].

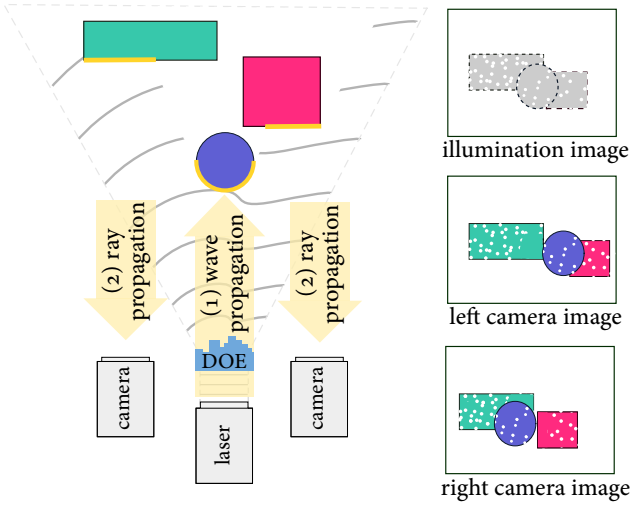

Figure 1. We simulate the illumination image projected by the laser and the DOE using wave optics. We then simulate the stereo images captured by cameras using geometric optics.

Phase Delay on the DOE. The phase of the emitted light wave is modulated when it passes through the DOE by ϕ_{delay} as $\phi \leftarrow \phi + \phi_{\text{delay}}$. The phase delay ϕ_{delay} is related to the height of the DOE h, the wavelength of the light λ, and the refractive index of the DOE for that wavelength η_λ, that is

$$\phi_{\text{delay}} = \frac{2\pi(\eta_\lambda - 1)}{\lambda} h. \tag{1}$$

Far-field Wave Propagation. Next, the light wave modulated by the DOE propagates into the scene. We model this propagation using Fraunhofer far-field wave propagation because we assume that scene depth ranges from $0.4\,\text{m}$ to $3\,\text{m}$ which is sufficiently larger than the wave spatial extent $uN = 1\,\text{mm}$ [14]. We implement this propagation operation by computing the Fourier transform $\mathcal{F}$ of the complex-valued light wave U of amplitude A and phase ϕ

$$U' \leftarrow \mathcal{F}(U), \tag{2}$$

where U' is the propagated complex light wave. Finally, the illumination pattern P in the scene is the intensity of the propagated light wave, a squared magnitude of U'

$$P \leftarrow |U'|^2. \tag{3}$$

The resolution of the pattern P remains the same as that of U, while the physical pixel pitch v of the pattern P changes accordingly as $v = \frac{\lambda z}{uN}$, where z is the propagation distance [14]. Refer to the Supplemental Document for the simulated illumination patterns corresponding to existing DOE designs.

[1] $u = 1\,\mu\text{m}$ and $N = 1000$ in our experiments.

Sampling the Illumination Pattern. A pixel in the simulated illumination image P has the physical width of $v = \frac{\lambda z}{uN}$ at a scene depth z. At the same time, a camera pixel maps to a width of $\frac{p}{f}z$ at the scene depth z via perspective unprojection, where f is the camera focal length, and p is the pixel pitch of the camera. We resample the illumination image P to have the same pixel pitch as a camera pixel pitch. We compute the corresponding scale factor as follows

$$\frac{\text{camera pixel size}}{\text{illumination pattern pixel size}} = \frac{\frac{p}{f}z}{\frac{\lambda}{uN}z} = \frac{puN}{f\lambda}. \quad (4)$$

The scale factor $\frac{puN}{f\lambda}$ is applied to the illumination image $P \leftarrow \mathsf{resample}(P, \frac{puN}{f\lambda})$, where $\mathsf{resample}$ is the bicubic resampling operator.

Note that the depth dependency for the pixel sizes for the illumination pattern and the camera disappears in the scaling factor, meaning that the scale factor is independent of the propagation distance of the light. This indicates that the illumination pattern P can be applied to any scene regardless of its depth composition, which facilitates efficient simulation of the light transport.

3.2. Synthesis of Stereo Images

Once the illumination image P is computed, we then simulate stereo images. While wave optics can describe this procedure using Wigner distribution functions and far-field wave propagation, this would be prohibitively expensive for the proposed end-to-end optimization procedure, which requires tens of thousands of iterations, each triggering multiple forward simulations. Instead, we use a geometric-optics model representing light using intensity only, instead of both phase and amplitude as in wave optics.

Light-matter Interaction and Measurement. Given the illumination image P at the viewpoint of the illumination module, we next simulate the light-matter interaction and sensor measurement by the stereo cameras. In the following model, we use disparity maps $D^{\text{L/R}}$, reflectance maps $I^{\text{L/R}}$, and occlusion masks $O^{\text{L/R}}$ at the left and the right camera viewpoints. Occlusion masks $O^{\text{L/R}}$ describe the visibility at the viewpoints of the left/right camera with respect to the illumination module.

We first warp the illumination image P to the left and the right camera viewpoints using the disparity $D^{\text{L/R}}$. We incorporate the occlusion maps $O^{\text{L/R}}$ through element-wise multiplication with the warped images, resulting in the final illumination images seen at the stereo camera viewpoints (P^{L} and P^{R}), that is,

$$P^{\text{L/R}} = O^{\text{L/R}} \odot \mathsf{warp}(P, D^{\text{L/R}}), \quad (5)$$

where $\odot$ is the element-wise product and the operator warp warps the illumination image P by the disparity $D^{\text{L/R}}$.

We then compute scene response and sensor measurement using a Lambertian reflectance model. We implement imaging parameters including sensor clipping, signal-independent Gaussian noise, camera exposure, illumination power, and ambient illumination. Altogether, this is described by

$$J^{\text{L/R}} = \sigma(\gamma(\alpha + \beta P^{\text{L/R}})I^{\text{L/R}} + \eta), \quad (6)$$

where $J^{\text{L/R}}$ are the simulated captured images for the left and the right camera viewpoints. The term γ is the scalar describing exposure and the sensor's spectral quantum efficiency, α is the ambient light, β is the power of the laser illumination, η is Gaussian noise, and σ is the intensity-cropping function.

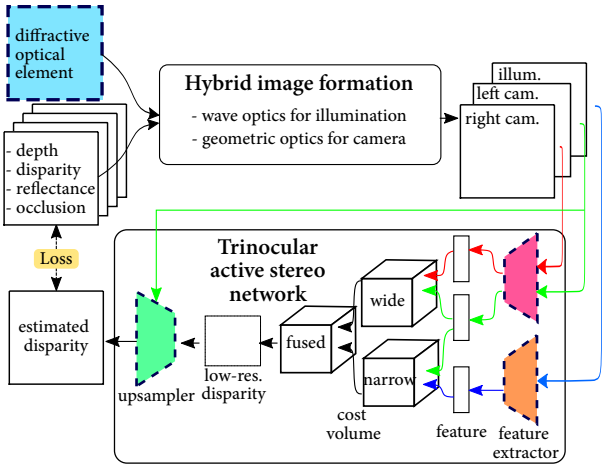

Figure 2. The proposed hybrid image formation model simulates the stereo images from which we reconstruct a depth map using a trinocular network. The loss is backpropagated to both the DOE and the network, enabling joint optimization. Dotted boxes indicate optimization parameters.

4. Trinocular Active Stereo Network

We depart from existing active stereo architectures that take stereo images or a single illumination image as inputs [55, 38]. Instead, we exploit the fact that an active stereo system provides stereo cues between the cameras but also the illumination and camera pairs. Specifically, we consider two baseline configurations in our active stereo camera: a narrow-baseline configuration between the illumination module and either of the two cameras, and one wide-baseline pair consisting of the left and right cameras. To take advantage of these two different baselines, we propose the following trinocular active stereo network, which is illustrated in Figure 2.

Reconstruction Network. The proposed reconstruction network receives the following inputs: a left-camera image $\mathbf{x}_{\text{L}}$, a right-camera image $\mathbf{x}_{\text{R}}$, and an illumination image $\mathbf{x}_{\text{illum}}$. During the training phase, our image formation model synthetically generates these trinocular inputs; during real-world testing, we directly use the calibrated sensor

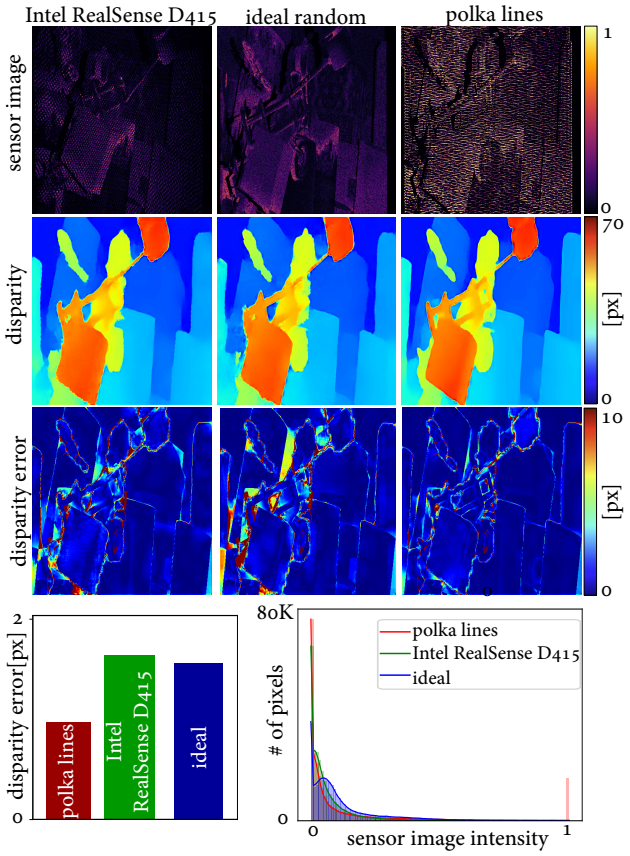

Figure 3. We evaluate our learned illumination pattern in simulation and we outperform the hand-crafted illumination pattern (Intel RealSense D415) and the ideal random pattern. Our learned Polka Line pattern effectively focuses energy to promote feature matching. The example shown here features an indoor environment.

inputs.

The proposed network first extracts feature tensors $\mathbf{y}_{\text{L/R/illum}}$ of the three input images using two convolutional encoders: FE_{cam} for the camera images and FE_{illum} for the illumination image, that is

$$\begin{aligned} \mathbf{y}_{\text{L}} &= \text{FE}_{\text{cam}}(\mathbf{x}_{\text{L}}), \mathbf{y}_{\text{R}} = \text{FE}_{\text{cam}}(\mathbf{x}_{\text{R}}), \\ \mathbf{y}_{\text{illum}} &= \text{FE}_{\text{illum}}(\mathbf{x}_{\text{illum}}). \end{aligned} \tag{7}$$

Next, we construct trinocular cost volumes for two separate baselines. We define a feature cost volume C_{wide} for the wide-baseline pair as

$$C^d_{\text{wide}}(x,y) = \mathbf{y}_{\text{L}}(x,y) - \mathbf{y}_{\text{R}}(x-d,y), \tag{8}$$

where d is a disparity candidate. Similarly, the narrow-baseline cost volume is defined between the left-camera features $\mathbf{y}_{\text{L}}$ and the illumination features $\mathbf{y}_{\text{illum}}$ as

$$C^d_{\text{narrow}}(x,y) = \mathbf{y}_{\text{L}}(x,y) - \mathbf{y}_{\text{illum}}(x-d,y). \tag{9}$$

We fuse the two cost volumes into a single cost volume

$$C^d_{\text{fused}} = C^d_{\text{wide}} + C^{\hat{d}}_{\text{narrow}}, \tag{10}$$

where $\hat{d} = d\frac{b_{\text{wide}}}{b_{\text{narrow}}}$ is the disparity scaled by the ratio between the wide baseline and the narrow baseline. Per-pixel disparity probability is computed using a soft-max layer, followed by disparity regression on the obtained probability resulting from the low-resolution disparity estimate [55]. Finally, an edge-aware convolutional upsampler estimates a disparity map $D^{\text{L}}_{\text{est}}$ for the left camera viewpoint at the original resolution. For network details, we refer the reader to the Supplemental Document.

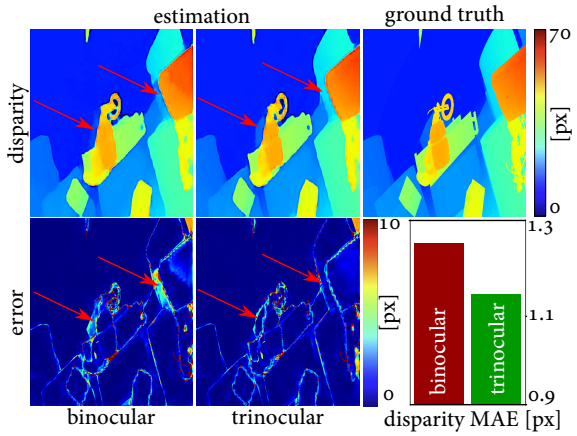

Figure 4. The proposed trinocular reconstruction approach is more robust at object boundaries than conventional binocular methods, as it exploits cues between several camera and illumination pairs in a single active stereo system.

Joint Learning. Denoting the network parameters as θ and the phase delay for the DOE as ϕ_{delay}, we solve the following end-to-end joint optimization problem

$$\underset{\phi_{\text{delay}},\theta}{\text{minimize}}\, \mathcal{L}_{\text{s}}(D^{\text{L}}_{\text{est}}(\phi_{\text{delay}},\theta), D^{\text{L}}), \tag{11}$$

where $\mathcal{L}_{\text{s}} = \mathsf{MAE}$ is the mean-absolute-error loss of the estimated disparity supervised by the ground-truth disparity D^{L}. Note that solving this optimization problem using stochastic gradient methods is only made possible by formulating the proposed image formation model and reconstruction method as fully differentiable operations. We also incorporate varying ambient illumination conditions into our learning framework by controlling the following simulation parameters: ambient light power α and scalar γ in Equation (6). We train three separate models for different illumination configurations of generic, indoor, and outdoor environments. For details, we refer the reader to the Supplemental Document.

Dataset. Our method requires an active-stereo dataset of disparity maps $D^{\text{L/R}}$, NIR reflectance maps $I^{\text{L/R}}$, and occlusion masks $O^{\text{L/R}}$ at the left and the right camera viewpoints. To obtain this dataset, we modify a synthetic passive-stereo RGB dataset [31] which provides disparity maps $D^{\text{L/R}}$ but not the NIR reflectance maps $I^{\text{L/R}}$ and the occlusion masks $O^{\text{L/R}}$. We obtain the NIR reflectance maps $I^{\text{L/R}}$ from the RGB stereo images using the RGB-inversion method from [15]. Next, we compute the occlusion masks $O^{\text{L/R}}$ of the

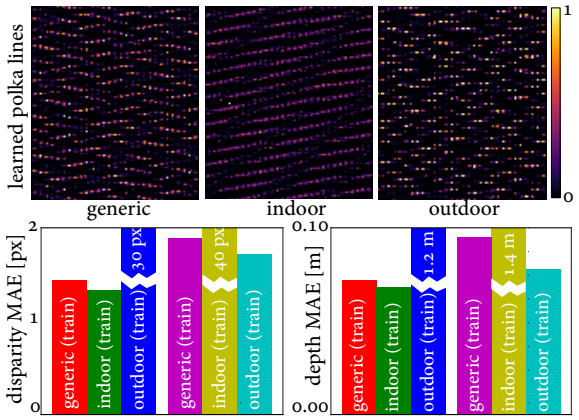

Figure 5. By changing simulation parameters, the proposed end-to-end optimization method can learn illumination patterns tailored to indoor, outdoor, and generic environments.

stereo cameras with respect to the illumination module. We horizontally shrink the stereo occlusion masks by half since the illumination module lies halfway between the stereo pair. Finally, we resize the images to the same resolution as the illumination images.

5. Self-supervised Finetuning

To compensate for fabrication inaccuracies of the optimized DOE and the domain gap between the simulated training images and the real captures, we finetune our reconstruction network using a real-world dataset captured by our prototype. To this end, we capture left and right IR image pairs $J^{\mathrm{L/R}}$ and obtain the illumination images $P^{\mathrm{L/R}}$ by projecting patterns onto a diffuse textureless wall. However, for the disparity maps and the occlusion masks, it is challenging to obtain corresponding ground truths in the real world. Therefore, we adopt the self-supervised learning approach previously proposed in [57, 55].

The key idea in the self-supervised training step is to find disparity maps $D_{\mathrm{est}}^{\mathrm{L/R}}$ and validity maps $V_{\mathrm{est}}^{\mathrm{L/R}}$ that provide the optimal reconstruction of the stereo images $J^{\mathrm{L/R}}$ by warping the other images $J^{\mathrm{L/R}}$ with the disparity $D_{\mathrm{est}}^{\mathrm{L/R}}$ in consideration of the validity $V_{\mathrm{est}}^{\mathrm{L/R}}$. The validity maps are defined as the opposite of the occlusion maps $V_{\mathrm{est}}^{\mathrm{L/R}} = 1 - O_{\mathrm{est}}^{\mathrm{L/R}}$. In addition to the reconstruction network described in the previous section, we introduce a validation network that estimates the validation maps. $V_{\mathrm{est}}^{\mathrm{L/R}}$ to account for occlusion. For the loss functions, $\mathcal{L}_{\mathrm{u}}$ encourages the network to estimate disparity maps that reconstruct one stereo view from the other view through disparity warping. $\mathcal{L}_{\mathrm{v}}$ is the regularization loss for the validity masks $V_{\mathrm{est}}^{\mathrm{L/R}}$ [55, 38]. $\mathcal{L}_{\mathrm{d}}$ is the disparity smoothness loss. We train the network parameters of the trinocular reconstruction network and the validation network on the captured stereo images and the illumination image of the prototype. At the inference time, we mask out the disparity estimates of pixels with low validity. For further details, refer to the Supplemental Document.

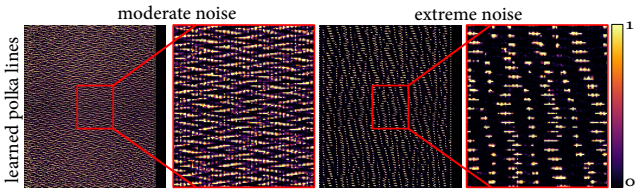

Figure 6. Optimized illumination for different noise levels. For scenarios with strong ambient light, leading to low illumination contrast, the illumination pattern is optimized to have higher-intensity sparse dots than the moderate noise environment.

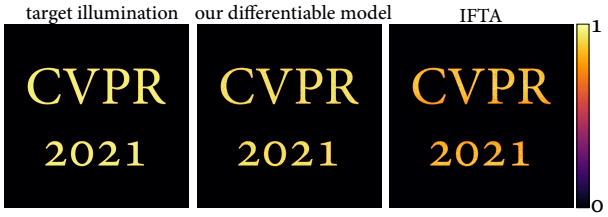

Figure 7. The proposed differentiable image formation can be used for designing a DOE that produces the desired illumination pattern. Our method improves on state-of-the-art iterative FFT methods [11] while allowing for design flexibility, see text.

6. Analysis

Before introducing our experimental prototype system, we first evaluate the proposed end-to-end framework using synthetic data.

Polka Lines Illumination Pattern. We evaluate the effectiveness of our learned illumination, the Polka Lines pattern, by comparing to heuristically-designed patterns: the pseudo-random dot and the regularly spaced dot [1]. For a fair comparison, we use our trinocular network architecture for all patterns and finetune the reconstruction network for each individual illumination pattern. The experiments in Figure 3 validate that the proposed Polka Lines pattern outperforms the conventional patterns in indoor environments. For these synthetic experiments, we ensure that equal illumination power is used for all illumination patterns. We refer to the Supplemental Document for analysis in outdoor environments. The proposed Polka Lines design is the result of the proposed optimization method. We can interpret the performance of this pattern by analyzing the structure of the Polka Lines patterns compared to heuristic patterns. First, each dot in a line of dots has varying intensity levels, in contrast to the constant-intensity heuristic patterns. We attribute the improved performance in large dynamic ranges to these varying dot intensities. Second, the orientations of Polka Lines are locally varying, which is a discriminative feature for correspondence matching. We refer to the Supplemental Document for further discussion.

Trinocular Reconstruction Ablation Study. We validate our trinocular reconstruction method by comparing it to binocular methods such as Zhang et al.[56]. We build a baseline model that ingests only binocular inputs of stereo camera images by removing the illumination feature ex-

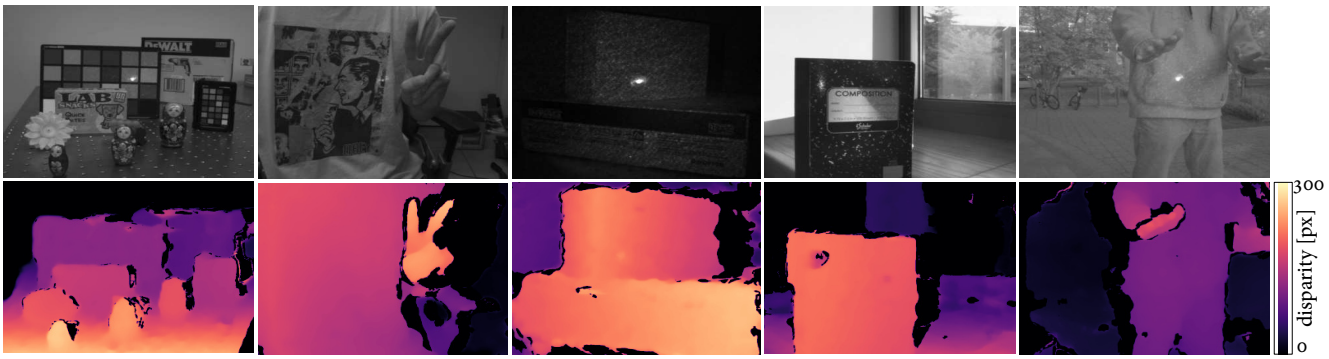

Figure 8. The described system acquires accurate disparity for challenging scenes. We show here examples containing complex objects including textureless surface under diverse environments from indoor illumination to outdoor sunlight.

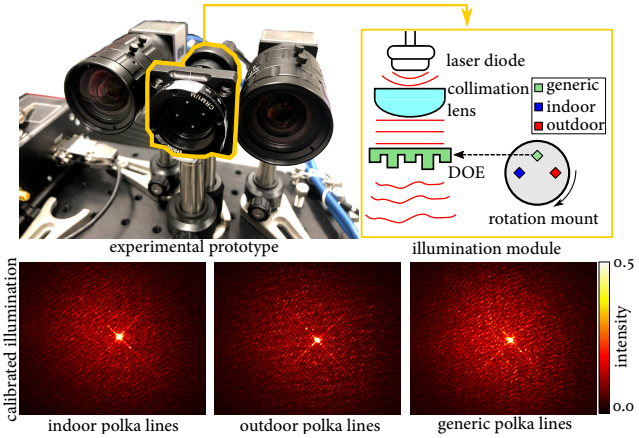

Figure 9. The proposed prototype system consists of stereo NIR cameras and an illumination module, where laser light is collimated and modulated by a DOE. We fabricated three DOEs designed for generic, indoor, and outdoor environments that can be switched by a rotational mount. Calibrated illumination images closely resemble our simulation; a dense low-intensity dot pattern for the indoor, a sparse high-intensity dot pattern for the outdoor, a dense varying-intensity dot pattern for the generic environment.

tractor. Figure 4 shows that the binocular reconstruction method struggles, especially in occluded regions, where the proposed trinocular approach provides stable estimates.

Environment-specific Illumination Design. Our end-to-end learning method readily facilitates the design of illumination patterns tailored to specific environments by changing the environment parameters in Equation (6) and solving Equation (11). We vary the ambient power α and the laser power β to simulate indoor, outdoor, and hybrid "generic" environments[2]. Figure 5 demonstrates that the illumination pattern becomes dense with low-intensity dots in the indoor case for dense correspondence, whereas the outdoor environment promotes a sparse pattern with high-intensity dots that stand out from the ambient light. In the generic environment, we obtain "Polka Lines" with varying intensities from low to high. We also evaluate the proposed method for

[2]We vary the parameter values depending on the environments: indoor ($\alpha = 0.0, \beta = 1.5$), outdoor ($\alpha = 0.5, \beta = 0.2$), generic ($\alpha \in [0, 0.5], \beta \in [0.2, 1.5]$)

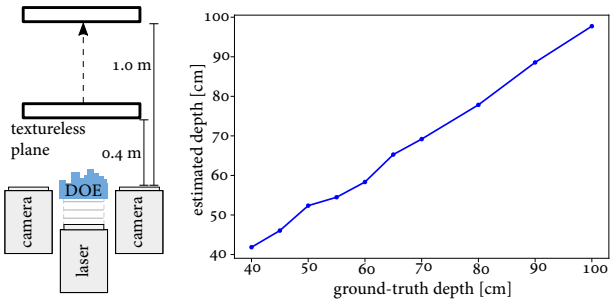

Figure 10. The experimental prototype accurately reconstructs the depth of a textureless plane at distances from 0.4 m to 1.0 m.

two different noise levels, e.g., under strong ambient illumination, using the standard deviation values of 0.02 and 0.6 for the Gaussian noise term η. Figure 6 shows that the illumination pattern becomes sparse with high intensity dotted lines for the severe noise.

DOE Phase Profile Design. We can repurpose the proposed method to design a DOE that produces a target far-field illumination pattern when illuminated by a collimated beam. Designing DOEs for structured illumination has applications beyond active stereo, including anti-fraud protection, projection marking, and surface inspection [48]. Figure 7 shows that we obtain reconstruction quality comparable to state-of-the-art iterative FFT methods [11]. One benefit of using our framework for DOE design is its flexibility. For example, any additional phase-changing optical element can readily be incorporated into the image formation model. Also, additional loss functions can be imposed, e.g., enforcing smoothness of the DOE to reduce potential fabrication inaccuracies. We refer to the Supplemental Document for the optimization details.

7. Experimental Prototype Results

Experimental Prototype. Figure 9 shows our experimental prototype along with captures of the proposed Polka Lines illumination pattern variants. We implement the proposed system with two NIR cameras (Edmund Optics 37-327) equipped with the objective lenses of 6 mm focal length (Edmund Optics 67-709). The pixel pitch of the

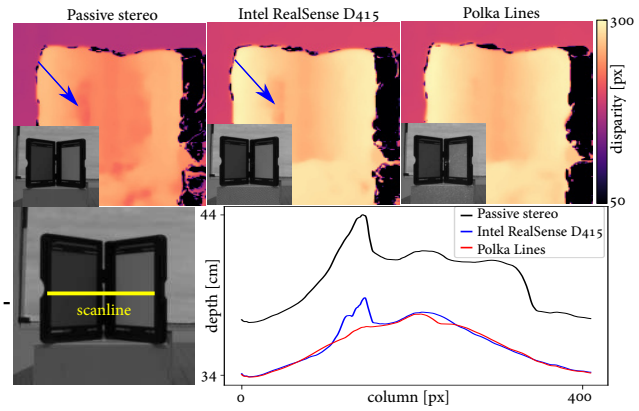

Figure 11. The learned illumination pattern with varying-intensity dots outperforms passive stereo and the commercial hand-engineered pattern (Intel RealSense D415) for high dynamic range scene conditions. Blue arrows indicate estimation artifacts. We capture a V-shaped reflectance target (x-rite Pro Photo Kit).

cameras is $5.3\,\mu\text{m}$, and the stereo baseline is $55\,\text{mm}$. We employ a NIR laser with a center wavelength 850 nm, and beam diameter of 1 mm. We use a laser diode (Thorlabs L850P200), a laser diode socket (Thorlabs S7060R), a collimation lens (Thorlabs LT200P-B), and a laser driver (Thorlabs KLD101). We fabricate the optimized DOE with a 16-level photolithography process. For fabrication details, we refer to the Supplemental Document. The illumination pattern from the fabricated DOE exhibits undiffracted zeroth-order components that are superposed with the diffracted pattern. While commercial mass-market lithography is highly optimized, our small-batch manual lithography did not meet the same fabrication accuracy. Although the fabrication accuracy is below commercial DOEs with high diffraction efficiency, the measured illumination patterns match their synthetic counterparts.

Depth Reconstruction. We measure the depth accuracy of our prototype system by capturing planar textureless objects at known distances as shown in Figure 10. The estimated depth using the Polka Lines pattern closely matches the ground truth, with a mean absolute error of 1.4 cm in the range from 0.4 m to 1 m. We demonstrate qualitative results on diverse real-world scenes in Figure 8, which includes complex objects, dynamic hand movement, textureless objects without ambient light, objects in sunlight, and moving person in dynamic outdoor environments. We showcase video-rate depth imaging in the Supplemental Video.

Comparison. We compare our learned Polka Lines pattern with the commercial Intel RealSense D415 pattern in Figure 11. The average illumination intensity of the Intel pattern is adjusted to match that of the proposed system via radiometric calibration using an integrating sphere (Thorlabs S142C). Figure 11 shows that our intensity-varying pattern is more robust to high dynamic range scenes than the Intel pattern, thanks to denser Polka dot patterns with a larger

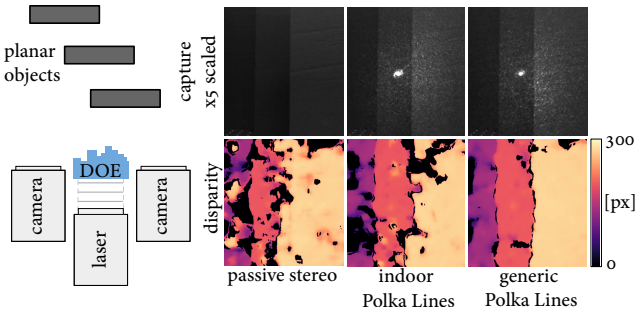

Figure 12. We capture a scene with low-reflectance planar objects. While passive stereo suffers at the textureless surface, the proposed learned illumination enables effective depth reconstruction. The DOE learned for the generic environment contains a wider range of pattern intensities than the DOE learned for indoor scenes, enabling better depth estimation for these objects.

dynamic range. We note that the Intel pattern is of high fabrication quality and does not exhibit a severe zeroth-order component (as does our fabricated DOE). We validate our learned Polka Line variants for generic environments and indoor environments in Figure 12. The generic variant features a wide intensity range of dots, resulting in accurate reconstruction for low-reflectance objects.

8. Conclusion

We introduce a method for learning an active stereo camera, including illumination, capture, and depth reconstruction. Departing from hand-engineered illumination patterns, we learn novel illumination patterns, the Polka Lines patterns, that provide state-of-the-art depth reconstruction and insights on the function of structured illumination patterns under various imaging conditions. To realize this approach, we introduce a hybrid image formation model that exploits both wave optics and geometric optics for efficient end-to-end optimization, and a trinocular reconstruction network that exploits the trinocular depth cues of active stereo systems. The proposed method allows us to design environment-specific structured Polka Line patterns tailored to the camera and scene statistics. We validate the effectiveness of our approach with comprehensive simulations and with an experimental prototype, outperforming conventional hand-crafted patterns across all tested scenarios. In the future, combined with a spatial light modulator, the proposed method may not only allow for ambient illumination specific patterns, but also semantically driven dynamic illumination patterns that adaptively increase depth accuracy.

Acknowledgements

The authors are grateful to Ethan Tseng and Derek Nowrouzezahrai for fruitful discussions. Felix Heide was supported by an NSF CAREER Award (2047359) and a Sony Young Faculty Award.

References

[1] *Intel® RealSense™ Depth Camera D415 https://www.intelrealsense.com/depth-camera-d415/ Accessed June 20, 2020.*

[2] udepth: Real-time 3d depth sensing on the pixel 4. Accessed September 19, 2020.

[3] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.

[4] Supreeth Achar, Joseph R Bartels, William L'Red' Whittaker, Kiriakos N Kutulakos, and Srinivasa G Narasimhan. Epipolar time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 36(4):1–8, 2017.

[5] Brian F. Aull, Andrew H. Loomis, Douglas J. Young, Richard M. Heinrichs, Bradley J. Felton, Peter J. Daniels, and Deborah J. Landers. Geiger-mode avalanche photodiodes for three-dimensional imaging. 13(2):335–349, 2002.

[6] Seung-Hwan Baek, Diego Gutierrez, and Min H Kim. Birefractive stereo imaging for single-shot depth acquisition. *ACM Transactions on Graphics*, 35(6):194, 2016.

[7] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, volume 11, pages 1–11, 2011.

[8] Ayan Chakrabarti. Learning sensor multiplexing design through back-propagation. In *Advances in Neural Information Processing Systems*, pages 3081–3089, 2016.

[9] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[10] H Dammann and K Görtler. High-efficiency in-line multiple imaging by means of multiple phase holograms. *Optics communications*, 3(5):312–315, 1971.

[11] Pei-Qin Du, Hsi-Fu Shih, Jenq-Shyong Chen, and Yi-Shiang Wang. Design and verification of diffractive optical elements for speckle generation of 3-d range sensors. *Optical Review*, 23(6):1017–1025, 2016.

[12] Sean Ryan Fanello, Julien Valentin, Christoph Rhemann, Adarsh Kowdle, Vladimir Tankovich, Philip Davidson, and Shahram Izadi. Ultrastereo: Efficient learning-based matching for active stereo systems. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6535–6544. IEEE, 2017.

[13] 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.

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

[15] Tobias Gruber, Frank Julca-Aguilar, Mario Bijelic, and Felix Heide. Gated2depth: Real-time dense lidar from gated images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1506–1516, 2019.

[16] Mohit Gupta, Qi Yin, and Shree K Nayar. Structured light in sunlight. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 545–552, 2013.

[17] 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.

[18] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Patrice Horaud. *Time-of-flight cameras: principles, methods and applications*. Springer Science & Business Media, 2012.

[19] Felix Heide, Steven Diamond, David B Lindell, and Gordon Wetzstein. Sub-picosecond photon-efficient 3d imaging using single-photon sensors. *Scientific reports*, 8(1):1–8, 2018.

[20] Felix Heide, Wolfgang Heidrich, Matthias Hullin, and Gordon Wetzstein. Doppler time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.

[21] Steven Hickson, Stan Birchfield, Irfan Essa, and Henrik Christensen. Efficient hierarchical graph-based segmentation of RGBD videos. pages 344–351, 2014.

[22] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2007.

[23] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011.

[24] Kaizhang Kang, Cihui Xie, Chengan He, Mingqi Yi, Minyi Gu, Zimin Chen, Kun Zhou, and Hongzhi Wu. Learning efficient illumination multiplexing for joint capture of reflectance and shape. *ACM Trans. Graph.*, 38(6):165–1, 2019.

[25] Andreas Kolb, Erhardt Barth, Reinhard Koch, and Rasmus Larsen. Time-of-flight cameras in computer graphics. In *Computer Graphics Forum*, volume 29, pages 141–159. Wiley Online Library, 2010.

[26] Adarsh Kowdle, Christoph Rhemann, Sean Fanello, Andrea Tagliasacchi, Jonathan Taylor, Philip Davidson, Mingsong Dou, Kaiwen Guo, Cem Keskin, Sameh Khamis, et al. The need 4 speed in real-time dense visual tracking. *ACM Transactions on Graphics (TOG)*, 37(6):1–14, 2018.

[27] Robert Lange. 3D time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology. 2000.

[28] 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–es, 2007.

[29] Julio Marco, Quercus Hernandez, Adolfo Muñoz, Yue Dong, Adrian Jarabo, Min H Kim, Xin Tong, and Diego Gutierrez.

Deeptof: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. *ACM Transactions on Graphics (ToG)*, 36(6):1–12, 2017.

[30] Manuel Martinez and Rainer Stiefelhagen. Kinect unleashed: Getting control over high resolution depth maps. In *MVA*, pages 247–250, 2013.

[31] N. Mayer, E. Ilg, P. Häusser, 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 *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134.

[32] C. Metzler, H. Ikoma, Y. Peng, and G. Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *Proc. CVPR*, 2020.

[33] Andreas Meuleman, Seung-Hwan Baek, Felix Heide, and Min H. Kim. Single-shot monocular rgb-d imaging using uneven double refraction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[34] Yinxiao Miao, Yongshun Zhao, Huiping Ma, Minwei Jiang, Jie Lin, and Peng Jin. Design of diffractive optical element projector for a pseudorandom dot array by an improved encoding method. *Applied Optics*, 58(34):G169–G176, 2019.

[35] Elias Nehme, Daniel Freedman, Racheli Gordon, Boris Ferdman, Lucien E Weiss, Onit Alalouf, Reut Orange, Tomer Michaeli, and Yoav Shechtman. Deepstorm3d: dense three dimensional localization microscopy and point spread function design by deep learning. *arXiv preprint arXiv:1906.09957v2*, 2019.

[36] Bingbing Ni, Gang Wang, and Pierre Moulin. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *Consumer Depth Cameras for Computer Vision*, pages 193–208. Springer, 2013.

[37] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[38] Gernot Riegler, Yiyi Liao, Simon Donne, Vladlen Koltun, and Andreas Geiger. Connecting the dots: Learning representations for active monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7624–7633, 2019.

[39] Sean Ryan Fanello, Christoph Rhemann, Vladimir Tankovich, Adarsh Kowdle, Sergio Orts Escolano, David Kim, and Shahram Izadi. Hyperdepth: Learning depth from structured light without matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5441–5450, 2016.

[40] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002.

[41] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. volume 1, 2003.

[42] John Sell and Patrick O'Connor. The xbox one system on a chip and kinect sensor. *IEEE Micro*, 34(2):44–53, 2014.

[43] 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.

[44] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[45] Shuochen Su, Felix Heide, Gordon Wetzstein, and Wolfgang Heidrich. Deep end-to-end time-of-flight imaging. pages 6383–6392, 2018.

[46] Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. Learning rank-1 diffractive optics for single-shot high dynamic range imaging. *IEEE CVPR*, 2020.

[47] Quilin Sun, Jian Zhang, Xiong Dun, Bernard Ghanem, Yifan peng, and Wolfgang Heidrich. End-to-end learned, optically coded super-resolution spad camera. *ACM Transactions on Graphics (TOG)*, 39, 2020.

[48] Jari Turunen and Frank Wyrowski. *Diffractive optics for industrial and commercial applications*. 1998.

[49] Ralf Vandenhouten, Andreas Hermerschmidt, and Richard Fiebelkorn. Design and quality metrics of point patterns for coded structured light illumination with diffractive optical elements in optical 3d sensors. In *Digital Optical Technologies 2017*, volume 10335, page 1033518. International Society for Optics and Photonics, 2017.

[50] Lizhi Wang, Tao Zhang, Ying Fu, and Hua Huang. Hyperreconnet: Joint coded aperture optimization and image reconstruction for compressive hyperspectral imaging. *IEEE Transactions on Image Processing*, 28(5):2257–2270, May 2019.

[51] George M. Williams. Optimization of eyesafe avalanche photodiode lidar for automobile safety and autonomous navigation systems. 56(3):1 – 9 – 9, 2017.

[52] Jiamin Wu, Bo Xiong, Xing Lin, Jijun He, Jinli Suo, and Qionghai Dai. Snapshot hyperspectral volumetric microscopy. *Scientific Reports*, 6:24624, 2016.

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

[54] Yicheng Wu, Vivek Boominathan, Xuan Zhao, Jacob T Robinson, Hiroshi Kawasaki, Aswin Sankaranarayanan, and Ashok Veeraraghavan. Freecam3d: Snapshot structured light 3d with freely-moving cameras. In *European Conference on Computer Vision*, pages 309–325. Springer, 2020.

[55] Yinda Zhang, Sameh Khamis, Christoph Rhemann, Julien Valentin, Adarsh Kowdle, Vladimir Tankovich, Michael Schoenberg, Shahram Izadi, Thomas Funkhouser, and Sean Fanello. Activestereonet: End-to-end self-supervised learning for active stereo systems. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 784–801, 2018.

[56] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

[57] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2017.