Single-shot Hyperspectral-Depth Imaging with Learned Diffractive Optics - Supplemental Document -

Seung-Hwan Baek[*†] Hayato Ikoma[‡] Daniel S. Jeon[*] Yuqi Li[§] Wolfgang Heidrich[§]
Gordon Wetzstein[‡] Min H. Kim[*]

[*]KAIST [†]Princeton University [‡]Stanford University [§]KAUST

This document provides additional details and results.

1. Network Architecture

For a given input image either generated by our wave-optics simulator or captured by our camera prototype, we reconstruct a hyperspectral image and a depth map using a convolutional neural network (CNN). Figure 2 shows our network architecture. Our CNN architecture is inspired by a U-net [10] with the difference of using two decoders instead one. The network takes a RGB sensor image as inputs. During training, 256×256 patches are fed through a basic block layer that consists of two pairs of a 3×3 convolutional layer with batch normalization and PReLU. We use max pooling after each of this basic block to implement the encoder architecture. The spatial resolution is reduced by half at each pooling stage and the number of channels is doubled. We obtain the feature with 1024 channels after the encoder. A counterpart decoder has the similar design with the encoder using a basic block at each spatial resolution and a transposed convolution for upscaling. We have skip connections from the encoders to the decoders at each resolution. The final convolutional layers enable us to match the number of output channels to the hyperspectral image and the depth map. For the spectral decoder, we added the spectrally upsampled image of to the output of the spectral decoder for residual learning.

2. Hyperspectral-Depth Dataset

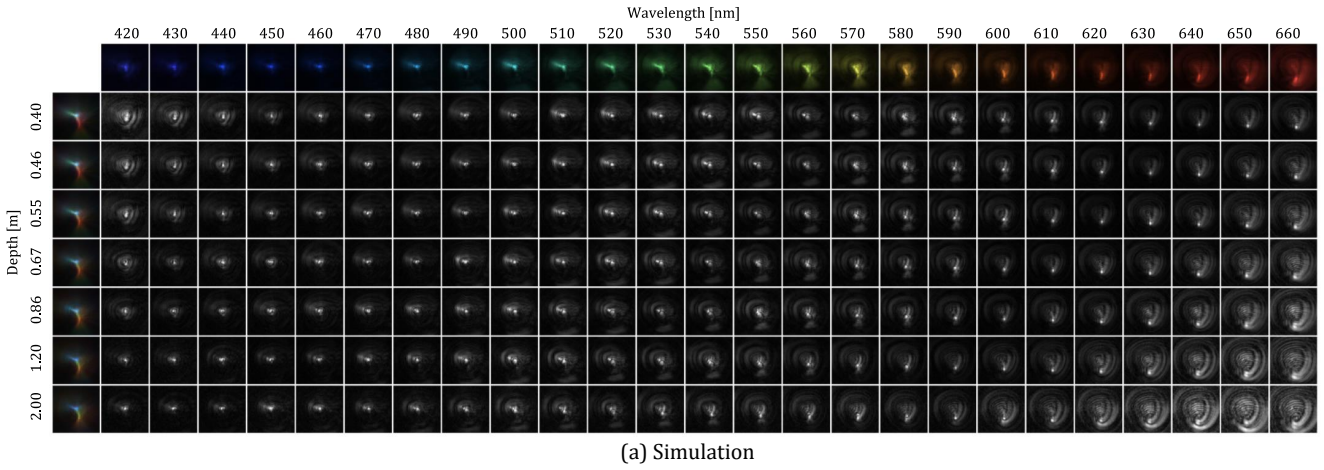
Our HS-D dataset has 16 indoor scenes. Each scene has a triplet of a hyperspectral image, a depth map, and a background mask with pixel-accurate alignment. The hyperspectral image is in reflectance domain, making it effective for spectral augmentation as shown in the main paper. Spectral range starts from 420 to 680 nm in 10 nm intervals, resulting in 27 spectral channels. The depth map has accurate values ranging from 0.4 to 2.0 m obtained by a structured light scanning. The background mask for invalid spectral and depth regions is also provided for selectively choosing valid patches for training. Every image has the spatial resolution of 2824×4240.

Data augmentation for HS-D training. We augment our HS-D dataset for robust learning. First, we spatially scale the images with the factors of 0.25, 0.50, and 1.00. We use bilinear interpolation for hyperspectral images and nearest-neighbor interpolation for depth maps and background masks. Second, we augment depth maps by globally translating the depth values along the z-axis by -0.2 m, 0.0 m, and 0.2 m. Third, we spectrally augment the dataset by multiplying the hyperspectral reflectance images with 29 different CIE standard illuminants, yielding radiance maps under various illuminations. In total, we have ∼20,000 patches.

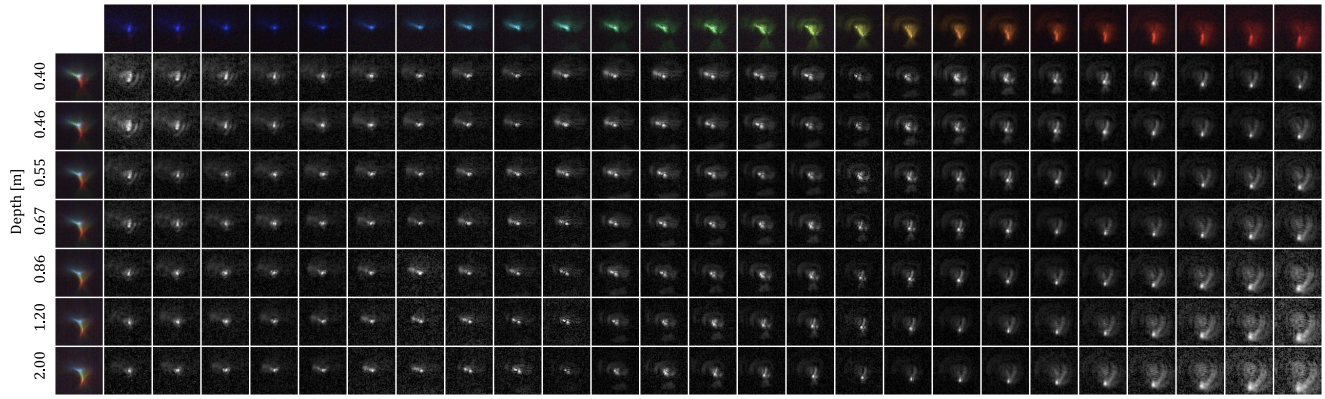
Benchtop combinational system for dataset acquisition. We build a benchtop combinational imaging system to capture our HS-D dataset. We use a projector (EPSON EB-X31) and a liquid-crystal-tunable-filter hyperspectral camera which consists of a machine vision camera (Pointgrey GS3-U3-91S6M-C), a liquid crystal tunable filter (VariSpec LCTF VIS), a relay lens (Sigma A, f/1.4, 50 mm), a collimating lens (Sigma A, f/1.4, 50 mm), and an imaging lens (Nikon, f/2.0, 35 mm). To capture each scene, we illuminate it with a solid-state plasma light source (Thorlabs HPLS-30-4) and capture spectral images with the f-number of 22 from 420 nm to 700 nm by the LCTF modulation. We capture and average the five spectral images for the low wavelengths (420 nm to 450 nm) to mitigate noise. We then turn off the plasma light source and sequentially illuminate the scene with the gray-code patterns using the projector and capture the images with the point gray camera. The LCTF is set to 600 nm for the structured-light capture.

HS-D dataset from raw captures. We obtain dark-level images of the hyperspectral camera by capturing spectral images while blocking incident light to the lens. We subtract the dark levels from every raw capture and denoise it with 3×3 median filtering. In each scene, we have a standard reflectance tile (Spectralon) that provides the calibrated illumination spectrum incident to the scene. We use this measurement to obtain reflectance images.

We estimate depth maps using triangulation from the gray-coded inputs [6]. Note that the structured-light scan-



(a) Simulation



(b) Calibration

Figure 1. (a) Our learned PSF shows depth and spectral dependency, allowing us to acquire both data from a single shot. (b) We calibrate the PSF of the fabricated DOE matched with the simulation counterpart.

ning often provide inaccurate reconstruction for image regions with low albedo. Therefore, we manually mask the black background out and apply depth refinement for the foreground objects [5], resulting in a dense depth map for each scene. The spatial resolution of each hyperspectral-depth image is 2824×4240.

Preprocessing for training and testing. We divide 18 scenes into 13 scenes for training and 5 for testing. We then collect 256×256-sized HS-D patches. Background dominant patches having invalid depth values or too low intensity are excluded for both training and testing. We also ensured that only one of the 13 training scenes includes a ColorChecker in a set of training minibatches in order to avoid overfitting to this target.

3. HS-D Encoded PSF

Our learned PSF changes its shape for spectrum and depth. Figure 1 shows the spectral-depth dependency of the simulated PSF as well as the calibrated PSF obtained from our prototype DOE.

Calibration. We calibrate the PSF for a combination of depth values and spectrum. To this end, we use a solid-state plasma light source (Thorlabs HPLS-30-04) covered with a high-power precision pinhole (Thorlabs P23C) with a 25 μm aperture. In a dark room, we place the illumination module at target distances from ∼0.4 to 2.0 m. We apply spectral filtering in front of the camera using a Varispec LCTF filter in 10 nm intervals and captured HDR hyperspectral images at each depth. We also calibrate the spectral response function of the Canon Mark III camera by measuring its response to the calibrated light source [1].

Diffraction efficiency. Prototype DOEs typically present a low-frequency component in its PSF as zeroth-order diffraction, due to the low diffraction efficiency [8, 4]. We examine the frequency response of our DOE prototype by capturing a black-white scene (Figure 3). As expected, our PSF consists of two-frequency components: one with high frequency and the other with low frequency, where each can be modeled as a Gaussian function of mean (1445/10.74) and standard deviation (1458/312.27) in pixels. We attempt to mitigate such low-frequency degradation of the captured images as detailed in Section 7.

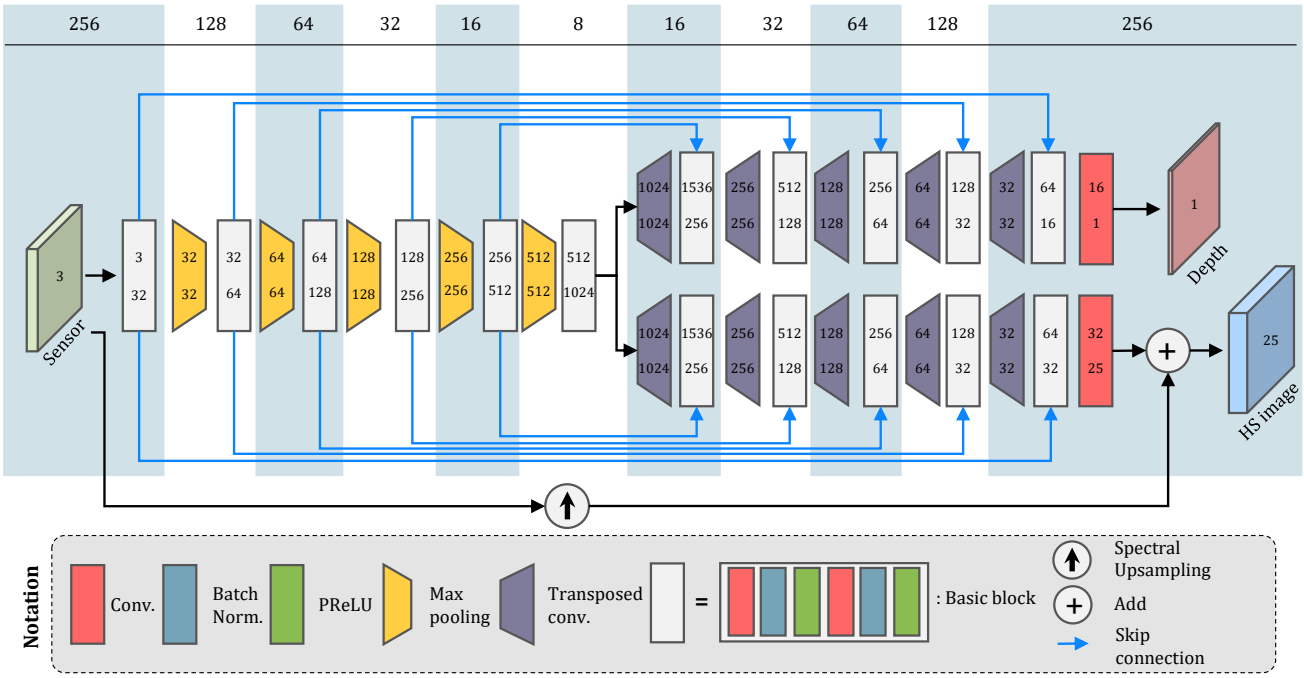


Figure 2. Reconstruction Network. The input to the network is a sensor image with three channels (RGB). The single-encoder-dual-decoder architecture with skip connections enable us to reconstruct a depth map and a hyperspectral image of 25 channels from 420 to 660 nm in 10 nm intervals. For details of the spectral upsampling, we refer to the main paper. The top row shows the width and height of the patch sizes during training. We denote the number of channels in each layer, where the input and output channels are shown in the top and the bottom of each block.

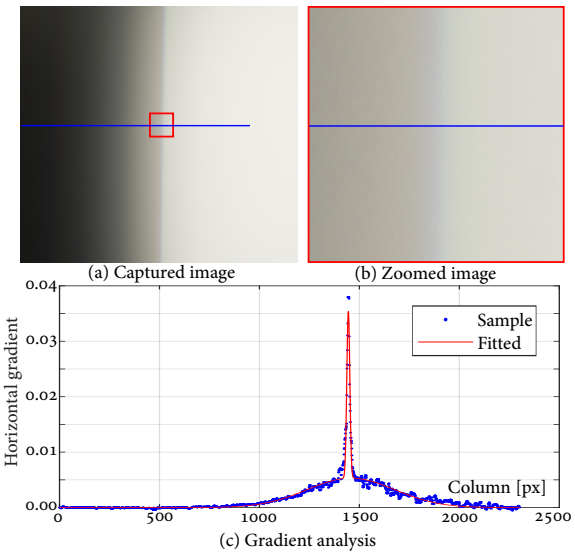


Figure 3. Analyzing the PSF of the prototype DOE. (a-b) Due to the low diffraction efficiency of the DOE prototype, we have a low-frequency component in the PSF in addition to the intended high-frequency component. (c) For a row on the image, we plot the gradient values, revealing the high-frequency and the low-frequency components.

4. Details on Hyperspectral Comparison

With consideration of computational resource and algorithms performance, we adjusted the spatial resolution of experiments. For Jeon et al. [4] and ours, we use the half resolution of the test images in 1412-by-2120. For Baek et al. [1], we reduce the resolution of the input image by one-eighth as their method takes about 45 minutes to process a 353-by-530 hyperspectral image.

5. Details on Depth Comparison

We compared our approach to two other DOE-based depth imaging methods: Wu et al. [13] and (c) Chang et al. [3]. The experimental configurations of these three methods are different including the effective pixel pitch, aperture diameter, the network design, the training dataset, and the camera response function. We implemented a PSF simulator for Chang et al. [3] and simulated PSFs with the same configuration of our prototype. For Wu et al. [13], we obtained the PSFs using the author-provided DOE height map that assumes the pixel pitch of 4.29 um and the aperture size of 0.8 mm. The simulated PSFs are shown in the main paper demonstrating that the simulated PSF shapes match those reported in the original works. Our U-net-based reconstruction network was used to train all DOE designs on our HS-D dataset. Note that the spectral decoder was deactivated in this experiment.

6. DOE initialization

Jointly optimizing a DOE and a CNN for hyperspectral-depth imaging is a challenging, non-convex inverse problem that aims at simultaneously solving multiple traditional problems, including phase retrieval, spectral super-resolution, monocular depth estimation, and deconvolution. This non-convex nature makes it crucial to find a good initialization of the optimization parameters. In particular, the initialization of the DOE has been shown to be important and is specific to a target application [13]. For hyperspectral-depth imaging, we therefore seek to find a proper initialization of the DOE through a Fisher-information-based optimization to obtain the initial DOE height field [11]. Since the Fisher information matrix for the general hyperspectral depth imaging problem is too large to evaluate, we consider a simpler subproblem where we estimate the location and wavelength of a monochromatic point-source emitter from its single RGB image (J_c). Its Fisher information matrix $\mathcal{I}$ then describes the sensitivity of the observed PSF to the spatial emitter positions (p_x, p_y, p_z) and wavelengths (p_λ). When the brightness of the point source is known, the Fisher information matrix under the Gaussian noise model is given as:

$$\mathcal{I}_{ij}\left(\delta\right) = \sum_{c,k} \frac{1}{\sigma^2} \frac{\partial J_c\left(k;\delta,h\right)}{\partial \delta_i} \frac{\partial J_c\left(k;\delta,h\right)}{\partial \delta_j}, \qquad (1)$$

where $\delta = \{p_x, p_y, p_z, p_\lambda\}$, σ is the standard deviation of the Gaussian noise, k is the pixel index, and h is the DOE height field. While the Fisher information depends on the position and wavelength of the point-source emitter, we aim to find a DOE height field that provides high Fisher information for all sources in our design space. To achieve this, we optimize the height of the DOE by minimizing the mean of the A-optimality of the Fisher information matrix over a set of monochromatic point sources located on the optical axis:

$$\underset{h}{\text{minimize}} \frac{1}{N} \sum_{p_\lambda \in \mathbf{\Lambda}} \sum_{p_z \in \mathbf{z}} \mathcal{A}\left(p_z, p_\lambda; h\right), \qquad (2)$$

where $\mathcal{A}$ is the A-optimality, which is the trace of the inverse of the Fisher information matrix $\mathcal{I}$. The design space of the imaging system is characterized by the set of wavelengths $\mathbf{\Lambda}$ and the set of the depth layers $\mathbf{z}$ where the point sources are placed. Since Equation (2) is not a convex problem, we solve it based on stochastic gradient descent optimization, using the Adam optimizer. This optimization itself requires an initialization, for which we choose a conventional Fresnel DOE lens pattern. We set the brightness of the point source so as to ensure the maximum intensity of the captured PSFs of a Fresnel lens is 0.8 of the maximum intensity of the image.

Evaluation. We tested three different initial DOE designs for end-to-end HS-D imaging: the Fresnel lens, the spiral DOE [4], and the Fisher-information-based DOE. Table 1 compares how much the end-to-end optimization process of optics improves the accuracy of reconstructed spectral and depth information for different initializations. Among the three candidates, we chose the Fisher-based initialization as it is superior to other initializations in terms of spectral and depth accuracy.

Initialization		Fresnel	Spiral [4]	Fisher [11]
Spec.	PSNR [dB]	27.96→28.68	26.90→27.67	28.51→**29.31**
	SSIM	0.74→0.78	0.64→0.75	0.79→**0.81**
Depth	RMSE [m]	0.21→**0.19**	0.32 → 0.26	0.23 → 0.20
	MAE [m]	0.15 → 0.12	0.20 → 0.18	0.15 → **0.12**

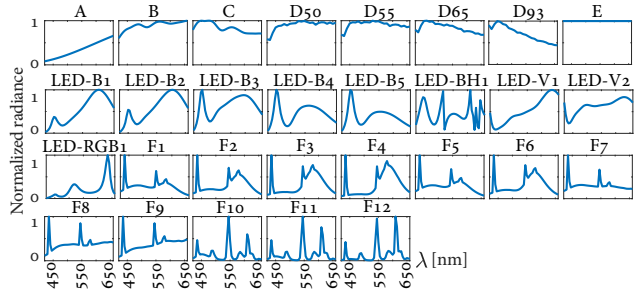
Table 1. DOE initialization for end-to-end learning. We used three different DOE initializations for our end-to-end optimization. The Fisher-initialized DOE optimization is superior to other initializations for spectral and depth reconstruction, and the Fresnel-lens-initialized optimization is the second best option.

7. End-to-end Optimization

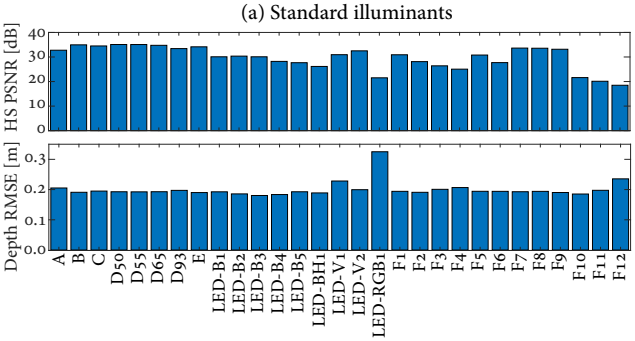
Phase vs. height. Instead of optimizing DOE height, we use its unwrapped phase for the optimization variable. This avoids additional complexity of employing physical-fabrication constraints as a loss function. Specifically, we optimize the unwrapped phase shift ϕ by the DOE at the wavelength of 550 nm. Once the phase is optimized, we apply phase wrapping to ϕ_{wrap} and convert this to a physical DOE height as $h = \frac{\lambda}{2\pi} \frac{\phi_{\text{wrap}}}{(\eta_\lambda - 1)}$.

Training details. We implement end-to-end optimization in Pytorch and uses the Adam optimizer [7]. The total number of network parameters is 39,484,378. The learning rates for the DOE phase and network weights are set as 10^{-4}. We decay the learning rates differently for the DOE and network by 0.1 per 10 epochs and 0.1 per 20 epochs, respectively, following [13]. Once the DOE shape is converged, we fix the DOE and keep training the network for training efficiency. The end-to-end training takes 12 epochs in about 48 hours, after which the reconstruction part was trained for an additional 30 epochs, also taking 48 hours. We use a workstation equipped with a 3.40 GHz Intel i7-3770 CPU, 32 GB of main memory, and an NVIDIA Titan Xp GPU with 12 GB memory. For testing, it takes about 1.45 seconds to reconstruct a hyperspectral-depth image with the resolution of 1412×2120.

Finetuning. After we built the prototype with the fabricated DOE and calibrated the PSF of the prototype, we found that low diffraction efficiency of the real DOE causes a long tail of PSF with low levels of intensity (similar to noise), forming a very large convolution kernel. To meet the requirement of memory footprint in GPU, we excluded the noisy long tail from our real PSF model. It results in com-



(a) Standard illuminants



(b) Spectral and depth accuracy

Figure 4. Reconstruction performance under different illuminants. (a) We evaluate our method on the HS-D test dataset augmented with 29 CIE standard illuminants. (b) Reconstruction accuracy of spectrum and depth is affected by the frequency of the illuminant as observed by degradation at fluorescent illuminants (F10-F12) and an LED illuminant (LED-RGB1).

mon hazy artifacts also observed in previous works of DOE engineering works [4, 12]. We extend a recent approach [9] that mitigates the hazy artifacts from DOE images at a single depth level. Instead, we capture a set of natural spectral images and the prototype camera input at different distance levels, yielding the real-DOE training dataset as shown in Figures 5.

Specifically, once we fabricated the DOE design, we captured 400 natural images (selected from the MIT-Adobe FiveK dataset [2]) displayed on a high-luminance 55-inch display (LG signage 55XS2B, peak luminance: 2,500 cd/m^2) using the real-DOE camera, at 7 different depths from 0.4 to 2.0 m, resulting in 2,800 images in total. See Figures 5(a). At the same time, we captured hyperspectral images using a custom-built hyperspectral camera (a machine vision camera equipped with a liquid-crystal bandpass filter in front of the objective lens.). These hyperspectral images are registered to the images taken by the prototype camera by deriving a set of homography matrices estimated by the checkerboard-calibration target.

By doing so, we can refine the parameters of our reconstruction network using the real-DOE training dataset. As each training patch consists of a constant depth value, we perform patch-wise reconstruction at test time and reconstructs the final output via the mean of overlapping patches. This additional refinement compensates the physical gap

(a) Acquisition setup for refinement (b) thumbnails

Figure 5. We build an acquisition setup to record a pair of HS-D information of natural images (Adobe FiveK), specifically used to mitigate the artifacts by diffraction inefficiency in the real-DOE prototype.

between the synthetically optimized DOE and the fabricated DOE, which causes low diffraction efficiency. See Figure 7.

8. Analysis

Comparison with depth imaging. The experimental configurations of these three methods (Chang et al. [3]/Wu et al. [13]/ours) are all different including the effective pixel pitch (4.29/9.60/6.75 um), aperture diameter (0.800/2.835/3.000 mm), the network design (three different variants of U-net), the training dataset (real RGB-D dataset/synthetic RGB-D dataset/real HS-D dataset), and the camera response function. Therefore, we varied the design parameters of their DOEs only while fixing the other configuration parameters to be the same as ours. We implemented the phase shift of the thin lens for Chang et al. [3] and used the DOE design provided by the authors for Wu et al. [13]. For fair comparison, we use our reconstruction network for all DOE designs. Note that the spectral decoder was deactivated in this experiment.

Illumination. We synthetically evaluate our end-to-end HS-D imaging under 29 different CIE standard illuminants by averaging the hyperspectral PSNR and the depth RMSE values of five different test scenes. Figure 4 shows the results. Our HS-D imaging estimates depth with high accuracy consistently under various illuminations, except for one LED illuminant with a sharp peak near the infrared wavelength (LED-RGB1). This illuminant is almost monochromatic (at the spectral resolution of our system), and hence the images lack the spectral cues needed to infer the depth with high accuracy. We note however, that the depth estimation works well for more natural types of illumination. Our method captures the spectral information with high accuracy under most illuminants in general. Under high-frequency illuminants, such as fluorescent F10, F11, F12, and LED-RGB1, our spectral reconstruction performs suboptimally due to the strong-peak illumination of fluorescent and LED light. We found that our end-to-end

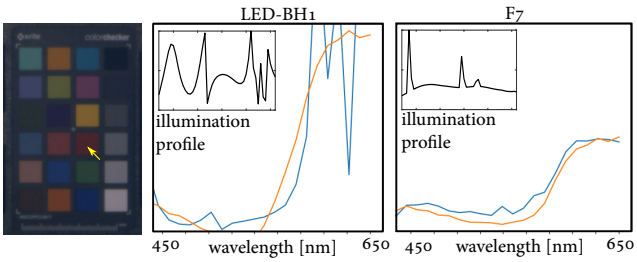


Figure 6. Spectral accuracy on synthetic data under high-frequency illumination of LED-BH1 and F7. The blue and orange curves show ground truth and our reconstruction.

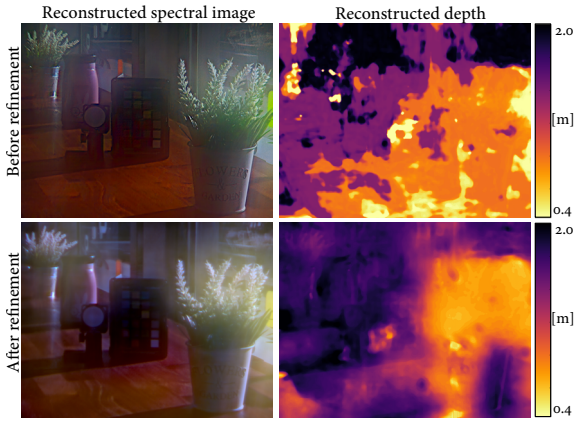


Figure 7. Comparison of reconstruction results with/without the additional refinement process for the real-DOE prototype. The additional refinement of the reconstruction network improves depth accuracy in particular.

HS-D imaging performs robustly under sun, tungsten, general fluorescent, and general LED lights.

Spectral resolution. Our method suffers from high-frequency spectral changes due to the low-frequency nature of hyperspectral-depth PSF kernels. We evaluate our spectral accuracy with a synthetic colorchecker under two high-frequency illuminations of LED-BH1 and F7 in Figure 6. While our method follows the low-frequency trend of spectral signatures, the estimates deviate from the ground truth in terms of spectral details. We refer to Figure 4 in Supplemental Document for quantitative analysis.

Spatial resolution. We evaluate the spatial resolution of our real-prototype results in terms of modulation transfer function (MTF) by capturing a spatial-resolution target as shown in Figure 8. These two input and output images are converted to luminance to compute MTFs. Qualitative and quantitative results show that the spatial resolution is improved by our reconstruction process for the real-DOE prototype.

Computational Efficiency. Our end-to-end design process of optimizing the optimal DOE profile takes place only once at the initial system design step. The required running time of our computational optimization is smaller than most conventional optics design approaches which require tedious semi-manual optimization.

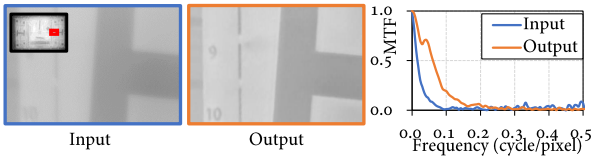


Figure 8. Spatial resolution analysis of input and output spectral images of our real prototype. These two images are converted to luminance to compute MTFs. The MTF of output is clearly improved by our reconstruction network.

End-to-end Learning for Distributions. Our method optimizes a pattern of DOE for given system parameters, such as target depth, wavelength samples, and camera parameters. While this may seem like a drawback requiring optimization for a particular system-design sample, our end-to-end learning method is not fundamentally limited in this respect. Our method can be straightforwardly tweaked based on the end goal to incorporate target "distributions," not a single sample, so that on-average performance can be optimized by randomizing system parameters per each training iteration of the stochastic gradient descent.

9. DOE Fabrication

The DOE height map is parameterized as a bitmap with a resolution of 375×375 features and a pixel pitch of 8 μm, resulting in a DOE aperture of 3 mm. Note that, for fabrication, we upsample the DOE height field to a resolution of 3000×3000 of 1 μm pixel pitch with nearest neighbor interpolation to match with the simulation process, and quantize the height range to 62 levels (21.5 nm/level). Our fabricated DOE exhibits real PSFs similar to the simulation (Figure 9).

The diffractive optical element is fabricated through soft lithography [14]. A master mold is made with positive photoresist (AZ-1512, MicroChemicals) spun on a titanium-coated glass substrate. The pattern is written by a direct-write gray-scale photolithography machine (MicroWriter ML3, Durham Magneto Optics Ltd) and developed with a MF-319 developer (Microposit). After the development, polydimethylsiloxane (PDMS, SYLGARD 184, Dow) is cast and cured at room temperature with the master mold to form another mold. This PDMS mold is used to transfer the pattern to a 3 mm-thick float glass substrate (30-773, Edmund Optics).

The glass substrate is preprocessed to form a circular aperture with the layers of chromium and gold through a lift-off process. A drop of UV-curable resin (NOA61, Norland Products, its refractive index at 546.1 nm is 1.5634) is then sandwiched between the glass substrate and the PDMS mold, and is exposed to a mercury-vapor lamp to cure the resin. After the PDMS mold is peeled off, the pattern is

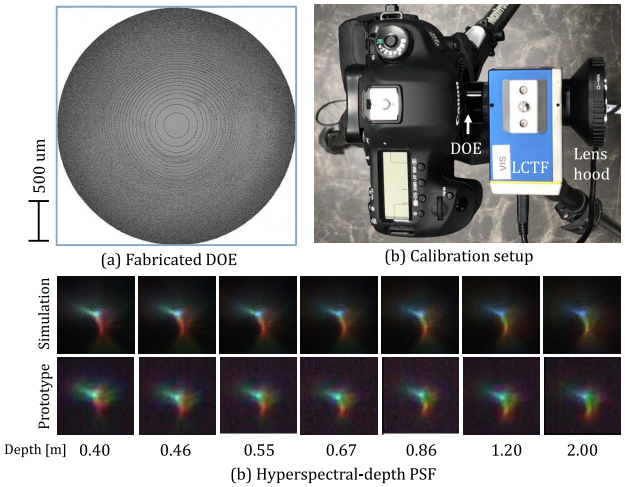


Figure 9. (a) We fabricate the learned DOE with the soft lithography. (b) To calibrate the PSFs of the prototype, we place a LCTF filter in front of the DOE and capture the hyperspectral images of a point light source at different depths. (c) Calibrated PSFs of the prototype show similar trend with the simulated PSFs.

replicated on the NOA61-resin layer which acts as a DOE. As the fabrication accuracy of the DOEs cannot be directly measured due to their microscale patterns, the accuracy of the fabrication system was indirectly measured on 15 reference holes which are designed to have different depths over 2 μm. The depths of the fabricated reference holes were measured with a profilometer (KLA Tencor Alpha Step D-500). The RMSE of the 15 sample points was 173.2 nm, and the estimated quantization scale was 20.5 nm/level. The remaining area of the glass substrate is covered by a chrome aperture mask of the same diameter (3 mm) placed on the same side where the DOE is printed.

10. Discussion

Spectral-depth tradeoff. Since we aim to estimate both spectrum and depth, the DOE and reconstruction network could be optimized favorably for one of them. While adjusting the weights of the main loss function of our method can balance this, it would be interesting to develop a method for handling this tradeoff in a fairer manner. Also, we observed that there are similarly shaped PSFs in the spectrum and depth slices of the optimized PSF. Even though we alleviate this with the reconstruction network by learning spatio-spectral priors, the ambiguity in the PSF is still challenging to resolve perfectly. To improve the reconstruction quality for spectrum and depth information, it would be worthwhile to optically resolve this ambiguity using multiple DOEs or other optical elements in future work.

Training strategy. The optimization problem for end-to-end optics includes many non-convex optimization problems and thus the initialization is critical and the training

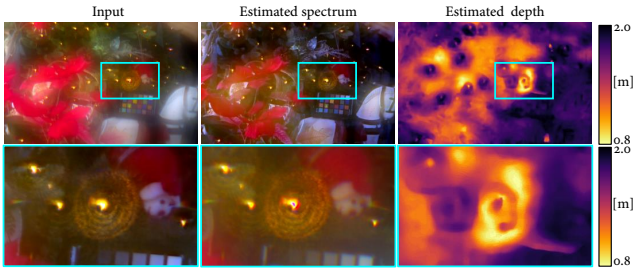


Figure 10. On regions where pixels are saturated by specular highlights, our method results in reconstruction artifacts in both the spectrum and depth map.

strategy of both optics and neural network is also important. In this work, we followed the existing conventions for the DOE initialization and network training in recent studies [3, 13]. Developing better training strategies for end-to-end optimization would be an interesting avenue of future work. Also, employing more advanced reconstruction schemes inspired by traditional optimization methods would be helpful to make the reconstruction interpretable in terms of end-to-end optics, even though it was not feasible in this work because of the demanding GPU memory for the problem of HS-D simulation and reconstruction.

Textureless and saturated regions. Our method falls in the regime of PSF engineering approaches, which fundamentally depend on texture information, similar to stereo imaging or depth-from-defocus imaging. This limits the accuracy of the reconstruction quality on texture-less or saturated surfaces as shown in the various results. In future work, it would be interesting to simultaneously estimate reconstruction confidence maps in addition to spectrum and depth, and then propagate the highly confident reconstruction to the regions with low confidence. Figure 10 shows a failure case of reconstruction on the saturated pixel regions.

Signal-dependent Poisson Noise. We use signal independent Gaussian noise for hyperspectral-depth image simulation. Employing a signal-dependent Poisson noise is challenging because of the discrete nature of the sampling procedure, making it non-differentiable. Reparameterization trick exists for the differentiable Poisson noise, however this induces instability of training. Therefore, following previous works on end-to-end optimization of optics and reconstruction, we use only the Gaussian noise.

Spectral range. In principle, our proposed method can be extended to a wider spectral range (e.g., infrared wavelengths). However, our target spectral range is limited to visible spectrum (420 to 660 nm) by the camera response function of a DSLR camera. It would be an interesting future work of extending our method to different spectral ranges.

Decoupled DOEs. Mechanically changeable DOEs spe-

cialized for depth and spectrum respectively is an interesting future work that could provide higher accuracy with the analogy to conventional multi-mode microscopes with rotating reconfigurable optics.

References

[1] Seung-Hwan Baek, Incheol Kim, Diego Gutierrez, and Min H Kim. Compact single-shot hyperspectral imaging using a prism. *ACM Trans. Graph. (TOG)*, 36(6):217, 2017. 2, 3

[2] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 5

[3] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 3, 5, 7

[4] Daniel S Jeon, Seung-Hwan Baek, Shinyoung Yi, Qiang Fu, Xiong Dun, Wolfgang Heidrich, and Min H Kim. Compact snapshot hyperspectral imaging with diffracted rotation. *ACM Trans. Graph. (TOG)*, 38(4):117, 2019. 2, 3, 4, 5

[5] Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. Image and depth from a conventional camera with a coded aperture. In *ACM Trans. Graph. (TOG)*, volume 26, page 70. ACM, 2007. 2

[6] Daniel Moreno and Gabriel Taubin. Simple, accurate, and robust projector-camera calibration. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 464–471. IEEE, 2012. 1

[7] 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. 4

[8] Yifan Peng, Qiang Fu, Felix Heide, and Wolfgang Heidrich. The diffractive achromat full spectrum computational imaging with diffractive optics. In *SIGGRAPH ASIA 2016 Virtual Reality meets Physical Reality: Modelling and Simulating Virtual Humans and Environments*, page 4. ACM, 2016. 2

[9] Yifan Peng, Qilin Sun, Xiong Dun, Gordon Wetzstein, Wolfgang Heidrich, and Felix Heide. Learned large field-of-view imaging with thin-plate optics. *ACM Trans. Graph. (TOG)*, 38(6):219, 2019. 5

[10] 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. 1

[11] Yoav Shechtman, Steffen Sahl, Adam Backer, and WE Moerner. Optimal point spread function design for 3d imaging. *Physical review letters*, 113(13):133902, 2014. 4

[12] Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Trans. Graph. (TOG)*, 37(4):114, 2018. 5

[13] 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. 3, 4, 5, 7

[14] Younan Xia and George M Whitesides. Soft lithography. *Annual review of materials science*, 28(1):153–184, 1998. 6