Single-shot Hyperspectral-Depth Imaging with Learned Diffractive Optics

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

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

Abstract

Imaging depth and spectrum have been extensively studied in isolation from each other for decades. Recently, hyperspectral-depth (HS-D) imaging emerges to capture both information simultaneously by combining two different imaging systems; one for depth, the other for spectrum. While being accurate, this combinational approach induces increased form factor, cost, capture time, and alignment/registration problems. In this work, departing from the combinational principle, we propose a compact single-shot monocular HS-D imaging method. Our method uses a diffractive optical element (DOE), the point spread function of which changes with respect to both depth and spectrum. This enables us to reconstruct spectrum and depth from a single captured image. To this end, we develop a differentiable simulator and a neural-network-based reconstruction method that are jointly optimized via automatic differentiation. To facilitate learning the DOE, we present a first HS-D dataset by building a benchtop HS-D imager that acquires high-quality ground truth. We evaluate our method with synthetic and real experiments by building an experimental prototype and achieve state-of-the-art HS-D imaging results.

1. Introduction

Spectral information is crucial for a plethora of applications in the fields of remote sensing, food/agriculture, medical imaging, and defense [28, 11, 1, 26, 5]. In parallel, depth imaging also has been developed for decades and now serves as a critical functionality for robotics, autonomous driving, mobile photography, and augmented/mixed reality [17, 36, 15]. These two imaging modalities recently started to be merged as hyperspectral-depth (HS-D) imaging that has various applications in ornithology, geology, biology, arts, and cultural heritage [22, 56, 23, 50, 12, 31, 35]. Specifically, single-shot HS-D imaging has applications such as analyzing living biological samples and geometric-spectral scene understanding from a moving camera on a vehicle or a hand-held camera.

To capture both spectrum and depth, existing HS-D imaging systems follow a combinational approach; they independently capture spectral and depth information with separate imaging systems, and combine the results after the fact [22, 47, 12, 50]. This enables accurate acquisition of both data by exploiting decade-long research in each regime. However, this combinational design fundamentally limits the potentials of HS-D imaging as it inevitably increases form factor, cost, capture time with additional alignment/registration problems.

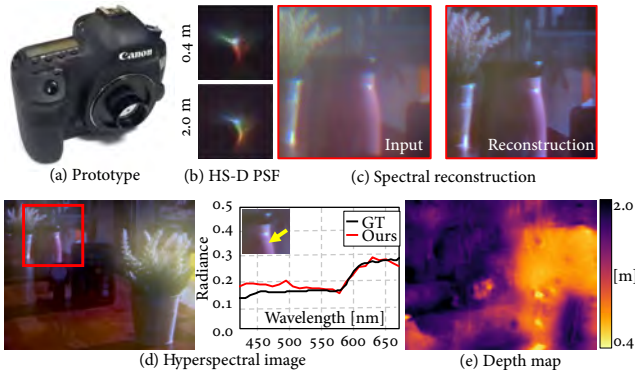


Figure 1. (a) Our compact single-shot HS-D imaging method uses an optimized DOE that creates (b) a PSF that varies with spectrum and depth. (c) – (e) It encodes spectral-depth information in the captured image, from which we simultaneously reconstruct a depth map and a hyperspectral image from 420 nm to 680 nm with 10 nm bandwidth.

In this work, instead of relying on two different imaging systems and combining them, we propose a first compact single-shot HS-D imaging method that uses a single diffractive optical element (DOE) in front of a conventional camera sensor. Our key intuition is that depth and spectrum are closely coupled in DOE-based imaging systems, thus allowing for single-shot capturing of an HS-D image. However, it is nontrivial to design a DOE that distinctively varies with scene depth and spectrum for HS-D imaging. To solve this problem, we implement a fully differentiable image simulator that synthesizes a sensor image for a given DOE height profile, building on recent advances in automatic differentiation and Fourier optics [38, 8, 51, 42, 41, 43, 49]. Combined with a convolutional neural network (CNN) that estimates depth and spectrum from a sensor image, we build an end-to-end differentiable pipeline from the DOE profile to

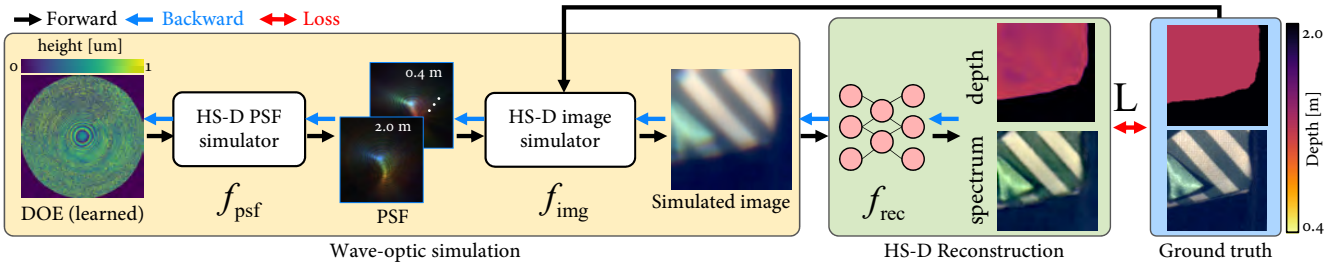


Figure 2. Our single-shot HS-D imaging is based on a differentiable pipeline that includes a wave-optics simulation and an HS-D image reconstruction part. For a DOE height map, we simulate its PSF at each depth and spectrum sample using a PSF simulator f_{psf}. Then, the HS-D image simulator f_{img} computes a sensor image with the aid of the ground-truth HS-D dataset. The CNN-based reconstructor f_{rec} estimates a hyperspectral image and a depth map. As the entire pipeline is differentiable, we optimize the DOE and the CNN by backpropagating the loss $\mathcal{L}$.

the reconstructed depth and spectrum, allowing us to jointly optimize DOE and CNN via backpropagation. The obtained DOE exhibits distinct variations in its shape with respect to both depth and spectrum.

For such joint optimization, one key missing part is the ground-truth HS-D dataset to supervise the optimization. As no such HS-D dataset exists, we build a benchtop HS-D imaging system that includes a structured light-based 3D scanning module and a bandpass filter-based hyperspectral imaging module. With this benchtop setup, we capture a first HS-D dataset that can be used for data-driven plenoptic imaging researches. We will make the dataset publicly available.

Our HS-D imaging method trained on the HS-D dataset outperforms the state-of-the-art single-shot HS-D imaging methods and alternative optical designs both in terms of form factor and accuracy. Our specific contributions are as follows.

- First compact monocular HS-D imaging method with a learned DOE that captures a depth map and a hyperspectral image from a single shot,
- First HS-D dataset of hyperspectral reflectance images and depth maps acquired by a benchtop HS-D imaging system that could fuel data-driven plenoptic imaging research, and
- Experimental verification in simulation and real experiments by fabricating an optimized DOE.

2. Related Work

Hyperspectral Imaging. Hyperspectral imaging has been extensively studied in the last decade. Scanning-based approaches capture multiple 1D spectral signals by isolating the spectral energy of each wavelength from others using a set of bandpass filters, a liquid crystal tunable filter (LCTF), or a slit with dispersive optics [4]. Compressive imaging techniques, a.k.a. coded aperture snapshot spectral imagers (CASSI), enable single-shot capture of hyperspectral images [45, 20, 21, 46, 9]. Recent approaches have demonstrated the potential of estimating hyperspectral images from spectrally varying point spread functions (PSFs) [3, 19] in a compact configuration that make use of edge information instead of using the modulated aperture mask. Our approach extends the capabilities of these spectrum-from-PSF methods by taking a first step towards snapshot imaging of higher-dimensional visual data: spectrum as well as depth.

Depth Imaging. Depth imaging is a widely studied topic. The approaches closest to ours include methods using the PSFs for depth estimation from a single image [2]. While traditional depth-from-defocus (DfD) analyzes the depth-dependent PSFs of a conventional camera to infer a depth map from two or more images [29, 39], depth-dependent spectral PSFs implemented by diffractive optical elements have also been exploited for all-in-focus particle imaging velocimetry [53]. Several groups have proposed computational photography approaches that employ amplitude-coded apertures [25, 44] and phase masks [33, 51] to simplify the depth estimation problem. Our work is inspired by these approaches, but we explore applications to HS-D imaging.

Hyperspectral-depth Imaging. HS-D imaging has been explored based on combinational paradigm that combines different imaging systems for spectrum and depth. For example, passive stereo [50, 18, 56] and active stereo [23, 31, 22] have been employed in conjunction with spectral cameras [50, 22, 56, 31] and spectral light sources [18, 23]. These approaches use two different imaging modalities for spectral and depth information, significantly increasing the device form factors and, in many cases, making it difficult to match stereo features across different spectral bands. CASSI systems have also been combined with light-field or time-of-flight (TOF) imaging to achieve snapshot monocular imaging [12, 35], but these systems use custom optical coding strategies which are restricted to indoor scenes only. To date, these systems have only been demonstrated on an optical table with a large form factor, limiting its applications. Furthermore, parallax and related alignment problems across modalities can negatively affect the reconstruction results. In contrast, we demonstrate a compact monocular HS-D imaging system with a learned DOE that support

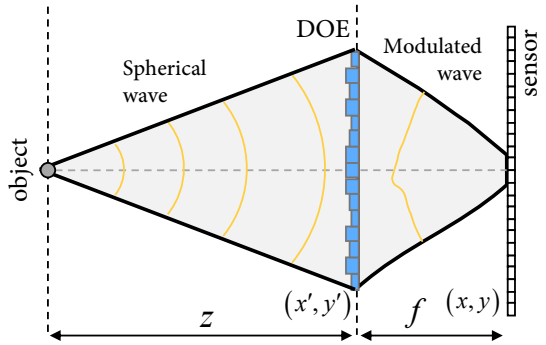


Figure 3. Schematic diagram of the light propagation from the object at depth z to the sensor with the focal length f. The phase of the spherical light wave coming from a scene point is modulated by the DOE and captured by a conventional sensor. The corresponding PSF varies as wavelength λ and depth z, enabling HS-D imaging from a single image.

single-shot capability, operating in a fully passive way.

Differentiable Optical Simulation. The idea of jointly optimizing optical elements with differentiable reconstruction algorithms has recently been explored for various applications [49], for instance, color filter design [6], spectral imaging [48], superresolution localization microscopy [30], super-resolution SPAD imaging [42], depth estimation [16, 51, 8], extended depth of field and super-resolution imaging [38], HDR imaging [27, 40], and image classification [7]. Based on this paradigm, we train a DOE for HS-D imaging while learning a reconstruction network. Our approach is the first to propose and demonstrate end-to-end optimization of a single DOE and a CNN for snapshot hyperspectral and also hyperspectral-depth imaging.

3. Diffraction-based HS-D Encoding

Our key intuition is that the wavefront of the diffracted light wave by a DOE changes with spectrum and scene depth. We examine this by establishing an HS-D image formation model based on Fourier optics [14] in a differentiable manner. This amounts to simulating a PSF and a sensor image for a given DOE, which comprise our pipeline shown in Figure 2.

Differentiable Point Spread Function. We denote $f_{\text{psf}}(\cdot)$ as our differentiable PSF simulator that computes a PSF for the given DOE height map h, wavelength λ, and depth z

$$P_{\lambda,z} = f_{\text{psf}}(h), \tag{1}$$

where $P_{\lambda,z}$ is the PSF. The wave field of wavelength λ originating from the scene point at depth z can be modeled as a spherical wave $U_{\lambda,z}$ in the aperture plane. Refer to Figure 3. Under the Fresnel approximation[1], this is modeled as $U^{(1)}_{\lambda,z} = \exp\left[i\frac{2\pi}{\lambda}\frac{x'^2+y'^2}{z}\right]$, where (x', y') is the spatial coordinate on the DOE plane. The wave field then passes through the camera aperture and the DOE resulting in changes of the amplitude and phase: $U^{(2)}_{\lambda,z} = A(x', y')\cdot\exp\left[i\frac{2\pi}{\lambda}\left(\frac{x'^2+y'^2}{z} + (\eta_\lambda - 1)h(x', y')\right)\right]$, where A is the amplitude aperture function, which is 0 for the blocked region and 1 elsewhere, and η_λ is the refractive index of the DOE material for wavelength λ.

The wave field propagates to the sensor by the focal length f, resulting in the point spread function $P_{\lambda,z}$, which is the squared magnitude of the complex wave field at the sensor plane:

$$P_{\lambda,z} = |\mathcal{F}\{A\cdot\exp[ik(\phi_{\text{scene}} + \phi_{\text{DOE}} + \phi_{\text{focal}}]\}|^2, \tag{2}$$

where $\mathcal{F}$ is the Fourier transform, k is the wave number $2\pi/\lambda$, and $\phi_{\{\text{scene/DOE/focal}\}}$ are the phase delays induced by propagating the scene to the DOE $\phi_{\text{scene}} = (x'^2 + y'^2)/z$, the DOE itself $\phi_{\text{DOE}} = (\eta_\lambda - 1)h(x', y')$, and propagating from the DOE to the sensor $\phi_{\text{focal}} = (x'^2 + y'^2)/(2f)$.

HS-D Encoding in PSF. Equation (2) shows that the PSF generated by a DOE depends on the wavelength λ and the depth z of a scene point as shown in the three phase terms $\phi_{\text{scene/DOE/focal}}$ augmented by the wave number k. The first term $k\phi_{\text{scene}}$ is inversely proportional to both wavelength λ and depth z, the second term $k\phi_{\text{DOE}}$ is proportional to the refractive index η_λ of the DOE material and inversely proportional to λ, and the last term $k\phi_{\text{focal}}$ is also inversely proportional to λ and the focal length f of the DOE. These terms are then summed up and converted to the frequency domain through Fourier transform $\mathcal{F}$. This analysis concludes that the point spread function $P_{\lambda,z}$ changes by wavelength λ and depth z. Therefore, we can in principle reconstruct these two pieces of information from PSF analysis in a single image.

However, factorizing depth and spectrum from a single image is a severely ill-posed problem, so making traditional PSF engineering approaches rarely attempt to solve this problem. To mitigate this challenge, we introduce a unified data-driven imaging solution that includes a learned DOE and a reconstruction network in an end-to-end manner.

Sensor Image Synthesis. Given the PSF $P_{\lambda,z}$, a pair of all-in-focus hyperspectral image I_λ and a depth map Z from our HS-D dataset, we simulate the corresponding sensor image $J_{c\in\{R,G,B\}}$ using an image simulator $f_{\text{img}}(\cdot)$. It is based on a convolutional PSF model and a layered scene representation, where a scene is modeled as a composition of multiple layers at different depths [8, 51]. The sensor image J_c is then modeled as follows:

$$\begin{aligned} J_c &= f_{\text{img}}(P_{\lambda,z}, I_\lambda, Z) \\ &= \sum_{\lambda\in\Lambda}\Omega_{c,\lambda}\sum_{z\in Z} M_z \odot (I_\lambda \otimes P_{\lambda,z}) + n, \end{aligned} \tag{3}$$

[1] It assumes that the wavelength λ is significantly smaller than the travel distance z: $\lambda \ll z$.

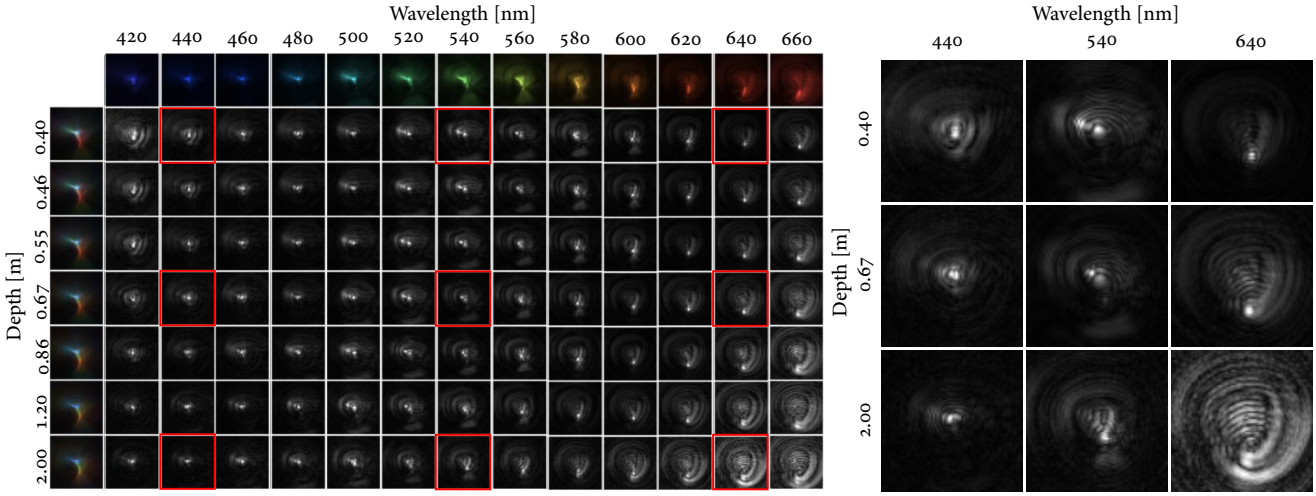


Figure 4. Our optimized PSF changes with respect to spectrum and depth over the spectral range from 420 to 660 nm and the depth range from 0.4 to 2.0 m. It enables the single-shot HS-D capture using a reconstruction network, simultaneously trained in the end-to-end manner.

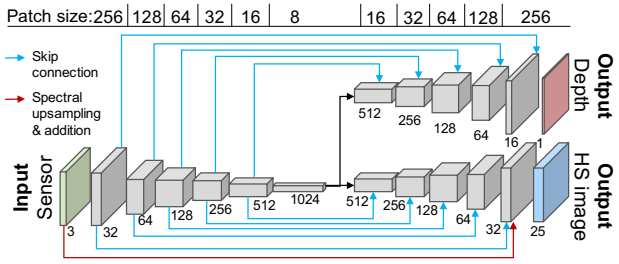


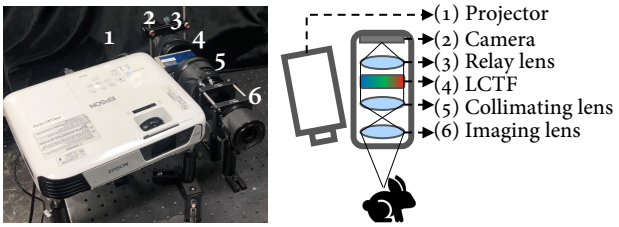
Figure 5. We take the sensor RGB image and extracts features using an encoder. Two decoders convert the features to a hyperspectral image and a depth map, respectively. In addition to the skip connection between the encoder and the decoders, we implement residual learning via a spectral-upsampling module.

where $\Omega_{c\in\{R,G,B\},\lambda\in\Lambda}$ is the camera response function, $\odot$ is an element-wise product operator, $\otimes$ is a convolution operator, and M_z is the weight map for each depth layer z. We compute M_z by applying a Gaussian filter to the binarized occupancy map that has value of one if the pixel depth is at z, and zero otherwise [8, 51]. Lastly, we apply signal-independent additive Gaussian noise n.We exclude Poisson noise due to its non-differentiability and even its reparameterization introduces training instability [32].

HS-D Sampling. We use dense sampling of 25 spectral channels from 420 to 660 nm in 10 nm intervals and seven depth levels from 0.4 to 2.0 m linearly spaced in disparity. This is essential not only for estimating an HS-D image from the spectral cue of PSFs, but also for accurately simulating the image formation model.

4. End-to-End HS-D Reconstruction

The spectrum-depth dependency of the PSF allows us to reconstruct a spectral image $\hat{I}_\lambda$ and a depth map $\hat{Z}$ from a captured image J_c. We use a deep convolutional neural



(a) Acquisition setup for HS-D dataset, schematic diagram

(b) Our HS-D dataset

Figure 6. (a) We built a benchtop HS-D imaging setup using structured-light 3D scanning and LCTF-based hyperspectral imaging. (b) We capture a first HS-D dataset consisting of hyperspectral reflectance images and depth maps using the benchtop setup.

network for the reconstruction algorithm $f_{\text{rec}}(\cdot)$:

$$\hat{I}_\lambda, \hat{Z} = f_{\text{rec}}\left(J_c\right). \tag{4}$$

Network Architecture. We design our network architecture based on the U-Net architecture [34] with a dual-decoder design: one for depth and the other for the HS image (Figure 5). This design is memory efficient essential for our end-to-end learning that consumes vast amount of GPU

HS-D imaging		Feng et al. [12]	Ours
Spec.	PSNR [dB]	23.62	**29.31**
	SSIM	0.76	**0.81**
Depth	RMSE [m]	0.57	**0.20**
	MAE [m]	0.30	**0.12**

Table 1. Our method outperforms the state-of-the-art HS-D imaging method [12] that combines the light-field imaging and the compressive spectral sensing.

DOE		Fresnel	Spiral [19]	Fisher [37]	E2E (ours)
Spec.	PSNR [dB]	27.96	26.90	28.51	**29.31**
	SSIM	0.74	0.64	0.79	**0.81**
Depth	RMSE [m]	0.21	0.32	0.23	**0.20**
	MAE [m]	0.15	0.20	0.15	**0.12**

Table 2. Our optimized DOE outperforms the alternative DOE designs (Fresnel, Spiral [19], Fisher [37]) for HS-D imaging.

memory for differentiable HS-D simulation. Further network details can be found in the Supplemental Document.

Residual Learning with Spectral Upsampling. To reduce the ill-posedness of the reconstruction problem, we apply residual learning [13] by adding the initial spectral tensor to the output HS image. However, different from conventional residual learning for super-resolution and demosaicing, the number of output spectral channels is larger than the input channels: from 3 to 25. As such, we propose to distribute the input RGB-channel intensity to the hyperspectral channels based on the camera response function Ω: $I_\lambda^{\text{up}} = \sum_c w(\lambda, c) J_c$, where $w(\lambda, c) = \Omega(\lambda, c) / \{\sum_{c'} \Omega(\lambda, c') \sum_{\lambda'} \Omega(\lambda', c)\}$ and $c, c' \in \{r, g, b\}$. The upsampled image I_λ^{up} is added to the output of the hyperspectral decoder, enabling effective residual learning (the red arrow in Figure 5).

Loss Function. Our loss function $\mathcal{L}$ is defined as the mean absolute error (MAE) of the inverse depth $D = 1/Z$ and the hyperspectral image I_λ, in addition to the total variation regularizer on depth:

$$\mathcal{L} = \alpha \frac{1}{N} \left\| \hat{I}_\lambda - I_\lambda \right\|_1 + \beta \frac{1}{M} \left\| \hat{D} - D \right\|_1 + \gamma \frac{1}{M} \left\| \nabla D \right\|_1, \tag{5}$$

where N and M are the number of total pixels in the hyperspectral image and the depth map, respectively, and ∇ is the spatial gradient operator. $\alpha = 1$, $\beta = 10^{-2}$, and $\gamma = 10^{-2}$ are the balancing weights.

Training. As the HS-D simulation and the reconstruction network are both differentiable, we jointly optimize the DOE and the network by solving a minimization problem via backpropagation:

$$\underset{h,\theta}{\text{minimize}}\, \mathcal{L}\left(\{\hat{Z}(h, \theta), \hat{I}_\lambda(h, \theta)\}, \{Z, I_\lambda\} \right), \tag{6}$$

where h is the DOE height, θ is a set of network parameters. We use patch-wise training with the patch size of 256×256. The DOE height h is initialized by Fisher information, the details of which are provided in the Supplemental Document.

Learned PSFs. If HS-D PSFs contain too complicated structures in spectral and depth dimensions, this inevitably leads to large spatial occupancy, limiting accurate deconvolution. If HS-D PSFs do not provide discriminative features, we suffer from metameric ambiguity of spectrum and depth. Our learned PSFs shown in Figure 4 are computationally obtained and exhibit spectral-depth variations with a compact size. Combining our neural network and the PSFs, we exploit both low-level optical PSF cue and higher-level contextual image cue to reconstruct a hyperspectral image and a depth map.

5. HS-D Dataset

To supervise the training, we present a *first* HS-D dataset. We provide 18 aligned pairs of a hyperspectral reflectance R_λ and a depth map Z (see Figure 6). Objects are carefully placed to avoid occlusion of the structured illumination. The reflectance R_λ is *spectrally* augmented to radiance I_λ using 29 CIE standard illuminants, resulting in 522 hyperspectral images. Our dataset covers visible spectrum from 420 nm to 680 nm with 10 nm interval and depth is in the range from 0.4 m to 2.5 m. Clear object boundary is guaranteed through manually-obtained background masks, assuring patch sampling from valid regions. For this data acquisition, we present a combinational HS-D imager combining structured light and LCTF-based hyperspectral imaging that acquires accurate spectral and depth data using brute-force scanning (Figure 6). Patch-wise training is employed for data efficiency, and 20,000 distinctive patches are used in total. Details on the imager can be found in the Supplemental Document.

6. Results

We perform synthetic evaluation of our method compared to the state-of-the-art methods and experimentally validate our prototype.

HS-D Imaging. Previous HS-D imaging methods take multi-sampling strategy in temporal domain [24], with multiple sensors [54, 55], or direct assorted filtering [10]. One notable exception is the work [12] that aims single-shot HS-D imaging by combining compressive spectral imaging and light-field imaging. Their system form factor is significantly larger than ours due to the complicated elements including a coded aperture mask, relay lenses, a prism, and a microlens array. Our method, with a much compact size, outperforms their method [12] as shown in Table 1 and Figure 7.

DOE Design. We compare our optimized DOE with alter-

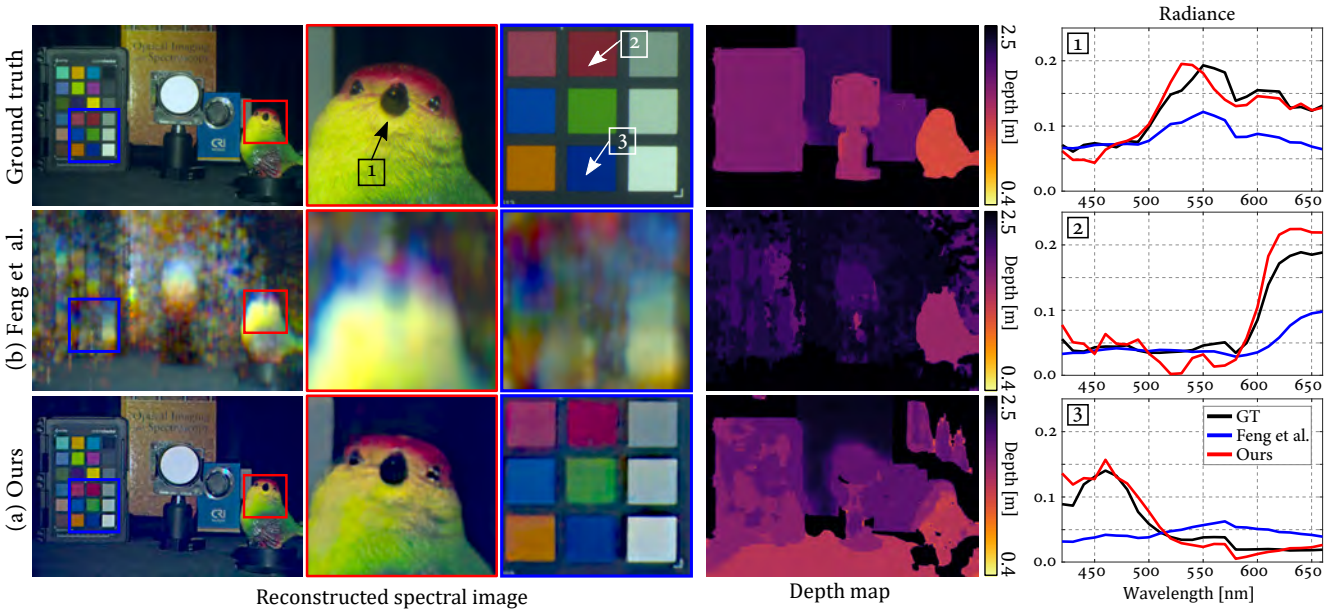


Figure 7. The spatial resolution of the spectral image and the depth map reconstructed by Feng et al. [12] is very low due to the fundamental architecture of light-field-based HS-D imaging. In contrast, our HS-D imaging can capture the spectral image and depth information with higher accuracy while our DOE-based camera has a significantly smaller form factor than the state-of-the-art system.

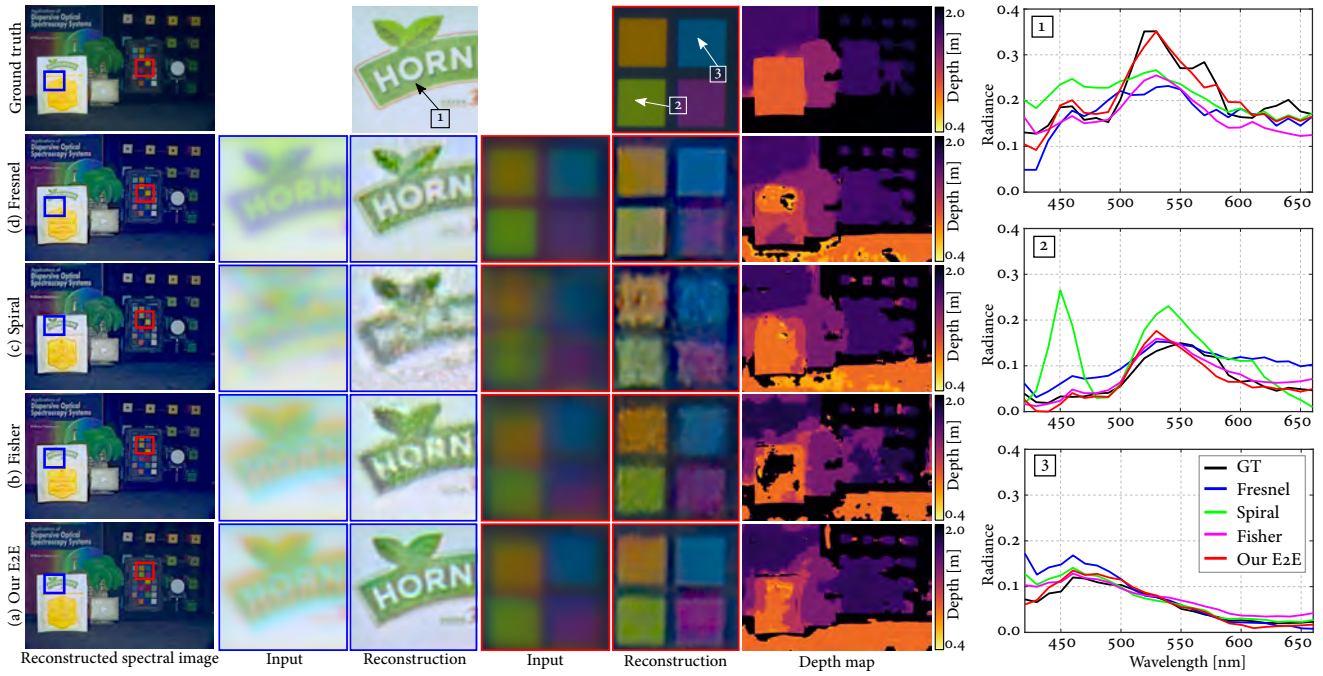


Figure 8. We compare our learned DOE with the Fresnel, the Spiral [19], the Fisher [37] DOEs. The spectral image and depth map reconstructed by our DOE method present significant improvements against results of the analytically-driven DOEs.

native DOE designs: a Fresnel DOE, a spiral DOE [19], and a Fisher DOE [37]. For fair comparison, we use our reconstruction network for all DOE designs. Table 2 and Figure 8 show that our DOE design outperforms the analytically driven designs. The Fresnel-lens DOE suffer from spectral metamerism with washed color artifacts (Figure 8d). The Fisher-information DOE and spiral DOE result in degraded spatial resolution (Figures 8b & 8c). While quantitative difference may appear small, the impact is clearly visible in qualitative reconstruction where high-frequency structures are preserved in our results.

Hyperspectral Imaging. Figure 10 compares our system with two hyperspectral-only imaging systems: a spiral DOE-based method [19] and a prism-based spectral imaging method [3]. It clearly shows that our method outperforms both methods. Baek et al. [3] suffers from smooth

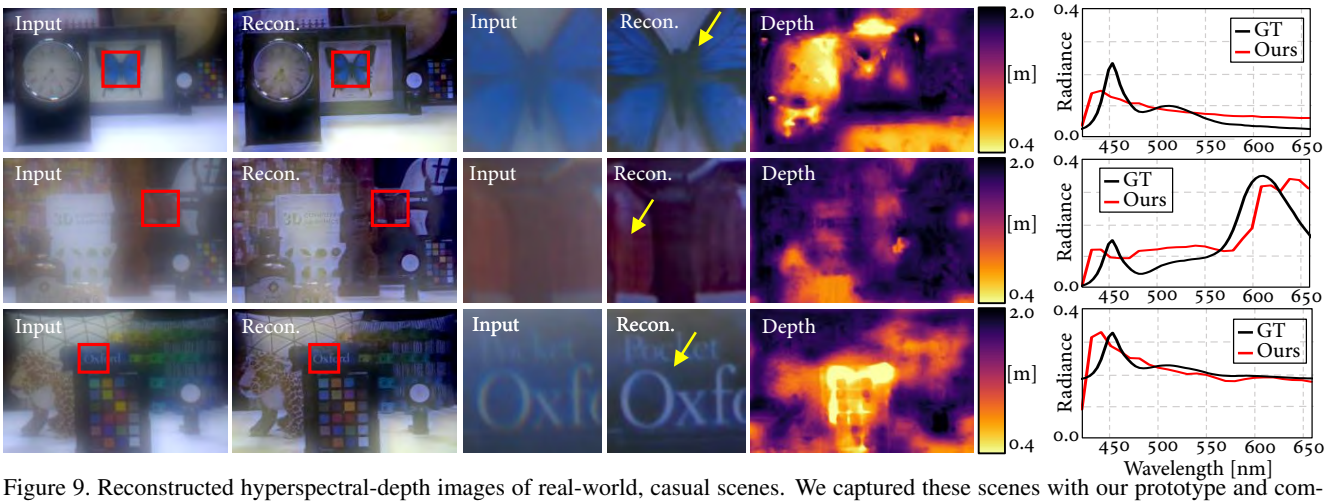
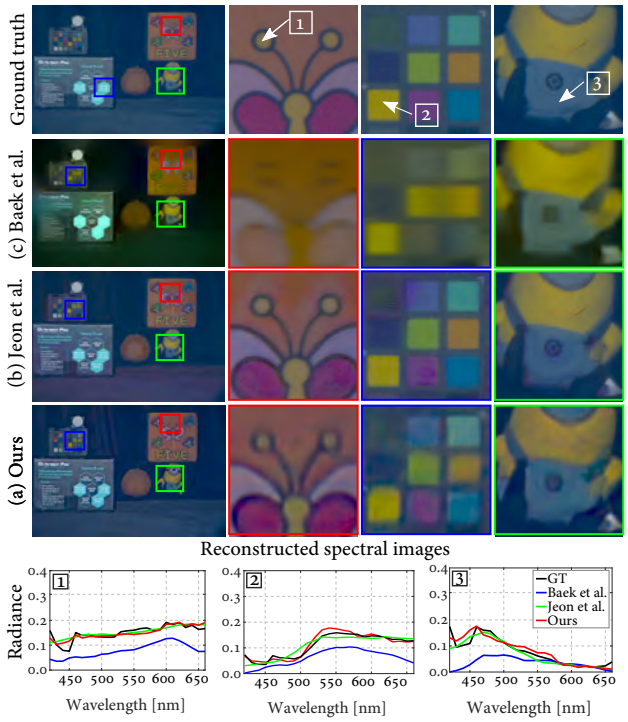


Figure 9. Reconstructed hyperspectral-depth images of real-world, casual scenes. We captured these scenes with our prototype and compare the normalized radiance of resulting HS-D data with the ground-truth measured by a spectroradiometer (SpectraScan 655) at points indicated by yellow arrows.



HS imaging	Baek et al. [3]	Jeon et al. [19]	Ours
HS PSNR [dB]	27.96	28.81	**29.31**
HS SSIM	0.75	**0.81**	**0.81**
Luminance PSNR [dB]	28	**40**	32
Luminance SSIM	0.91	**0.96**	0.94

Figure 10. We compare our method with the state-of-the-art HS imaging methods [3, 19]. Our proposed method outperforms both approaches in spectral accuracy (PSNR and SSIM computed on the hyperspectral cube) while achieving second-best performance in terms of spatial structure (PSNR and SSIM computed on the luminance image of the hyperspectral cube).

edge structures. Jeon et al. [19] produces overly smoothed spectral variation. Furthermore, different from their methods, our method estimate a hyperspectral image as well as a depth map. Refer to the supplemental document for the experimental details.

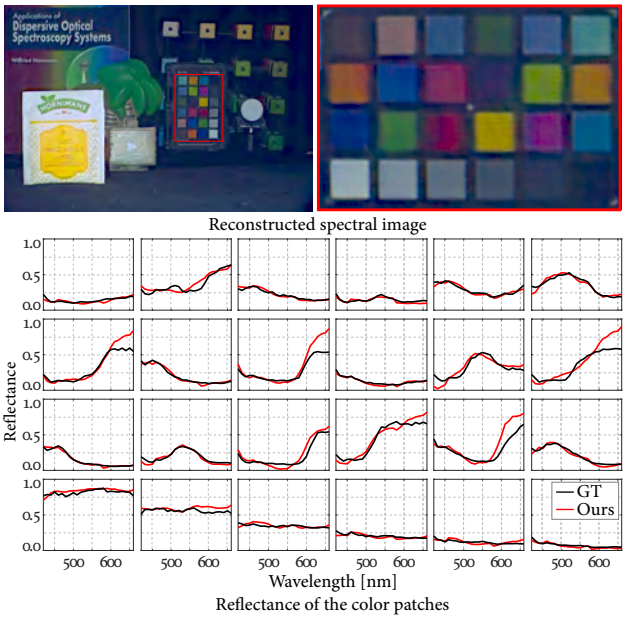
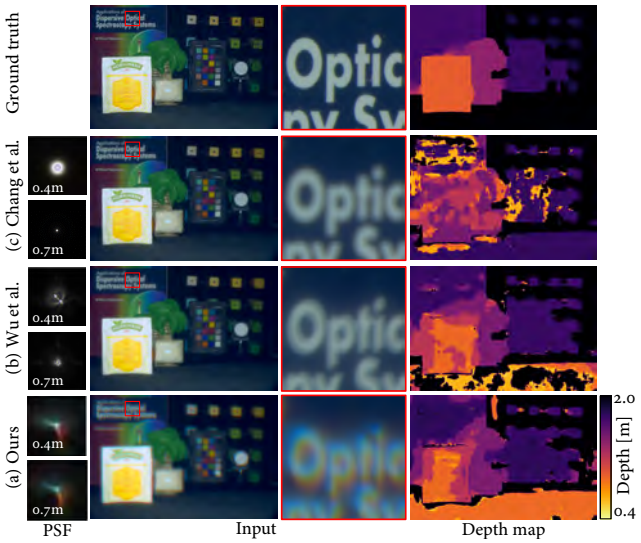


Figure 11. Quantitative evaluation of our HS-D imaging with a ColorChecker in a simulation with the ground truth. Spectral plots of 24 patches in the ColorChecker present spectral reconstruction of our method with high accuracy.

Depth Imaging. We compare our method to two DOE-based depth-only imaging methods: Wu et al. [51] and Chang et al. [8], by retraining our reconstruction network for each PSF. Figure 12 shows that our method can reconstruct clearer depth maps than other methods while it also reconstruct spectral information in addition. Refer to the supplemental document for the experimental details.



Depth imaging	Chang et al. [8]	Wu et al. [51]	Ours
RMSE [m]	0.45	0.31	**0.20**
MAE [m]	0.25	0.19	**0.12**

Figure 12. We compare our method with the state-of-the-art DOE-based depth imaging methods [8, 51]. Note that these methods are not designed for hyperspectral imaging. Our learned DOE outperforms the other two alternative designs in terms of depth accuracy.

PSF Dependency on Incident Angle. We found that our learned PSF exhibits insignificant variation with respect to incident angle. For validation, we simulate two PSFs at zero and eight degrees of incident angles, where the eight degree amounts to $\sim$60% of the vertical FOV (27 degrees).

Discriminating Power of PSF. We evaluate the discriminating powers of DOE using Cramer-Rao Lower Bound (CRLB) [37]. The CRLBs of the Fresnel/Spiral/Fisher DOEs are 2.73/1.89/1.36. As the Fisher DOE provides the highest discriminating power, we use this as an initialization point. However, the CRLB metric was originally designed only for a single sample at a 3D location and an impulse spectral peak. As we aim to capture natural scenes, not a single point, our end-to-end learning optimizes the DOE with the initialization from the Fisher DOE.

Spectral Evaluation. In order to evaluate the spectral accuracy of our system, we compare the reconstructed radiance of 24 patches in the standard ColorChecker under the CIE D65 illuminant with the ground truth in the simulation. As shown in Figure 11, our results closely match the spectral power distributions of every patch in the ground truth data, although we intentionally excluded the ColorChecker images in the training process to avoid overfitting of the network parameters to this target. The mean RMSE of reflectance (0.0–1.0) of all 24 patches is just 0.0478.

HS-D Imaging Prototype. We build our HS-D camera prototype using a Canon 5D Mark III camera sensor that has a pixel pitch of 6.22 μm and a resolution of 3840×5760 pixels. The focal length of the DOE is 50 mm. We fabricate the optimized DOE through soft lithography [52] of which detail is in the Supplemental Document. The DOE is then mounted to a C-mount tube, which is attached to the camera body with an EOS-C adapter. We made an additional C-mount extender with a 3D printer to place the DOE at the exact distance from the sensor. Calibration details, fine-tuning with the real PSFs, and the mismatch between the real-simulation PSFs are in the Supplemental Document. With our compact experimental prototype, we captured five real-world scenes shown in Figures 1 and 9, captured under sunlight and office-led lighting respectively. Ground-truth spectrum and depth for scene points are measured using a spectroradiometer and a laser distance meter.

Limitations. Our single-shot HS-D imaging method struggles with low-light and intensity-saturated scenes. Further, long-tail PSFs of the fabricated DOEs result in low contrast for real captured images, which is a prevailing artifact of existing DOE-based imaging methods due to the fabrication inaccuracy. A network architecture with more channels and depths would be beneficial to improve reconstruction quality. However, this is prohibited in our method due to the HS-D image simulation as a memory hog. Our method customizes DOEs for a target camera configuration. Its generalization to various camera parameters leaves as an interesting future work. Lastly, incorporating reconfigurable multiple DOEs could enhance depth and spectral accuracy.

7. Conclusion

We have presented a single-shot HS-D imaging system that consists of a learned DOE and a conventional DSLR camera. In contrast to the existing combinational approaches, the spectral and depth dependency of PSF is carefully analyzed and exploited in an end-to-end learning manner. To enable the joint learning procedure, we created the first HS-D dataset. Various experiments with the simulation and the real system validate the quality and accuracy of our method. Also, the proposed method can be used to reduce the form factor of HS-D imaging significantly, enabling compact and portable HS-D imaging without compromising quality and accuracy.

Acknowledgements

Min H. Kim acknowledges the support of Korea NRF grant (2019R1A2C3007229) in addition to Samsung Electronics, MSIT/IITP of Korea (2017-0-00072), National Research Institute of Cultural Heritage of Korea (2021A02P02-001), and Samsung Research Funding Center (SRFC-IT2001-04) for developing partial 3D imaging algorithms.

References

[1] Telmo Adao, Jonas Hruska, Luis Padua, Jose Bessa, Emanuel Peres, Raul Morais, and Joaquim Joao Sousa. Hyperspectral imaging: A review on uav-based sensors, data processing and applications for agriculture and forestry. *Remote Sensing*, 9(11):1110, 2017. 1

[2] Seung-Hwan Baek, Diego Gutierrez, and Min H. Kim. Birefractive stereo imaging for single-shot depth acquisition. *ACM Trans. Graph. (Proc. SIGGRAPH Asia 2016)*, 35(6), 2016. 2

[3] 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, 6, 7

[4] David J. Brady. *Optical Imaging and Spectroscopy*. John Wiley & Sons, Inc., 2009. 2

[5] X Briottet, Y Boucher, A Dimmeler, A Malaplate, A Cini, Marco Diani, HHPT Bekman, P Schwering, T Skauli, I Kasen, et al. Military applications of hyperspectral imagery. In *Targets and backgrounds XII: Characterization and representation*, volume 6239, page 62390B. International Society for Optics and Photonics, 2006. 1

[6] Ayan Chakrabarti. Learning sensor multiplexing design through back-propagation. In *Advances in Neural Information Processing Systems*, pages 3081–3089, 2016. 3

[7] 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. 3

[8] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 3, 4, 7, 8

[9] Inchang Choi, Daniel S. Jeon, Giljoo Nam, Diego Gutierrez, and Min H. Kim. High-quality hyperspectral reconstruction using a spectral prior. *ACM Trans. Graph. (Proc. SIGGRAPH Asia 2017)*, 36(6):218:1–13, 2017. 2

[10] Qi Cui, Jongchan Park, R Theodore Smith, and Liang Gao. Snapshot hyperspectral light field imaging using image mapping spectrometry. *Optics letters*, 45(3):772–775, 2020. 5

[11] Laura M Dale, André Thewis, Christelle Boudry, Ioan Rotar, Pierre Dardenne, Vincent Baeten, and Juan A Fernández Pierna. Hyperspectral imaging applications in agriculture and agro-food product quality and safety control: a review. *Applied Spectroscopy Reviews*, 48(2):142–159, 2013. 1

[12] Weiyi Feng, Hoover Rueda, Chen Fu, Gonzalo R Arce, Weiji He, and Qian Chen. 3d compressive spectral integral imaging. *Optics express*, 24(22):24859–24871, 2016. 1, 2, 5, 6

[13] Michaël Gharbi, Gaurav Chaurasia, Sylvain Paris, and Frédo Durand. Deep joint demosaicking and denoising. *ACM Trans. Graph.*, 35(6):191:1–191:12, Nov. 2016. 5

[14] Joseph W Goodman. *Introduction to Fourier optics*. Roberts and Company Publishers, 2005. 3

[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. 1

[16] 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. 3

[17] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Patrice Horaud. *Time-of-flight cameras: principles, methods and applications*. Springer Science & Business Media, 2012. 1

[18] Shuya Ito, Koichi Ito, Takafumi Aoki, and Masaru Tsuchida. A 3d reconstruction method with color reproduction from multi-band and multi-view images. In *Asian Conference on Computer Vision*, pages 236–247. Springer, 2016. 2

[19] 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, 5, 6, 7

[20] Daniel S Jeon, Inchang Choi, and Min H Kim. Multisampling compressive video spectroscopy. *Computer Graphics Forum*, 35(2):467–477, 2016. 2

[21] William R Johnson, Daniel W Wilson, Wolfgang Fink, Mark Humayun, and Greg Bearman. Snapshot hyperspectral imaging in ophthalmology. *J. biomedical optics*, 12(1):014036–014036, 2007. 2

[22] Min H Kim, Todd Alan Harvey, David S Kittle, Holly Rushmeier, Julie Dorsey, Richard O Prum, and David J Brady. 3d imaging spectroscopy for measuring hyperspectral patterns on solid objects. *ACM Trans. Graph. (TOG)*, 31(4):38, 2012. 1, 2

[23] Masahiro Kitahara, Takahiro Okabe, Christian Fuchs, and Hendrik PA Lensch. Simultaneous estimation of spectral reflectance and normal from a small number of images. In *VISAPP (1)*, pages 303–313, 2015. 1, 2

[24] Raimund Leitner, Andreas Kenda, and Andreas Tortschanoff. Hyperspectral light field imaging. In *Optical Sensors 2015*, volume 9506, page 950605. International Society for Optics and Photonics, 2015. 5

[25] 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

[26] Guolan Lu and Baowei Fei. Medical hyperspectral imaging: a review. *Journal of biomedical optics*, 19(1):010901, 2014. 1

[27] C. Metzler, H. Ikoma, Y. Peng, and G. Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *Proc. CVPR*, 2020. 3

[28] Roope Näsi, Eija Honkavaara, Päivi Lyytikäinen-Saarenmaa, Minna Blomqvist, Paula Litkey, Teemu Hakala, Niko Viljanen, Tuula Kantola, Topi Tanhuanpää, and Markus Holopainen. Using uav-based photogrammetry and hyperspectral imaging for mapping bark beetle damage at tree-level. *Remote Sensing*, 7(11):15467–15493, 2015. 1

[29] Shree K Nayar and Yasuo Nakagawa. Shape from focus. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16(8):824–831, 1994. 2

[30] 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. *Nature Methods*, 17:734–740, 2020. 3

[31] Keisuke Ozawa, Imari Sato, and Masahiro Yamaguchi. Hyperspectral photometric stereo for a single capture. *JOSA A*, 34(3):384–394, 2017. 1, 2

[32] 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

[33] Sri Rama Prasanna Pavani, Michael A Thompson, Julie S Biteen, Samuel J Lord, Na Liu, Robert J Twieg, Rafael Piestun, and WE Moerner. Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. *Proceedings of the National Academy of Sciences*, 106(9):2995–2999, 2009. 2

[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. 4

[35] Hoover Rueda-Chacon, Juan F Florez, Daniel Leo Lau, and Gonzalo R Arce. Snapshot compressive tof+ spectral imaging via optimized color-coded apertures. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1, 2

[36] 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. 1

[37] Yoav Shechtman, Steffen Sahl, Adam Backer, and WE Moerner. Optimal point spread function design for 3d imaging. *Physical review letters*, 113(13):133902, 2014. 5, 6, 8

[38] 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. 1, 3

[39] Murali Subbarao and Gopal Surya. Depth from defocus: a spatial domain approach. *Int. J. Computer Vision (IJCV)*, 13(3):271–294, 1994. 2

[40] 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. 3

[41] Qilin Sun, Congli Wang, Fu Qiang, Dun Xiong, and Heidrich Wolfgang. End-to-end complex lens design with differentiable ray tracing. *ACM Trans. Graph.*, 40(4), 2021. 1

[42] Quilin Sun, Jian Zhang, Xiong Dun, Bernard Ghanem, Yifan peng, and Wolfgang Heidrich. End-to-end learned, optically coded super-resolution spad camera. *ACM Trans. Graph. (TOG)*, 39, 2020. 1, 3

[43] Ethan Tseng, Ali Mosleh, Fahim Mannan, Karl St-Arnaud, Avinash Sharma, Yifan Peng, Alexander Braun, Derek Nowrouzezahrai, Jean-Francois Lalonde, and Felix Heide. Differentiable compound optics and processing pipeline optimization for end-to-end camera design. *ACM Trans. Graph. (TOG)*, 40(4), 2021. 1

[44] 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 Trans Graph. (SIGGRAPH)*, volume 26, page 69, 2007. 2

[45] Ashwin Wagadarikar, Renu John, Rebecca Willett, and David Brady. Single disperser design for coded aperture snapshot spectral imaging. *Applied optics*, 47(10):B44–B51, 2008. 2

[46] Lizhi Wang, Chen Sun, Ying Fu, Min H. Kim, and Hua Huang. Hyperspectral image reconstruction using a deep spatial-spectral prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[47] Lizhi Wang, Zhiwei Xiong, Guangming Shi, Wenjun Zeng, and Feng Wu. Simultaneous depth and spectral imaging with a cross-modal stereo system. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016. 1

[48] 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. 3

[49] Gordon Wetzstein, Aydogan Ozcan, Sylvain Gigan, Shanhui Fan, Dirk Englund, Marin Soljačić, Cornelia Denz, David A. B. Miller, and Demetri Psaltis. Inference in artificial intelligence with deep optics and photonics. *Nature*, 588(7836):39–47, 2020. 1, 3

[50] Jiamin Wu, Bo Xiong, Xing Lin, Jijun He, Jinli Suo, and Qionghai Dai. Snapshot hyperspectral volumetric microscopy. *Scientific Reports*, 6:24624, 2016. 1, 2

[51] 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. 1, 2, 3, 4, 7, 8

[52] Younan Xia and George M Whitesides. Soft lithography. *Annual review of materials science*, 28(1):153–184, 1998. 8

[53] Jinhui Xiong, Ramzi Idoughi, Andres A Aguirre-Pablo, Abdulrahman B Aljedaani, Xiong Dun, Qiang Fu, Sigurdur T Thoroddsen, and Wolfgang Heidrich. Rainbow particle imaging velocimetry for dense 3d fluid velocity imaging. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 36(4):36, 2017. 2

[54] Zhiwei Xiong, Lizhi Wang, Huiqun Li, Dong Liu, and Feng Wu. Snapshot hyperspectral light field imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3270–3278, 2017. 5

[55] Kang Zhu, Yujia Xue, Qiang Fu, Sing Bing Kang, Xilin Chen, and Jingyi Yu. Hyperspectral light field stereo matching. *IEEE transactions on pattern analysis and machine intelligence*, 41(5):1131–1143, 2018. 5

[56] Ali Zia, Jie Liang, Jun Zhou, and Yongsheng Gao. 3d reconstruction from hyperspectral images. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 318–325. IEEE, 2015. 1, 2