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.

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...
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...
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], super-resolution 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 DO 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 $f_{psf}(\cdot)$ as our differentiable PSF simulator that computes a PSF for the given DOE height map $h$, wavelength $\lambda$, and depth $z$

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

where $P_{\lambda,z}$ is the PSF. The wave field of wavelength $\lambda$ 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, this is modeled as $U_{\lambda,z}^{(1)} = \exp \left[ i \frac{2\pi}{\lambda} \frac{x'^2 + y'^2}{z} \right]$, 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: $U_{\lambda,z}^{(2)} = A(x', y') \exp \left[ i \frac{2\pi}{\lambda} \frac{x'^2 + y'^2}{z} + (\eta \lambda - 1) h(x', y') \right]$, where $A$ is the amplitude aperture function, which is 0 for the blocked region and 1 elsewhere, and $\eta \lambda$ is the refractive index of the DOE material for wavelength $\lambda$.

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:

$$P_{\lambda,z} = \left| \mathcal{F} \left( A \cdot \exp[i k(\phi_{\text{scene}} + \phi_{\text{DOE}} + \phi_{\text{focal}})] \right) \right|^2,$$

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 $\lambda$ 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 $\lambda$ and depth $z$, the second term $k\phi_{\text{DOE}}$ is proportional to the refractive index $\eta \lambda$ of the DOE material and inversely proportional to $\lambda$, and the last term $k\phi_{\text{focal}}$ is also inversely proportional to $\lambda$ 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 $\lambda$ 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 $I_\lambda$ and a depth map $Z$ from our HS-D dataset, we simulate the corresponding sensor image $J_{c(\{R,G,B\})}$ using an image simulator $f_{img}(\cdot)$. 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:

$$J_c = f_{img}(P_{\lambda,z}, I_\lambda, Z) = \sum_{\lambda \in \Lambda} \Omega_{c,\lambda} \sum_{z \in Z} M_z \otimes (I_\lambda \otimes P_{\lambda,z}) + n,$$

1It assumes that the wavelength $\lambda$ is significantly smaller than the travel distance $z$: $\lambda \ll z$. 
where $\Omega_{c, \lambda} \in \{R, G, B\}, \lambda \in \Lambda$ 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 network for the reconstruction algorithm $f_{\text{rec}}(\cdot)$:

$$\hat{I}_\lambda, \hat{Z} = f_{\text{rec}}(J_c).$$  \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
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 metamerism 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 $R_\lambda$ and a depth map $Z$ (see Figure 6). Objects are carefully placed to avoid occlusion of the structured illumination. The reflectance $R_\lambda$ is spectrally augmented to radiance $I_\lambda$ 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 significant 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.

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).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HS PSNR [dB]</td>
<td>27.96</td>
<td>28.81</td>
<td>29.31</td>
</tr>
<tr>
<td>HS SSIM</td>
<td>0.75</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td>Luminance PSNR [dB]</td>
<td>28</td>
<td>40</td>
<td>32</td>
</tr>
<tr>
<td>Luminance SSIM</td>
<td>0.91</td>
<td>0.96</td>
<td>0.94</td>
</tr>
</tbody>
</table>

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.

Figure 12. 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. 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.
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 ~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 radiation 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.


