Deep Hyperspectral-Depth Reconstruction Using Single Color-Dot Projection

Chunyu Li, Yusuke Monno, and Masatoshi Okutomi
Tokyo Institute of Technology, Tokyo, Japan
{lchunyu,ymonno}@ok.sc.e.titech.ac.jp, mxo@ctrl.titech.ac.jp

Abstract

Depth reconstruction and hyperspectral reflectance reconstruction are two active research topics in computer vision and image processing. Conventionally, these two topics have been studied separately using independent imaging setups and there is no existing method which can acquire depth and spectral reflectance simultaneously in one shot without using special hardware. In this paper, we propose a novel single-shot hyperspectral-depth reconstruction method using an off-the-shelf RGB camera and projector. Our method is based on a single color-dot projection, which simultaneously acts as structured light for depth reconstruction and spatially-varying color illuminations for hyperspectral reflectance reconstruction. To jointly reconstruct the depth and the hyperspectral reflectance from a single color-dot image, we propose a novel end-to-end network architecture that effectively incorporates a geometric color-dot pattern loss and a photometric hyperspectral reflectance loss. Through the experiments, we demonstrate that our hyperspectral-depth reconstruction method outperforms the combination of an existing state-of-the-art single-shot hyperspectral reflectance reconstruction method and depth reconstruction method.

1. Introduction

Depth reconstruction and hyperspectral reflectance reconstruction (spectral reconstruction, for short) are two active research areas in the fields of computer vision and image processing. Depth reconstruction aims at obtaining a scene's depth map, which presents the distances from the camera to each scene point. On the other hand, spectral reconstruction aims at acquiring scene's spectral reflectance information, which provides the wavelength-by-wavelength reflectance of each scene point. Since the depth and the spectral reflectance provides the scene's geometric and photometric properties, respectively, simultaneously acquiring them, which we refer to as hyperspectral-depth reconstruction, has various potential applications such as cultural heritage [6, 19], artwork authentication [30], material classification [5, 25], plant modeling [26], and relighting [35].

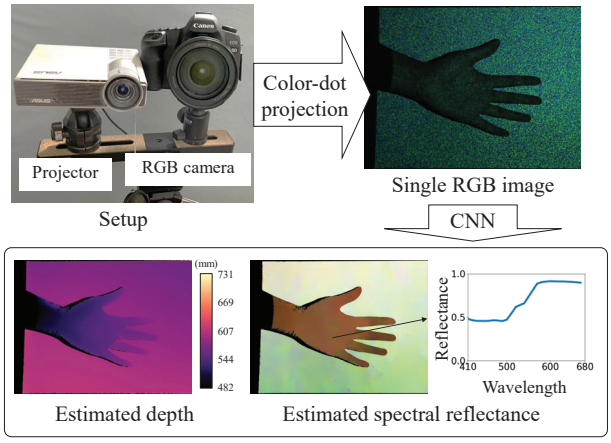


Figure 1. The overview of our system. From a single RGB image captured with a random color-dot projection, we simultaneously reconstruct the depth and the spectral reflectance for each pixel.

Although depth reconstruction and spectral reconstruction have been studied separately, some systems are recently designed to simultaneously acquire both the depth and the spectral reflectance. They typically combine a conventional depth-sensing technology with a hyperspectral camera [9, 13, 32, 34, 36, 39]. However, the requirement of a hyperspectral camera makes the system high cost. Some other systems use a standard RGB camera in conjunction with a variable and controllable light source, which emits temporally-changing illuminations to acquire multiband spectral observations [15,21–23,27,28,37]. However, these systems require multiple shots and thus are not applicable to dynamic scenes. Very recently, Baek et al. have proposed a single-shot system that uses a standard RGB camera and a diffractive optical element attached in front of the camera [3]. Although this system realizes a compact design using existing optical components, it still requires customized hardware design.

In this paper, we propose a novel single-shot system to simultaneously acquire the depth and the spectral reflectance using a standard RGB camera and an off-the-shelf RGB projector (see Fig. 1). Our system is based

on a single random color-dot projection, which simultaneously acts as structured light for depth reconstruction and spatially-varying color illuminations for spectral reconstruction. Since the random color dots provide a unique code pattern and three distinct RGB color illuminations for each local region, we exploit these cues for the hyperspectral-depth reconstruction. To effectively reconstruct the depth and the spectral reflectance from a single color-dot image, we propose a novel end-to-end deep learning method. Since the location of an observed color-dot pattern depends on the scene depth, we perform the joint learning of the depth and the spectral reflectance to improve the accuracy of each other, by considering the geometric warping of the color-dot pattern. Furthermore, to address the difficulty of constructing a real-world hyperspectral-depth dataset, we develop a spectral renderer to generate a synthetic dataset using a spectral rendering model under the color-dot illumination. Main contributions of this work are summarized as follows.

1. We propose the first single-shot hyperspectral-depth reconstruction system using a standard RGB camera and an off-the-shelf RGB projector without any hardware modifications.
2. We propose a novel network architecture and end-to-end learning method using a spectral renderer to simultaneously reconstruct the depth and the spectral reflectance from a single color-dot pattern image.
3. We experimentally validate the effectiveness of our system for synthetic and real-world scenes.

2. Related Works

Existing systems for hyperspectral-depth reconstruction are roughly classified into hyperspectral camera-based systems [9, 13, 32, 34, 36, 39] and controllable lighting-based systems [15, 21–23, 27, 28, 37].

Most of the hyperspectral camera-based systems acquire the depth and the spectral reflectance data by replacing the RGB camera of an existing depth-sensing technology, such as structured light [9, 13], ToF [32], stereo [34], and light fields [36, 39], with a hyperspectral camera. Although these systems can realize the single-shot acquisition of the depth and the spectral reflectance, the necessity of a hyperspectral camera brings high cost. Also, the integration of a hyperspectral camera into a depth-sensing device requires highly complicated and dedicated hardware design.

Controllable lighting-based systems are based on a traditional 3D reconstruction method that uses extra light sources, such as structured light [15, 22, 37] and photometric stereo [21, 23, 27, 28]. These systems use a standard RGB camera and observe spectral measurements by temporally changing illumination spectrum. However, since these systems require multiple shots, they are limited to static scenes.

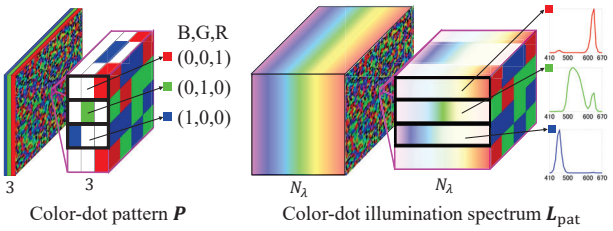


Figure 2. Color-dot representations. Left: Color-dot pattern $\boldsymbol{P}$, which is generated by randomly filling each projector pixel with one of three binary codes: R (0,0,1), G (0,1,0), and B (1,0,0). Right: Color-dot illumination spectrum $\boldsymbol{L}_{\mathrm{pat}}$, which has N_λ-dimensional illumination spectrum at each pixel.

There are two other classes of closely related methods: lighting-based hyperspectral imaging methods using an RGB camera [8, 12, 14, 29] and deep-learning-based active stereo methods [2, 10, 31, 38] (especially, Connecting the Dots [31], which learns to reconstruct the depth from a single gray-scale-dot pattern image, is the closest work to ours). Although these methods inspired us, they only reconstruct either the depth or the spectral reflectance. In contrast, our method simultaneously reconstructs the depth and the spectral reflectance from a single color-dot image based on end-to-end network learning, which consequently enables us to improve the accuracy of each other.

3. Proposed Method

3.1. Random Color-Dot Projection

In our system, we use an off-the-shelf RGB projector and a standard RGB camera to capture a single color-dot pattern image. The extrinsic and intrinsic parameters of the projector-camera system are pre-calibrated. The spectral sensitivity of the RGB camera and the spectral power distributions of the projector's RGB primaries are assumed to be known or pre-estimated.

As shown in the left figure of Fig. 2, the projector is used to project a single color-dot pattern to acquire geometric and spectral observations. The color-dot pattern $\boldsymbol{P}$ is generated by randomly filling each projector pixel with one of the three code words representing the projector's RGB primaries: R (0,0,1), G (0,1,0), and B (1,0,0).

As the geometric observation, the random pattern provides a locally unique code for establishing correspondences between the captured color-dot image and the reference color-dot pattern $\boldsymbol{P}$. As the spectral observation, the projector's RGB primaries provide three distinct illuminations, which results in the information from nine spectral bands (i.e., 3 illuminations $\times$ 3 color channels) by assuming locally uniform spectral reflectance. The illumination spectrum representation of the projected color-dot pattern is denoted by $\boldsymbol{L}_{\mathrm{pat}}$ and shown in the right figure of Fig. 2, where N_λ denotes the dimension discretized from the con-

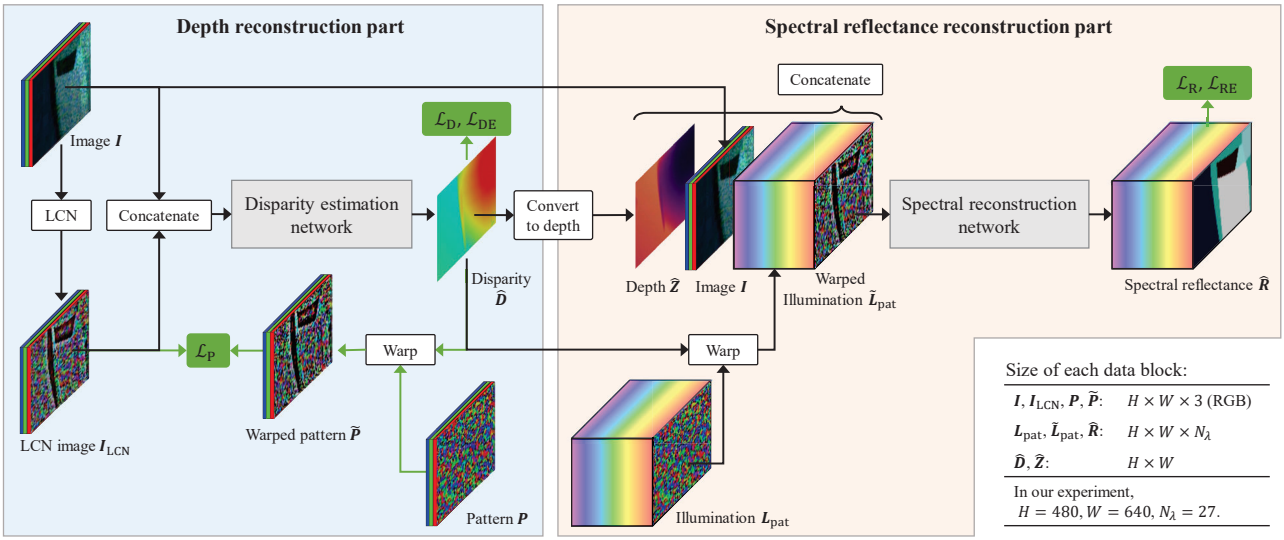


Figure 3. The overview of our end-to-end network architecture. As the first part, the disparity estimation network estimates the disparity map $\hat{\boldsymbol{D}}$ from the captured image $\boldsymbol{I}$ and the local contrast normalization (LCN) image $\boldsymbol{I}_{\mathrm{LCN}}$. Then, the estimated disparity map $\hat{\boldsymbol{D}}$ is converted to the depth map $\hat{\boldsymbol{Z}}$. As the second part, the spectral reconstruction network estimates the spectral reflectance image $\hat{\boldsymbol{R}}$ from the inputs of the captured image $\boldsymbol{I}$, the estimated depth map $\hat{\boldsymbol{Z}}$, and the warped illumination spectrum $\tilde{\boldsymbol{L}}_{\mathrm{pat}}$. The two networks are trained in an end-to-end manner using both geometric losses ($\mathcal{L}_{\mathrm{D}}$, $\mathcal{L}_{\mathrm{DE}}$ and $\mathcal{L}_{\mathrm{P}}$) and photometric losses ($\mathcal{L}_{\mathrm{R}}$ and $\mathcal{L}_{\mathrm{RE}}$).

tinuous wavelength domain (specifically, we used the sampling of every 10nm from 410nm to 670nm, i.e., N_λ=27).

3.2. End-to-End Network Architecture

Figure 3 illustrates the overview of our end-to-end network architecture. Based on the geometric and the photometric cues that can be observed from the color-dot projection, we reconstruct a disparity map $\hat{\boldsymbol{D}}$ and a spectral reflectance image $\hat{\boldsymbol{R}}$ from a single color-dot image $\boldsymbol{I}$, where the image $\boldsymbol{I}$ is rectified in advance using the extrinsic and intrinsic parameters of the projector-camera system. For the reconstruction, we apply two deep convolutional neural networks: disparity estimation network and spectral reconstruction network.

Firstly, the disparity estimation network estimates the disparity map $\hat{\boldsymbol{D}}$ with the inputs of the captured image $\boldsymbol{I}$ and the local contrast normalization (LCN) image $\boldsymbol{I}_{\mathrm{LCN}}$. LCN is applied to extract the color-dot pattern from $\boldsymbol{I}$. The estimated disparity is then converted to the depth map $\hat{\boldsymbol{Z}}$ using the calibrated parameters of the projector-camera system.

Then, the spectral reconstruction network estimates the spectral reflectance image $\hat{\boldsymbol{R}}$ with the inputs of the captured image $\boldsymbol{I}$, the estimated depth map $\hat{\boldsymbol{Z}}$, and the warped illumination spectrum $\tilde{\boldsymbol{L}}_{\mathrm{pat}}$. In this study, our aim is to reconstruct the spectral reflectance, which is inherent to a target object and irrelevant to the illumination and the scene geometry such as the shading. Since the depth provides an important cue to eliminate the effect of the shading from the estimated spectral reflectance, we input the depth map into the spectral reconstruction network. In addition, to provide the correct illumination information for each camera pixel, we input the illumination spectrum $\tilde{\boldsymbol{L}}_{\mathrm{pat}}$, which can be generated by warping the color-dot illumination spectrum $\boldsymbol{L}_{pat}$ from the projector viewpoint to the camera viewpoint based on the estimated disparity map.

The two networks are trained in a supervised and end-to-end manner using both geometric losses ($\mathcal{L}_{\mathrm{D}}$, $\mathcal{L}_{\mathrm{DE}}$ and $\mathcal{L}_{\mathrm{P}}$) and photometric losses ($\mathcal{L}_{\mathrm{R}}$ and $\mathcal{L}_{\mathrm{RE}}$). In our training process, the error of the estimated disparity will lead to wrong shading inference and wrong illumination warping for the spectral reconstruction network, meaning that the accuracy of the disparity affects the accuracy of the spectral reflectance. Thus, jointly training the two networks contributes to the improvement of the accuracy for both the disparity and the spectral reflectance, as we will demonstrate in the experimental result section.

3.2.1 Disparity Estimation Network

The disparity estimation network produces an output image of the same resolution as the input with left-right disparity information. Since the appearance of the color-dot pattern in the captured image $\boldsymbol{I}$ depends on various spatially-varying factors such as the shading and the texture, we preprocess the captured image to extract the projected color-dot pattern by applying LCN [16, 31, 38]. Following [31], for each pixel (u, v) and each color channel n, we compute the local mean μ and the standard deviation σ of a small local region (11×11 in our experiments) centered at pixel co-

ordinate (u, v). These local statistics are used to normalize the current pixel intensity as

$$\boldsymbol{I}_{\mathrm{LCN}}(u,v,n) = \frac{\boldsymbol{I}(u,v,n) - \boldsymbol{\mu}(u,v,n)}{\boldsymbol{\sigma}(u,v,n) + \eta}, \tag{1}$$

where η is a small constant to avoid numerical instabilities.

Then, we concatenate the LCN image $\boldsymbol{I}_{\mathrm{LCN}}$ with the original image $\boldsymbol{I}$ to form a six-channel input for the disparity estimation network, where the disparity is defined by the x-coordinate difference of the corresponding pixels between the captured image and the reference color-dot pattern. Given the estimated disparity map $\hat{\boldsymbol{D}}$, the scene depth $\hat{\boldsymbol{Z}}$ can be calculated as

$$\hat{\boldsymbol{Z}}(u,v) = bf/\hat{\boldsymbol{D}}(u,v), \tag{2}$$

where b is the baseline of the projector-camera system and f is the focal length of the rectified camera.

We design the network architecture based on the Disparity Decoder presented in [31]. This network consists of a contractive part and an expanding part with long-range links between them. In total, the network has 32 convolution layers and each of them is followed by ReLU. The final layer is followed by a scaled sigmoid non-linearity which constrains the output disparity map to the range between 0 and the maximum of the disparity. The network details can be found in the supplementary document.

3.2.2 Spectral Reconstruction Network

We next estimate the spectral reflectance image $\hat{\boldsymbol{R}}$ using the spectral reconstruction network. The captured image $\boldsymbol{I}$, the predicted scene depth $\hat{\boldsymbol{Z}}$ and the warped illumination spectrum to the camera viewpoint $\tilde{\boldsymbol{L}}_{\mathrm{pat}}$ are concatenated and passed to the spectral reconstruction network. As the disparity map provides the pixel correspondences between the captured image and the reference color-dot pattern, the warped illumination spectrum $\tilde{\boldsymbol{L}}_{\mathrm{pat}}$ can be calculated as

$$\tilde{\boldsymbol{L}}_{\mathrm{pat}}(u,v,\lambda) = \boldsymbol{L}_{\mathrm{pat}}(u - \hat{\boldsymbol{D}}(u,v), v, \lambda). \tag{3}$$

Since the disparity is estimated with sub-pixel accuracy, we apply bilinear interpolation for the resampling of the warped illumination spectrum.

The network architecture of the spectral reconstruction network is similar to that of the disparity estimation network with the difference of the input and the output channels. The range of the output spectral reflectance is constrained between 0 and 1 by the scaled sigmoid non-linearity.

3.3. Loss Function

The loss function for end-to-end training is described as

$$\mathcal{L} = \sum_{(u,v)\in\mathcal{V}} \mathcal{L}_{\mathrm{D}} + \omega_{\mathrm{DE}}\mathcal{L}_{\mathrm{DE}} + \omega_{\mathrm{P}}\mathcal{L}_{\mathrm{P}} + \omega_{\mathrm{R}}\mathcal{L}_{\mathrm{R}} + \omega_{\mathrm{RE}}\mathcal{L}_{\mathrm{RE}}, \tag{4}$$

including geometric losses (disparity loss $\mathcal{L}_{\mathrm{D}}$, disparity edge loss $\mathcal{L}_{\mathrm{DE}}$, and pattern loss $\mathcal{L}_{\mathrm{P}}$) and photometric losses (spectral reflectance loss $\mathcal{L}_{\mathrm{R}}$ and spectral reflectance edge loss $\mathcal{L}_{\mathrm{RE}}$). The balance of each loss is determined by the parameters ω_{DE}, ω_{P}, ω_{R}, and ω_{RE}. As the cast shadows that are apparent in the input image are meaningless in the network training, we binarize the input image to mask out the shadows and calculate the losses only for the non-shadow pixel set $(u, v) \in \mathcal{V}$.

Disparity loss $\mathcal{L}_{\mathrm{D}}$ and spectral reflectance loss $\mathcal{L}_{\mathrm{R}}$ compute the mean squared error between the ground truth and the estimated value as

$$\begin{aligned} \mathcal{L}_{\mathrm{D}} &= \|\hat{\boldsymbol{D}}(u,v) - \boldsymbol{D}_{\mathrm{gt}}(u,v)\|^2, \\ \mathcal{L}_{\mathrm{R}} &= \sum_{\lambda} \|\hat{\boldsymbol{R}}(u,v,\lambda) - \boldsymbol{R}_{\mathrm{gt}}(u,v,\lambda)\|^2, \end{aligned} \tag{5}$$

where $\boldsymbol{D}_{\mathrm{gt}}$ is the ground-truth disparity and $\boldsymbol{R}_{\mathrm{gt}}$ is the ground-truth spectral reflectance.

For pattern loss $\mathcal{L}_{\mathrm{P}}$, we take the advantage of the projector-camera setup to strengthen geometric constraints. To this end, we warp the reference color-dot pattern $\boldsymbol{P}$ to the camera viewpoint using the estimated disparity $\hat{\boldsymbol{D}}$ as

$$\tilde{\boldsymbol{P}}(u,v) = \boldsymbol{P}(u - \hat{\boldsymbol{D}}(u,v), v), \tag{6}$$

where $\tilde{\boldsymbol{P}}$ is the warped pattern. Then, we calculate the loss between the LCN image $\boldsymbol{I}_{\mathrm{LCN}}$ and the warped color-dot pattern $\tilde{\boldsymbol{P}}$ as

$$\mathcal{L}_{\mathrm{P}} = \|\boldsymbol{I}_{\mathrm{LCN}}(u,v) - \tilde{\boldsymbol{P}}(u,v)\|_C, \tag{7}$$

where $\|\cdot\|_C$ denotes the smooth Census transform [11].

As the color-dot pattern is relatively sparse, we further add disparity edge loss $\mathcal{L}_{\mathrm{DE}}$ and spectral reflectance edge loss $\mathcal{L}_{\mathrm{RE}}$ for predicting accurate and sharp boundaries. To this end, we use Sobel operator [18] to perform 2D spatial gradient calculation on the disparity and the spectral reflectance to enhance the boundaries. We apply a pair of Sobel convolution kernels to produce the approximate gradients of each pixel in the vertical and horizontal directions. Then, we calculate the errors of vertical gradients and horizontal gradients separately and add then up together. We formulate the losses $\mathcal{L}_{\mathrm{DE}}$ and $\mathcal{L}_{\mathrm{RE}}$ as

$$\begin{aligned} \mathcal{L}_{\mathrm{DE}} =& \|\hat{\boldsymbol{D}}^{\mathrm{V}}(u,v) - \boldsymbol{D}^{\mathrm{V}}_{\mathrm{gt}}(u,v)\|^2 \\ &+ \|\hat{\boldsymbol{D}}^{\mathrm{H}}(u,v) - \boldsymbol{D}^{\mathrm{H}}_{\mathrm{gt}}(u,v)\|^2, \end{aligned} \tag{8}$$

$$\begin{aligned} \mathcal{L}_{\mathrm{RE}} =& \sum_{\lambda} \|\hat{\boldsymbol{R}}^{\mathrm{V}}(u,v,\lambda) - \boldsymbol{R}^{\mathrm{V}}_{\mathrm{gt}}(u,v,\lambda)\|^2 \\ &+ \|\hat{\boldsymbol{R}}^{\mathrm{H}}(u,v,\lambda) - \boldsymbol{R}^{\mathrm{H}}_{\mathrm{gt}}(u,v,\lambda)\|^2, \end{aligned} \tag{9}$$

where the values with superscripts V and H denote the vertical gradient and the horizontal gradient, respectively.

3.4. Hyperspectral-Depth Dataset Generation

Since it is difficult to simultaneously acquire accurate depth and spectral reflectance as a large-scale ground-truth dataset in real-world situations, we developed a spectral renderer to generate a synthetic dataset with rendered RGB color-dot images, ground-truth disparity maps, and ground-truth spectral reflectance images by extending the algorithm of a structured-light renderer [31].

We render the scene with randomly populated 3D models using spectral reflectance samples. For simplicity, we first obtain the corresponding 3D point $\boldsymbol{x}$ for each pixel by computing the intersection of the camera ray and the 3D surface, and then acquire the ground-truth depth value as the z-coordinate of the 3D point in the camera coordinate system. The ground-truth spectral reflectance $\boldsymbol{r}$ of this 3D point is also obtained. According to Eq. (2), we can obtain the ground-truth disparity from the depth value. The illumination spectrum $\boldsymbol{l}$ for the 3D point is determined by the corresponding pattern code which can be obtained by reprojecting the 3D point to the projector's image plane.

Suppose that the camera response is linear and inter-reflection and ambient illumination are negligible, the camera's pixel intensity I of n-th color channel is calculated based on the spectral rendering model [22] as

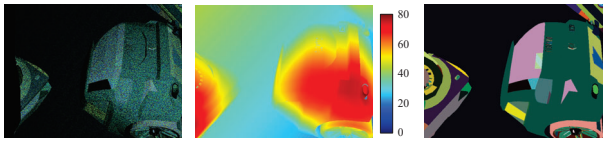
$$I(n) = s \int_{\Omega_\lambda} c(n,\lambda) l(\lambda) r(\lambda) d\lambda, \tag{10}$$

where $c(n,\lambda)$ is the n-th channel camera spectral sensitivity and λ represents the wavelength. Ω_λ is the wavelength range that the projector emits the illumination (410nm to 670nm for our used projector). s is the shading factor, which describes the proportion of the reflected radiance leaving the surface point $\boldsymbol{x}$ with respect to the intensity of the emitted light from the projector at position $\boldsymbol{x}_{\mathrm{pro}}$. Assuming that the 3D point has Lambertian reflectance and the projected illumination follows the inverse-square law, we define the shading factor s as

$$s = \frac{1}{\|\boldsymbol{x}_{\mathrm{pro}} - \boldsymbol{x}\|^2} \times \frac{\boldsymbol{x}_{\mathrm{pro}} - \boldsymbol{x}}{\|\boldsymbol{x}_{\mathrm{pro}} - \boldsymbol{x}\|} \cdot \boldsymbol{n}, \tag{11}$$

where $\boldsymbol{n}$ is the normal of the point $\boldsymbol{x}$. The first term represents the quadratic attenuation with respect to the distance of the object point $\boldsymbol{x}$ from the projector $\boldsymbol{x}_{\mathrm{pro}}$. The second term represents the inner product of the normalized lighting vector and the point normal $\boldsymbol{n}$. In practice, the continuous wavelength domain Ω_λ is discretized to N_λ dimension (we sampled at every 10nm from 410nm to 670nm, i.e., N_λ=27). The observed RGB intensity $[I(R), I(G), I(B)]^T$ can be computed by the matrix form as

$$\begin{bmatrix} I(R) \\ I(G) \\ I(B) \end{bmatrix} = s\boldsymbol{c}^T \mathrm{Diag}(\boldsymbol{l})\boldsymbol{r}, \tag{12}$$



(a) Color-dot image (b) Disparity map (c) sRGB

Figure 4. Spectral rendering examples: (a) A rendered input RGB image with the projected color-dot pattern. (b) A ground-truth disparity map, and (c) An sRGB image converted from the ground-truth spectral reflectances.

where $\boldsymbol{r} \in \mathbb{R}^{N_\lambda}$ represents the spectral reflectance, $\boldsymbol{l} \in \mathbb{R}^{N_\lambda}$ is the illumination spectrum corresponding one of the projector's RGB primaries, $\boldsymbol{c} \in \mathbb{R}^{N_\lambda \times 3}$ is the camera sensitivity matrix, and $\mathrm{Diag}(\cdot)$ is a square diagonal matrix function.

We used the same camera spectral sensitivity, projector illumination spectrum, and geometrically calibration parameters as our actual setup, which is described in Sec4.1.

4. Experimental Results

4.1. Setup and Implementation Details

We used an ASUS P3B projector and Canon EOS 5D Mark-II digital camera for our projector-camera system. The spectral power distributions of the projector's RGB primaries were measured by using a StellarNet BlueWave-VIS Spectrometer and they are shown in Fig. 2. The spectral sensitivity of EOS 5D camera was obtained from the public database of [17]. To calibrate the projector-camera system geometrically, we used the calibration method of [33].

For the synthetic dataset generation described in Sec.3.4, we used ShapeNet Core dataset [7] as the 3D models. We randomly placed the models at a distance from 0.3m to 1m and then randomly assigned the ground-truth spectral reflectance data to different texture parts of each 3D model. We generated 8,192 scenes for training, and 256 scenes for testing. For the training data, we used chair and car models in ShapeNet Core and the spectral reflectance data of 1,269 Munsell color chips [1]. For the testing data, we used camera, airplane, and watercraft models in ShapeNet Core and the spectral reflectance data of a standard X-Rite colorchart with 24 patches, which are unseen in the training data. The images were rendered with the resolution of 640×480. Rendering examples of an RGB color-dot image, a ground-truth disparity map, and an sRGB color image converted from the ground-truth spectral reflectances are shown in Fig. 4.

We implemented the proposed method in PyTorch and trained our model using Adam optimizer [20]. The learning rate was set as 1.0×10^{-4}. The loss weights in Eq. (4) were empirically set as $\omega_{\mathrm{DE}} = 100$, $\omega_{\mathrm{P}} = 0.2$, $\omega_{\mathrm{R}} = 1$, and $\omega_{\mathrm{RE}} = 8$. We used full-size 640x480 images for training. The total number of training epochs is 200 with batch size of 8. Training our model takes around 57 hours in total with one NVIDIA GeForce RTX 2080 Ti 11G GPU.

Table 1. Evaluation metrics.

Depth:

$$\mathrm{RMSE} = \sqrt{\mathrm{mean}\left[\left(\hat{\boldsymbol{Z}} - \boldsymbol{Z}_{\mathrm{gt}}\right)^2\right]}$$

$$\theta_i = \% \text{ of } \hat{\boldsymbol{Z}}(u,v) \text{ subject to } \max\left(\frac{\hat{\boldsymbol{Z}}(u,v)}{\boldsymbol{Z}_{\mathrm{gt}}(u,v)}, \frac{\boldsymbol{Z}_{\mathrm{gt}}(u,v)}{\hat{\boldsymbol{Z}}(u,v)}\right) < 1.03^i$$

Spectral reflectance:

$$\mathrm{RMSE} = \sqrt{\mathrm{mean}\left[\left(\hat{\boldsymbol{R}} - \boldsymbol{R}_{\mathrm{gt}}\right)^2\right]}$$

$$\mathrm{MRAE} = \mathrm{mean}\left(|\hat{\boldsymbol{R}} - \boldsymbol{R}_{\mathrm{gt}}|/\boldsymbol{R}_{\mathrm{gt}}\right)$$

Table 2. Quantitative comparison with the state-of-the-art methods on all the test scenes.

	Depth				Spectral reflectance	
	$\theta_1 \uparrow$	$\theta_2 \uparrow$	$\theta_3 \uparrow$	RMSE $\downarrow$	MRAE $\downarrow$	RMSE ($\times 10^{-2}$) $\downarrow$
AdaBins [4]	53.00	82.61	93.52	24.60	-	-
Connecting [31]	98.02	98.72	99.11	8.83	-	-
Basis [12]	-	-	-	-	0.38	8.02
AWAN [24]	-	-	-	-	0.34	7.93
Ours	**98.18**	**99.17**	**99.58**	**6.10**	**0.32**	**5.31**

4.2. Evaluation on Synthetic Data

We first qualitatively and quantitatively evaluate our proposed method on the test set of the synthetic dataset generated using the spectral renderer.

For the depth evaluation, we use root mean squared errors (RMSE) and threshold accuracy (θ_i) used in [4]. For the spectral reflectance evaluation, we use RMSE and mean relative absolute error (MRAE) used in [24]. These metrics are formulated in Table 1, where $\mathrm{mean}(\cdot)$ computes the arithmetic mean.

4.2.1 Comparison with State-of-the-Art Methods

Since there is no existing single-shot hyperspectral-depth reconstruction method directly applicable to our setup, we compare our method with state-of-the-art single-shot depth reconstruction methods and single-shot spectral reconstruction methods, respectively. For the depth evaluation, we compare our method with state-of-the-art AdaBins [4], which learns the depth from a standard RGB image without any dot pattern, and Connecting the Dots [31], which learns the depth from a single gray-scale image with Kinect dot pattern. Their networks were retrained using our dataset. Since their input images are different from ours, we re-rendered RGB images without the dot pattern (under white illumination) and gray-scale Kinect dot pattern images, respectively. The examples of the input images are shown in Fig. 5(a). For the spectral reflectance evaluation, we compare our method with two spectral reconstruction methods from a single RGB image: A widely-applied method using spectral reflectance basis functions (Basis) [12] and a state-of-the-art deep learning-based AWAN [24]. We used the RGB image without the dot pattern (under white illumination) as their inputs, which is the same input as AdaBins. For Basis, the spectral basis functions were calculated from 1,269 Munsell spectral reflectances of our training data. For AWAN, we retrained the model using our dataset.

Table 2 summarizes the overall quantitative evaluation on all 256 test scenes. We can observe that our method yields the best results for both the depth and the spectral reflectance. In contrast to the compared methods that only focus on a single property, our method jointly reconstructs the depth and the spectral reflectance by training the network model using both geometric losses and photometric losses, leading the improved performance to each other.

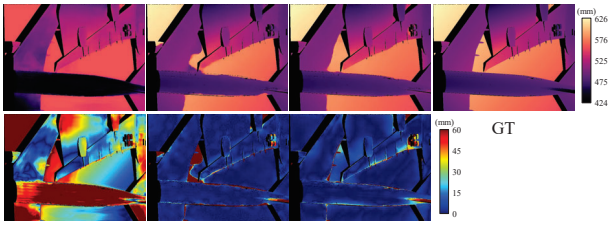
Figure 5(b) provides the qualitative results of the depth reconstruction, where our method provides a more accurate depth map. Figure 5(c) shows the visual comparison of sRGBs (top row), which was converted from the estimated spectral reflectances, and the error maps for the estimated spectral reflectances (bottom row), where RMSE for all wavelengths is visualized for each pixel. We can confirm that our sRGB result is the closest to the ground truth and represents the object's inherent spectral reflectance less affected by the shading, compared with the sRGB results of the existing methods that do not consider the shading (depth) information. Figure 5(d) shows the spectral reflectance results on eight sample points. Our method can reconstruct accurate spectral reflectances representing correct spectral shapes as well as correct relative scales. This is because that our method can benefit from the color-dot projection to acquire nine-band information and depth information, while the existing single-shot spectral reconstruction methods only rely on a standard three-band RGB image.

4.2.2 Ablation Study

Loss comparison for disparity estimation: To confirm the contribution of the individual geometric loss of the disparity estimation network (disparity loss $\mathcal{L}_{\mathrm{D}}$, disparity edge loss $\mathcal{L}_{\mathrm{DE}}$, and pattern loss $\mathcal{L}_{\mathrm{P}}$), we trained the disparity estimation network only and compared different loss combinations. Table 3 shows the depth reconstruction performance with the four loss combinations. We can observe that, if we add the disparity edge loss $\mathcal{L}_{\mathrm{DE}}$, the depth accuracy is significantly improved especially for large errors evaluated in the metrics θ_3, indicating that the edge loss contributes to reducing the boundary errors. The pattern loss $\mathcal{L}_{\mathrm{P}}$ significantly reduces overall RMSE, demonstrating the effectiveness of the color-dot pattern matching. We can also confirm that the best result can be achieved by using all the losses.

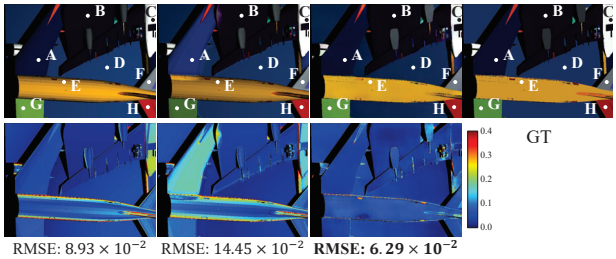
Input for AdaBins, Basis, and AWAN | Input for Connecting | Input for ours

(a) Input images for each method.



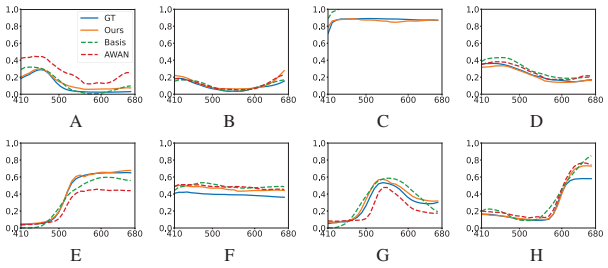
RMSE: 30.67 AdaBins | RMSE: 16.04 Connecting | **RMSE: 8.50 Ours**

(b) Depth results. Top: Estimated depth maps. Bottom: Visualized errors.



RMSE: 8.93×10^{-2} Basis | RMSE: 14.45×10^{-2} AWAN | **RMSE: 6.29×10^{-2} Ours**

(c) Spectral reflectance results. Top: sRGB color representation converted from the estimated spectral reflectances. Bottom: Visualized errors.



(d) Estimated spectral reflectance results on 8 sample points.

Figure 5. Synthetic comparison with the state-of-the-art methods.

Effectiveness of joint training: To demonstrate the effectiveness of the joint training of the depth and the spectral reflectance, we compare our full model with the following network models. (i) Disparity estimation network, which applies only the disparity estimation network for the disparity training, as compared in the previous paragraph (i.e., the best result of Table 3). (ii) Spectral reconstruction network, which applies only the spectral reconstruction network using only the single color-dot image input. The networks (i) and (ii) mean the cases of the separated network training at our setup. (iii) Joint network training without the depth input, which applies both the disparity estimation network and the spectral reconstruction network, but does not apply the depth input for the spectral reconstruction network. (iv) Joint network training without the illumination input, which applies both the networks, but does not apply the warped illumination input for the spectral reconstruction network. The networks (iii) and (iv) are compared to confirm the importance of the depth and the illumination inputs for estimating the spectral reflectance.

Table 3. Depth accuracy comparison with respect to different loss combinations for the disparity estimation network.

	$\theta_1 \uparrow$	$\theta_2 \uparrow$	$\theta_3 \uparrow$	RMSE $\downarrow$
$\mathcal{L}_{\mathrm{D}}$	97.89	98.71	99.04	8.01
$\mathcal{L}_{\mathrm{D}} + \mathcal{L}_{\mathrm{DE}}$	97.93	98.89	99.32	7.38
$\mathcal{L}_{\mathrm{D}} + \mathcal{L}_{\mathrm{P}}$	98.00	98.90	99.10	7.22
$\mathcal{L}_{\mathrm{D}} + \mathcal{L}_{\mathrm{DE}} + \mathcal{L}_{\mathrm{P}}$	**98.03**	**99.12**	**99.38**	**6.80**

Table 4. Effectiveness of joint training.

		Depth RMSE $\downarrow$	Reflectance RMSE $\downarrow$ ($\times 10^{-2}$)
Disparity estimation network		6.80	-
Spectral reconstruction network		-	5.79
Joint	w/o depth input	6.24	5.69
	w/o illumination input	6.32	5.75
	full model	**6.10**	**5.30**

Table 4 shows the results of the comparison. From the results, we can observe that joint training certainly provides better performance compared with the separated training of the depth and the spectral reflectance. In addition, both the depth and the warped illumination inputs to the spectral reconstruction network contribute to the performance improvement for estimating object-inherent (shading- and illumination-irrelevant, in other words) spectral reflectance. Interestingly, the depth result also can be significantly improved by using the warped illumination input, because the illumination spectrum pattern corresponding to the input image is accurate only when the disparity is correct, suggesting that the errors of the spectral reflectances can be back-propagated to update the disparity estimation network.

4.3. Results on Real Scenes

We next evaluate our method for real scenes. Because our method can realize the single-shot reconstruction of the depth and the spectral reflectance, we applied our method to a dynamic scene with a moving hand using a successive capturing mode of the camera. Figure 6 shows the captured input color-dot images (top row), the estimated depth maps from each input image (middle row) and the 3D point clouds converted from each depth map (bottom row). Each 3D point of the point cloud is colored by sRGB,

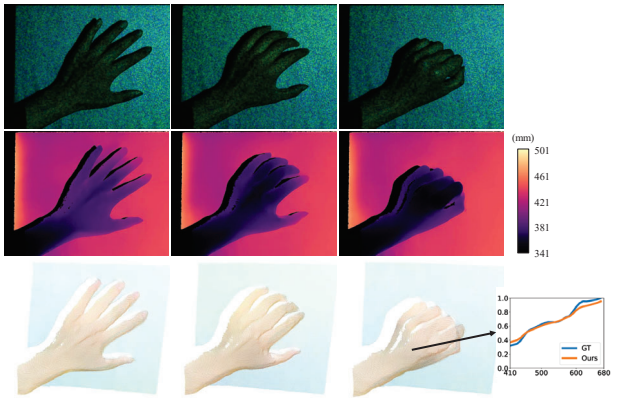


Figure 6. The reconstruction results for a dynamic scene (hand). Top: The sequentially captured input color-dot images. Middle: The estimated depth maps from each input image. Bottom: The 3D point clouds converted from each estimated depth map and the spectral reflectance result for one sample point (rightmost). Each 3D point is colored by sRGB, which is converted from the corresponding estimated spectral reflectance.

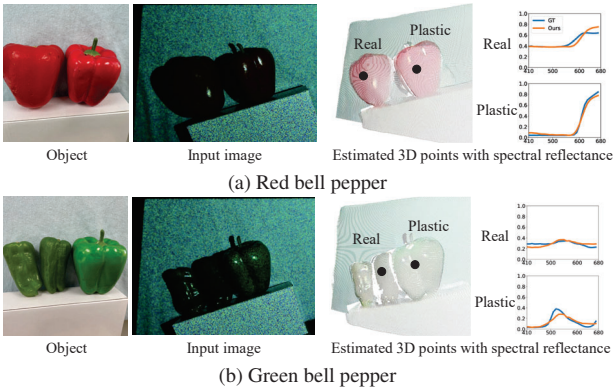


Figure 7. Which is real? We captured real (a) red and (b) green bell peppers, as well as plastic models, to show an example application for material discrimination. From left to right, we show the object images for reference, the input images to our system, the resultant 3D point clouds with sRGB color representation, and the estimated spectral reflectances for sampled points.

which is converted from the corresponding estimated spectral reflectance. We also show the spectral reflectance result for one sample point in the right-bottom figure. From the results, we can confirm that our method performs well in the dynamic real scene.

One important application of the spectral reflectance reconstruction is to differentiate the materials that have similar colors, but different spectral reflectances, because the spectral reflectances provide much richer information than the RGB tristimulus values. To demonstrate this, we captured real and plastic bell peppers. As the objects shown in the left column of Fig. 7, it is hard to differentiate the real and the plastic bell peppers only from the color appearance. In contrast, we can confirm the difference of the spectral reflectances from our spectral reconstruction results, as shown in the right column. Beyond that, our system can reconstruct dense 3D points using the estimated depth maps. Additional results on real scenes reconstructed by our proposed system can be seen in the supplemental video.

4.4. Discussion and Limitations

To offer a theoretical insight on the spectral reflectance estimation, we conducted a condition-number analysis on the system matrix consisting of the products of the projector's RGB illumination spectrums and RGB camera sensitivities. The condition number is 1275.3, which indicates that the direct linear inverse problem is highly ill-posed. As commonly performed [12, 29], if we introduce a spectral basis model (e.g., 8 bases) and a smoothness constraint to the spectral reflectance, the condition number reduces to 13.0, which means that the problem is solvable. Although we solved the ill-posed problem without such constraints by exploiting deep-learning-based reconstruction, we consider that bridging the theoretical analysis and the learning-based method could be one of the important future directions.

Our method has several limitations. First, our method will degrade the performance under the existence of strong ambient illumination because it makes the color-dot extraction by LCN more difficult and it also changes the illumination spectrum of the color dot. Second, heavy occlusions will lead to a large area of cast shadow, resulting in a highly incomplete depth map. Third, similar to other structured-light methods, our method is difficult to reconstruct the dark objects that do not reflect the projector light sufficiently.

5. Conclusion

In this paper, we have proposed a novel single-shot system to simultaneously acquire scene depth and spectral reflectance using a standard RGB camera and an off-the-shelf projector. Our system utilizes a single color-dot projection to simultaneously provide geometric and spectral observations. To effectively reconstruct the depth and the spectral reflectance in a joint training manner, we have built an end-to-end deep neural network architecture by incorporating a geometric color-dot pattern loss and a photometric spectral reflectance loss. Experimental results using both synthetic and real-world data have demonstrated the potential of our system for a high-fidelity 3D sensing technology. Our dataset and spectral renderer for the dataset generation are available in our project page (http://www.ok.sc.e.titech.ac.jp/res/DHD/).

Acknowledgment This work was partly supported by JSPS KAKENHI Grant Number 17H00744 and 21K17762. We thank Tatsuhiko Tezuka for assisting in our experiments.

References

[1] Munsell colors matt. https://sites.uef.fi/spectral/munsell-colors-matt-spectrofotometer-measured/. 5

[2] Seung-Hwan Baek and Felix Heide. Polka lines: Learning structured illumination and reconstruction for active stereo. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5757–5767, 2021. 2

[3] Seung-Hwan Baek, Hayato Ikoma, Daniel S Jeon, Yuqi Li, Wolfgang Heidrich, Gordon Wetzstein, and Min H Kim. Single-shot hyperspectral-depth imaging with learned diffractive optics. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 2651–2660, 2021. 1

[4] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4009–4018, 2021. 6

[5] Nicola Brusco, S Capeleto, M Fedel, Anna Paviotti, Luca Poletto, Guido Maria Cortelazzo, and G Tondello. A system for 3D modeling frescoed historical buildings with multispectral texture information. *Machine Vision and Applications*, 17(6):373–393, 2006. 1

[6] Camille Simon Chane, Alamin Mansouri, Franck S Marzani, and Frank Boochs. Integration of 3D and multispectral data for cultural heritage applications: Survey and perspectives. *Image and Vision Computing*, 31(1):91–102, 2013. 1

[7] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. *arXiv preprint: 1512.03012*, 2015. 5

[8] Cui Chi, Hyunjin Yoo, and Moshe Ben-Ezra. Multi-spectral imaging by optimized wide band illumination. *Int. Journal of Computer Vision*, 86:140–151, 2010. 2

[9] Elkin Díaz, Jaime Meneses, and Henry Arguello. Hyperspectral + depth imaging using compressive sensing and structured light. In *Proc. of 3D Image Acquisition and Display: Technology, Perception and Applications*, pages 3M3G–6, 2018. 1, 2

[10] Sean Ryan Fanello, Christoph Rhemann, Vladimir Tankovich, Adarsh Kowdle, Sergio Orts Escolano, David Kim, and Shahram Izadi. Hyperdepth: Learning depth from structured light without matching. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 5441–5450, 2016. 2

[11] David Hafner, Oliver Demetz, and Joachim Weickert. Why is the census transform good for robust optic flow computation? In *Proc. of Int. Conf. on Scale Space and Variational Methods in Computer Vision*, pages 210–221, 2013. 4

[12] Shuai Han, Imari Sato, Takahiro Okabe, and Yoichi Sato. Fast spectral reflectance recovery using DLP projector. *Int. Journal of Computer Vision*, 110(2):172–184, 2014. 2, 6, 8

[13] Stefan Heist, Chen Zhang, Karl Reichwald, Peter Kühmstedt, Gunther Notni, and Andreas Tünnermann. 5D hyperspectral imaging: Fast and accurate measurement of surface shape and spectral characteristics using structured light. *Optics Express*, 26(18):23366–23379, 2018. 1, 2

[14] Hironori Hidaka, Yusuke Monno, and Masatoshi Okutomi. Spectral reflectance estimation using projector with unknown spectral power distribution. In *Proc. of Color Imaging Conference (CIC)*, pages 205–209, 2020. 2

[15] Keita Hirai, Ryosuke Nakahata, and Takahiko Horiuchi. Measuring spectral reflectance and 3D shape using multi-primary image projector. In *Proc. of Int. Conf. on Image and Signal Processing (ICISP)*, pages 137–147, 2016. 1, 2

[16] Kevin Jarrett, Koray Kavukcuoglu, Marc'Aurelio Ranzato, and Yann LeCun. What is the best multi-stage architecture for object recognition? In *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, pages 2146–2153, 2009. 3

[17] Jun Jiang, Dengyu Liu, Jinwei Gu, and Sabine Süsstrunk. What is the space of spectral sensitivity functions for digital color cameras? In *Proc. of Workshop on Applications of Computer Vision (WACV)*, pages 168–179, 2013. 5

[18] N. Kanopoulos, N. Vasanthavada, and R.L. Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of Solid-State Circuits*, 23(2):358–367, 1988. 4

[19] Min H Kim, Holly Rushmeier, John Ffrench, Irma Passeri, and David Tidmarsh. Hyper3D: 3D graphics software for examining cultural artifacts. *ACM Journal on Computing and Cultural Heritage*, 7(3):14:1–19, 2014. 1

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint: 1412.6980*, 2014. 5

[21] Masahiro Kitahara, Takahiro Okabe, Christian Fuchs, and Hendrik PA Lensch. Simultaneous estimation of spectral reflectance and normal from a small number of images. In *Proc. of Int. Conf. on Computer Vision Theory and Applications (VISAPP)*, pages 303–313, 2015. 1, 2

[22] Chunyu Li, Yusuke Monno, Hironori Hidaka, and Masatoshi Okutomi. Pro-Cam SSfM: Projector-camera system for structure and spectral reflectance from motion. In *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, pages 2414–2423, 2019. 1, 2, 5

[23] Chunyu Li, Yusuke Monno, and Masatoshi Okutomi. Spectral MVIR: Joint reconstruction of 3D shape and spectral reflectance. In *Proc. of Int. Conf. on Computational Photography (ICCP)*, pages 1–12, 2021. 1, 2

[24] Jiaojiao Li, Chaoxiong Wu, Rui Song, Yunsong Li, and Fei Liu. Adaptive weighted attention network with camera spectral sensitivity prior for spectral reconstruction from RGB images. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 462–463, 2020. 6

[25] Haida Liang, Andrei Lucian, Rebecca Lange, Chi Shing Cheung, and Bomin Su. Remote spectral imaging with simultaneous extraction of 3D topography for historical wall paintings. *ISPRS Journal of Photogrammetry and Remote Sensing*, 95:13–22, 2014. 1

[26] Jie Liang, Ali Zia, Jun Zhou, and Xavier Sirault. 3D plant modelling via hyperspectral imaging. In *Proc. of IEEE Int. Conf. on Computer Vision Workshops (ICCVW)*, pages 172–177, 2013. 1

[27] Giljoo Nam and Min H Kim. Multispectral photometric stereo for acquiring high-fidelity surface normals. *IEEE Computer Graphics and Applications*, 34(6):57–68, 2014. 1, 2

[28] Keisuke Ozawa, Imari Sato, and Masahiro Yamaguchi. Hyperspectral photometric stereo for a single capture. *Journal of the Optical Society of America A*, 34(3):384–394, 2017. 1, 2

[29] Jong-Il Park, Moon-Hyun Lee, Michael D. Grossberg, and Shree K. Nayar. Multispectral imaging using multiplexed illumination. In *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, pages 1–8, 2007. 2, 8

[30] Adam Polak, Timothy Kelman, Paul Murray, Stephen Marshall, David JM Stothard, Nicholas Eastaugh, and Francis Eastaugh. Hyperspectral imaging combined with data classification techniques as an aid for artwork authentication. *Journal of Cultural Heritage*, 26:1–11, 2017. 1

[31] Gernot Riegler, Yiyi Liao, Simon Donne, Vladlen Koltun, and Andreas Geiger. Connecting the dots: Learning representations for active monocular depth estimation. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7624–7633, 2019. 2, 3, 4, 5, 6

[32] Hoover Rueda-Chacon, Juan F Florez, Daniel Leo Lau, and Gonzalo R Arce. Snapshot compressive ToF+spectral imaging via optimized color-coded apertures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 42(10):2346–2360, 2020. 1, 2

[33] Marjan Shahpaski, Luis Ricardo Sapaico, Gaspard Chevassus, and Sabine Susstrunk. Simultaneous geometric and radiometric calibration of a projector-camera pair. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 4885–4893, 2017. 5

[34] Lizhi Wang, Zhiwei Xiong, Guangming Shi, Wenjun Zeng, and Feng Wu. Simultaneous depth and spectral imaging with a cross-modal stereo system. *IEEE Trans. on Circuit and Systems for Video Technology*, 28(3):812–817, 2016. 1, 2

[35] Kate Devlin1 Alan Chalmers1 Alexander Wilkie and Werner Purgathofer. Tone reproduction and physically based spectral rendering. *Eurographics*, 2002. 1

[36] Zhiwei Xiong, Lizhi Wang, Huiqun Li, Dong Liu, and Feng Wu. Snapshot hyperspectral light field imaging. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 3270–3278, 2017. 1, 2

[37] Yibo Xu, Anthony Giljum, and Kevin F Kelly. A hyperspectral projector for simultaneous 3D spatial and hyperspectral imaging via structured illumination. *Optics Express*, 28(20):29740–29755, 2020. 1, 2

[38] Yinda Zhang, Sameh Khamis, Christoph Rhemann, Julien Valentin, Adarsh Kowdle, Vladimir Tankovich, Michael Schoenberg, Shahram Izadi, Thomas Funkhouser, and Sean Fanello. ActiveStereoNet: End-to-end self-supervised learning for active stereo systems. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 784–801, 2018. 2, 3

[39] Kang Zhu, Yujia Xue, Qiang Fu, Sing Bing Kang, Xilin Chen, and Jingyi Yu. Hyperspectral light field stereo matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 41(5):1131–1143, 2018. 1, 2