IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 16, 2018, accepted May 18, 2018, date of publication May 24, 2018, date of current version June 26, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2840138

Demultiplexing Colored Images for Multispectral Photometric Stereo via Deep Neural Networks

YAKUN JU[1], LIN QI[1], HUIYU ZHOU[2], JUNYU DONG[1], (Member, IEEE), AND LIANG LU[1]
[1]Department of Computer Science and Technology, Ocean University of China, Qingdao 266071, China
[2]Department of Informatics, University of Leicester, Leicester LE1 7RH, U.K.

Corresponding author: Junyu Dong (dongjunyu@ouc.edu.cn)

This work was supported in part by the International Science and Technology Cooperation Program of China under Grant 2014DFA10410, and in part by the National Natural Science Foundation of China under Grant 61501417 and 41576011. The work of H. Zhou was supported in part by UK EPSRC under Grants EP/N508664/1, EP/R007187/1, and EP/N011074/1, and in part by the Royal Society-Newton Advanced Fellowship under Grant NA160342.

ABSTRACT Recovering fine-scale surface shapes is a challenging task in computer vision. Multispectral photometric stereo is one of the popular methods as it can handle non-rigid/moving objects and produces per-pixel dense results. However, the colored images captured by practical multispectral photometric stereo setups are aliased in RGB channels. Existing solutions require prior information to calibrate few points and estimates whole surface normal by the calibration, while prior information is not always available and accurate. Differing from previous solutions which require calibration or other prior information, we first formulate the problem in a learning framework, which directly seeks the per-pixel mapping of the aliased and spectrum-multiplexed pixel response to the anti-aliased and demultiplexed counterpart. In this paper, we propose to use a novel deep neural networks framework as the ''demultiplexer''. By using ''demultiplexer'' and classic photometric stereo, our method can reconstruct a dense and accurate surface normal from a single-frame colored image without any prior information nor extra information injected. We build an imaging device to collect images of different materials under colored lights and white lights. We conducted extensive experiments on our data set and a public data set. The results show that the proposed fully connected network successfully demultiplexes the colorful image and produces satisfactory surface estimation.

INDEX TERMS Multispectal photometric stereo, spectrum demultiplexing, normal estimation, deep neural networks.

I. INTRODUCTION

Recovering 3D shapes of objects is a challenging problem in computer vision. In the past few decades, many reconstruction algorithms and improvements were proposed, including photometric stereo [1], structured light [2], binocular stereo vision [3], [4], and structure from motion [5]–[7]. Among these methods, photometric stereo is highlighted by its per-pixel resolutions and finer reconstruction details. However, traditional photometric stereo methods require multiple images captured with different illumination directions while the camera and the target object should hold stationary. This limits its use in dynamic applications such as non-rigid objects and moving objects. To solve this problem, multispectral photometric stereo technologies were proposed [8], [9], where three colorful lights (red, green and blue) simultaneously illuminate the target from different directions and one single color image is captured. Ideally, the image intensity in each RGB channel corresponds to the reflected radiance from the respective colorful lights. From the communication perspective, the traditional photometric stereo takes a time-division multiplexing strategy, whereas the multispectral photometric stereo takes a spectral-division multiplexing strategy. In practice, the intensity in each channel of the captured image is aliased, which is the tangle of illumination, surface reflectance and camera response (more details will be discussed in the later sections). There is no satisfactory solutions to demultiplex this aliased signal yet, and the existing methods utilize image itself [8], employ pre-calibration [10] or require initial surface normal [11].

In many cases, such pre-calibration or initial inputs are expensive or impractical. These previous methods use priori information to calibrate the reflectance properties and

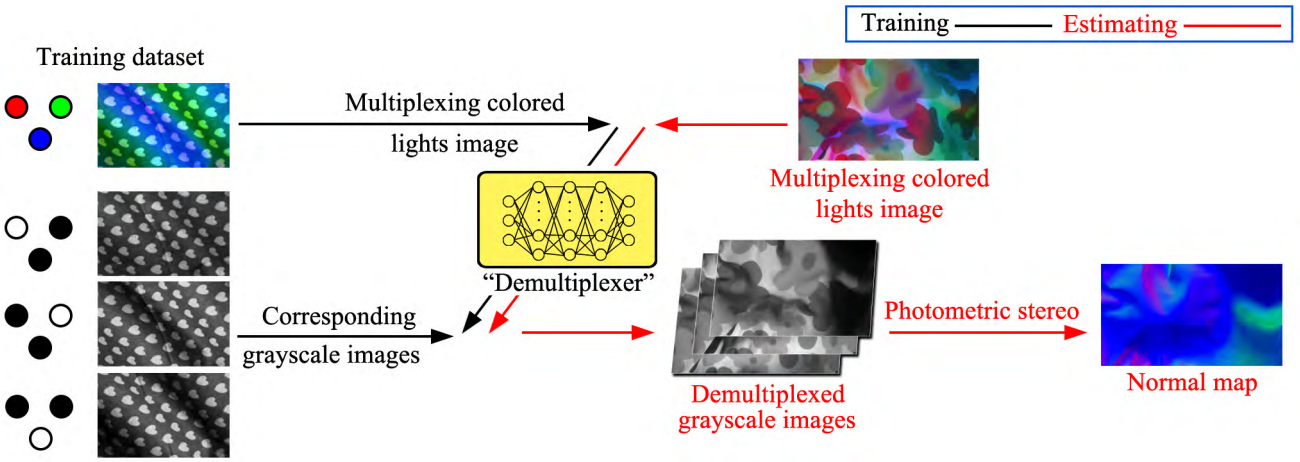

FIGURE 1. The overview of the proposed method. It consists of training and estimating stages. The circles on the left of the dataset represent the illumination situations, the color of circles represent the color of illumination, while black circles mean the lights are turn off. The first one means the red, green and blue colored lights turn on simultaneously, and the last three mean turn on one white light respectively.

illumination of some pixels, and use these pixels to estimate the whole surface normal, while prior information is not always available and accurate. It is therefore still an open question to accurately and robustly recover surface shapes only from a single colorful image in multispectral photometric stereo. Is it possible to transform a single colored image into three grayscale images as captured by three lights under different directions? We therefore propose to build a "demultiplexer" which can divide a single colored image into three grayscale images illuminated by the three white lights at the same position with minimum errors. In this paper, we employ deep neural networks (DNN) as the "demultiplexer." DNN has been broadly used for various computer vision tasks and has shown state-of-the-art performance due to its powerful learning capability. Researchers in 3D reconstruction have used it to estimate surface normal [12], [13] and scene depth [14]. Motivated by these successes, we here propose to use it as "demultiplexer."

Given an image of a target object illuminated by three colored lights (red, green and blue), we want to learn a per-pixel mapping to three corresponding grayscale images of the object illuminated by white lights rather than calibrating some points by priori information. We can then recover surface normals using traditional photometric stereo techniques. The overview of the proposed method is shown in Figure 1. This mapping was realized by employing the high regression power of DNN. We collected an image dataset for training and estimating the network, which consists of many materials under colored lights and white lights. The illumination situations are shown on the left of the examples. The images of the colored lights are the input, and the corresponding images of the white lights are the output. A public dataset DiLiGent [15] was also used to test the generalization of our method.

The main contributions of this work are the following:

1) A learning based framework is proposed to solve the demultiplexing problem in multispectral photometric stereo firstly, establishing the per-pixel mapping of the aliased and spectral multiplexed pixel response to the anti-aliased and demultiplexed counterpart.
2) A novel fully connected deep neural network is proposed as the "demultiplexer". Given an image of the target object illuminated by colorful lights, the network predicts three grayscale images of the object illuminated by an identical light separately.
3) A new and a modified image dataset are built for the multispectral photometric stereo process, which were used to train the proposed network and evaluate the existing methods.

The rest of the paper is organized as follows. We introduce the related work in Section 2, followed by the basic theory summarized in Section 3. Section 4 presents our proposed method. The training datasets for multispectral photometric stereo are introduced in Section 5. Section 6 reports the experimental results. Section 7 concludes this paper.

II. RELATED WORK

Although traditional photometric stereo [1] provides dense surface reconstruction, it cannot well handle deforming surfaces or dynamic scenes. To solve this problem, multispectral photometric stereo was proposed [9]. However, the problem in multispectral photometric stereo is that the intensity in each channel of the captured image is aliased, related to the tangle of illumination, surface reflectance and camera response. Early methods demultiplexed the aliasing using the image itself [8], but it is difficult to handle multicolor objects. Kim *et al.* [16] optimized system implementation exploiting the physical properties of typical cameras and LEDs.

Hernandez *et al.* [10] utilized a calibration method which is planar with special marking that allows the plane orientation to be estimated. By placing the fabric in the center of the tool, accurate surface normal can be obtained using this pre-calibration method. However, this method requires

to calibrating every kind of reflectance properties on the surface as prior knowledge. On the one hand, this calibration is very complex and needs very high environmental requirements. On the other hand, it is difficult to separate the surface with different reflectance properties in practical application.

Other than calibration, some researchers employed a coarse surface estimation as the initial input and iteratively searched for an optimized solution, where the initial input can be the depth obtained by Kinect or binocular stereo [11], [17], [18]. However, the initial input is not always available and in some circumstances is inaccurate, such as underwater.

Previous methods heavily rely on prior information to solve the aliasing problem in multispectral photometric stereo. The challenge is to robustly and accurately estimate surface normals without such priors. There is no close-form solution to this ill-posed problem yet, because the single colorful image is not enough to calculate all the unknown parameters.

Instead of solving the underdetermined problem directly, many works adopted learning-based solutions. Santo *et al.* [13] proposed the Deep Photometric Stereo Network (DPSN) to estimate surface normal as a regression learning problem rather than the traditional constrained model. A dropout layer was used to handle shadows and non-Lambertian conditions. Nguyen *et al.* [19] and Jia *et al.* [20] used neural networks to predict hyperspectral image from an rgb-channels image.

In this paper, we employed deep neural network to demultiplex the color image in multispectral photometric stereo. We propose a novel fully connected network as the ''demultiplexer'', which outputs three corresponding grayscale images from the colorful image. Then the normal map of the object can be recovered using a traditional three-source photometric stereo method.

III. THEORY BACKGROUND

Unless otherwise stated, we use boldfaced uppercase and lowercase letters to denote matrices and column vectors respectively.

When we take a picture of a Lambertian surface illuminated from direction $\mathbf{l}$, the intensity of pixel (x, y) in each channel can be written as:

$$c_i = \mathbf{l}^{\mathbf{T}}\mathbf{n} \int E(\lambda)R(\lambda)S_i(\lambda)d\lambda \tag{1}$$

where c_i is the intensity of pixel (x, y) in channel i ($i \in \{\mathrm{r}, \mathrm{g}, \mathrm{b}\}$), $E(\lambda)$ represents the energy distribution of illumination as a function of wavelength λ, $R(\lambda)$ represents the spectral reflectance function of the surface., $S_i(\lambda)$ is the camera sensor for channel i and $\mathbf{n}$ is the surface normal at pixel (x, y).

As $E(\lambda)$ and $S(\lambda)$ are difficult to measure, we usually use the 'scaled albedo' $\boldsymbol{\rho}$ to represent the integration in Eq.1:

$$c_i = \rho_i \mathbf{l}^{\mathbf{T}}\mathbf{n} \tag{2}$$

The normal $\boldsymbol{n}$ can be recovered given at least three images captured under illuminations whose directions are non-coplanar.

In multispectral photometric stereo, the objects are illuminated simultaneously by three non-coplanar lights which are red (R), green (G) and blue (B). The direction of the three lights can be written as $\mathbf{l}_{\boldsymbol{k}}$ ($k \in \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}$). Let $E_k(\lambda)$ ($k \in \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}$) be the energy distribution of three lights respectively. Then, the intensity of the pixel (x, y) in channel i can be expressed as the summation of the contributions from all the lights [8]:

$$c_i = \sum_k \mathbf{l}_{\boldsymbol{k}}^{\mathbf{T}}\mathbf{n} \int E_k(\lambda)R(\lambda)S_i(\lambda)d\lambda \tag{3}$$

Let $\rho_{ik} = \int E_k(\lambda)R(\lambda)S_i(\lambda)d\lambda$, where ρ_{ik} represents the $(i, k)_{th}$ element of the matrix $\mathbf{P_m}$. c_i can be combined into a vector $\mathbf{c}$=$[c_\mathrm{r}, c_\mathrm{g}, c_\mathrm{b}]$. $\mathbf{l}_{\boldsymbol{k}}$ can be combined into a matrix $\mathbf{L}$=$[\mathbf{l_R}, \mathbf{l_G}, \mathbf{l_B}]^{\mathrm{T}}$. Then, Eq.3 can be written as follows:

$$\mathbf{c} = \mathbf{P_m L n} \tag{4}$$

When a Lambertian surface with normal $\mathbf{n}$ is illuminated by three identical lights $\mathbf{L_w}$=$[\mathbf{l_1}, \mathbf{l_2}, \mathbf{l_3}]^{\mathrm{T}}$ respectively, the measurement $\mathbf{c_w}$ can be described as follows:

$$\mathbf{c_w} = \mathbf{P_d L_w n} \tag{5}$$

where $\mathbf{c_w}$=$[c_1, c_2, c_3]$, c_1, c_2, c_3 represent the grayscale value of pixel (x, y) in the image illuminated by $\mathbf{l_1}$, $\mathbf{l_2}$, $\mathbf{l_3}$ respectively. $\mathbf{P_d}$=$\boldsymbol{diag}_{\mathbf{3\times 3}}(||\mathbf{L}^{-\mathbf{1}}\mathbf{c_w}||)$ is used to substitute $\mathbf{P_m}$ [1].

IV. PROPOSED METHOD

If we used classic photometric stereo to calculate the colored image directly, the results would be wrong. We cannot demultiplex the colored image by directly separating the three channels. In fact, the intensity in each channel is not only affected by the corresponding colored light but also contaminated by the other two colorful lights. We show this problem by feeding the colored image as three grayscale images (by channel) to the classic photometric stereo. Figure 2 shows the results. It can be seen the errors in the recovered normal are significant.

To establish a per-pixel ''demultiplexer'', we change the illumination from RGB lights to white lights while ensuring that there is no change in the relative position between the camera, lights and the target. Let white lights illuminate the surface respectively, then the measurement $\mathbf{c_w}$ can be described as Eq.5. Since we do not change the direction of the illuminations, we can replace $\mathbf{L}$ with $\mathbf{L_w}$. Therefore, the measurement $\mathbf{c_w}$ can be written as:

$$\mathbf{c_w} = \mathbf{P_d P_m^{-1} c} \tag{6}$$

We confirm that every surface reflectance property corresponds to an unique demultiplexing matrix $\mathbf{\Phi}$=$\mathbf{P_d P_m^{-1}}$. Thus, a straight thought is to build a mapping function f: $\mathbb{R}^3 \rightarrow \mathbb{R}^3$, which can transform the single multispectral colored image $\mathbf{c}$

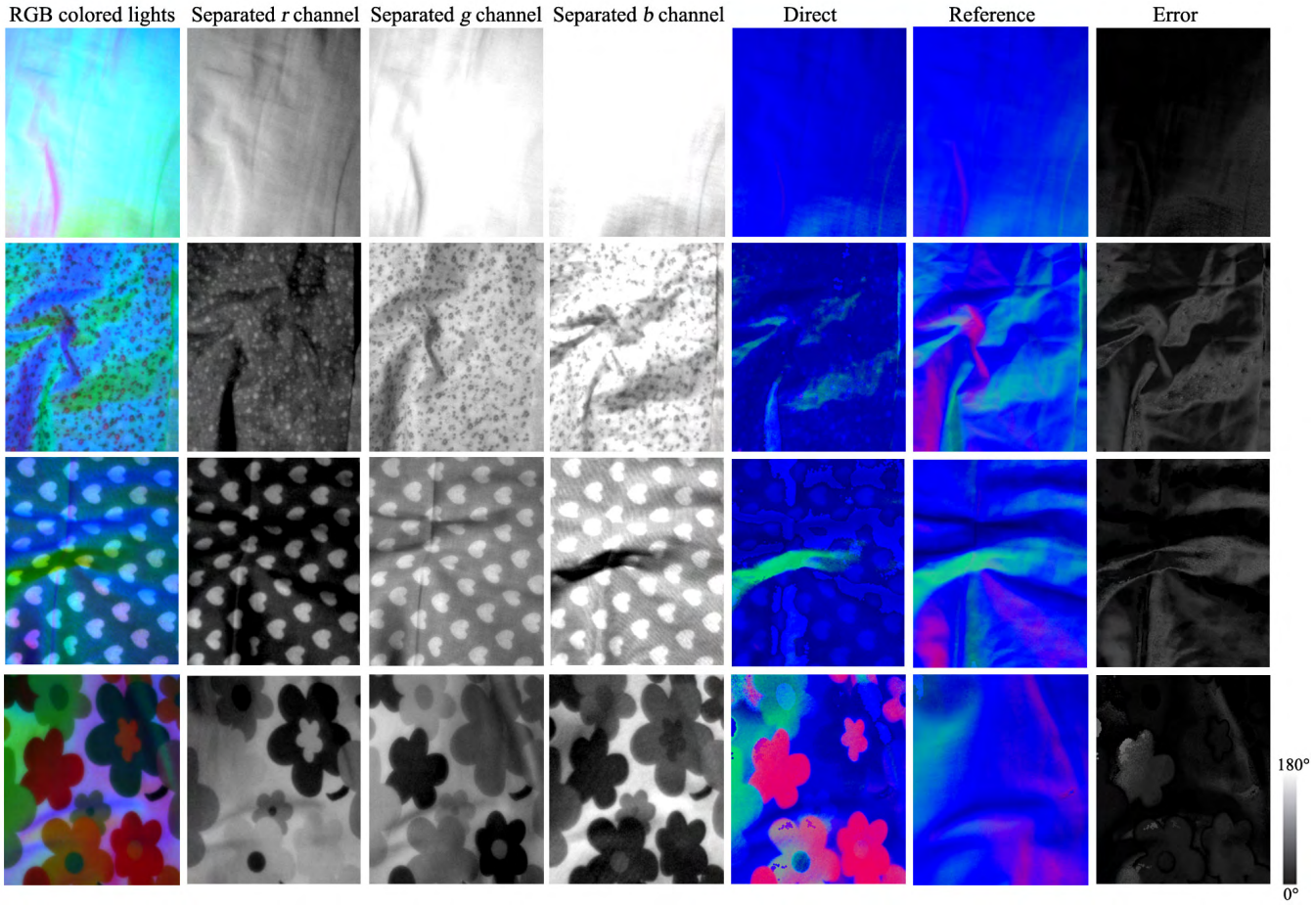

FIGURE 2. Fabric examples under the tri-colored light. In the experiment, we choose four kinds of fabric including white fabric, floral fabric, dichromatic and fabric with color patches. 'Direct' means photometric stereo using multiplexing tri-colored lights image directly, while the Reference means photometric stereo using three white lights illuminated serially on the same position of the tri-colored lights.

into three grayscale images $\mathbf{c_w}$ illuminated by the three white lights at the same position pixel by pixel. Faced with the mapping containing almost unlimited surface reflectance properties, there is a complex nonlinear relationship between $\mathbf{c}$ and $\mathbf{c_w}$. The classical machine learning algorithms such as support vector regression have limited the generalization ability to represent complex problems with massive samples, while DNN can be used to approximate such mapping function obviously. Through experimental validation, we employ the DNN to learn a nonlinear transformation f.

A. NETWORK ARCHITECTURE

We present the details of our model in this section. Like DPSN [13], we employ fully connected deep neural networks to learn the per-pixel mapping from $\mathbf{c}$ to $\mathbf{c_w}$ which fits demultiplexing matrix $\mathbf{\Phi}$.

The structure of our model is summarized in Figure 3. The model consists of 7 fully connected layers. Each fully connected layer includes a sigmoid activation function. We introduce direct connections from the input layer to the subsequent fully connected layers inspired by DenseNet [21]. Concatenating the input layer improves the robustness of the networks. Furthermore, we observe that concatenation can accelerate training convergence and enhance training optimization. A more detailed discussion will be shown in the network architectures' comparison.

Our model is trained with mean squared errors (MSE) as the loss function:

$$\mathcal{L}(\beta) = \frac{1}{n}\sum_{i=1}^{n} ||f(\mathbf{x_i}, \beta) - \mathbf{y_i}||^2 \tag{7}$$

where n is the number of the training samples, each $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$ is a pair of $\mathbf{c}$ vector and corresponding $\mathbf{c_w}$ in the training set, and β are the network parameters that require to be estimated for the $f(\mathbf{x}, \beta)$ to fit the training pairs $(\mathbf{x_i}, \mathbf{y_i})$. The loss is minimized using stochastic gradient descent with the standard backpropagation [22].

B. PREDICTION AND CALCULATE SURFACE NORMAL

In order to train the model, we learn the mapping from multiplexing a tri-colorful lights image $\mathbf{c}$ to white light grayscale images $\mathbf{c_w}$. In the prediction phase, given a set of $\mathbf{c}$, our model

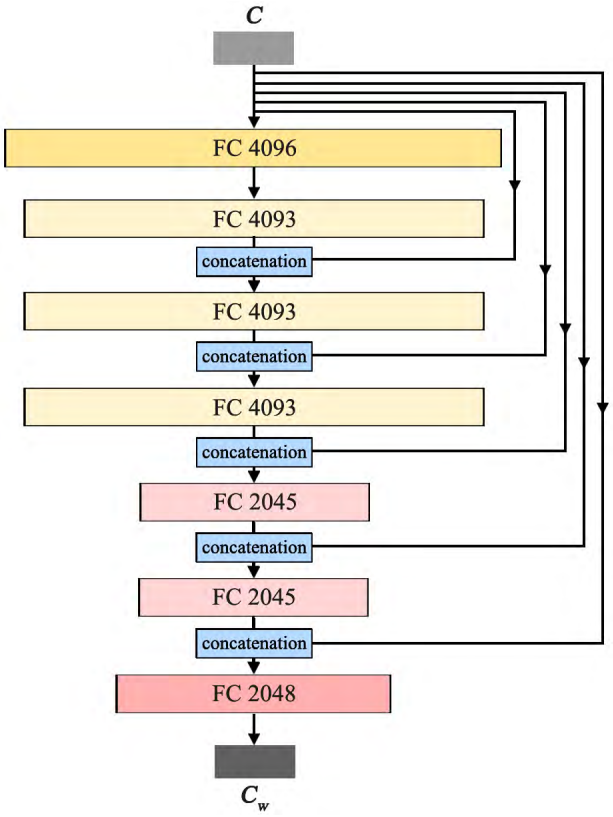

FIGURE 3. The structure of our model. FC means fully connected layers, the number behind FC represents the number of dimension. In concatenation, we concatenate the input layer to the fully connected layer as a totality and propagate it to the next fully connected layer.

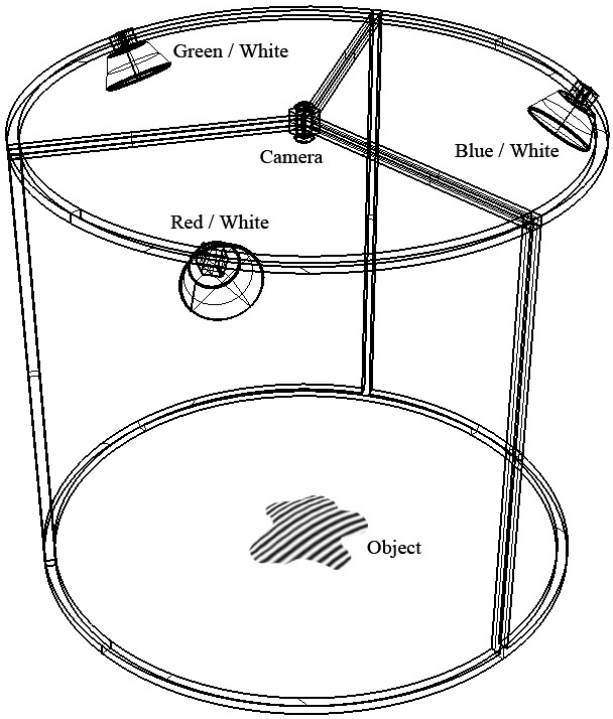

FIGURE 4. Schematic structure drawing of our device.

estimates the demultiplexed $\widehat{\mathbf{c}_\mathbf{w}}$. Our model estimates $\widehat{\mathbf{c}_\mathbf{w}}$ per-pixel and enables them to obtain the integrated grayscale images.

With the estimated $\widehat{\mathbf{c}_\mathbf{w}}$, we can calculate the surface normal. In the proposed method, we use the classic photometric stereo [1]. Although there are many improved algorithms of photometric stereo, using the classic photometric stereo shows better fitting ability of our method.

V. TRAINING DATASET

A. OUR DATASET

The training set for our model consists of colorful light measurement $\mathbf{c}$ and the white lights measurement $\mathbf{c_w}$. The white lights measurement $\mathbf{c_w}$ is composed of three grayscale values illuminated by three white lights respectively. To establish the dataset, we build a lighting device, composed of a camera and three stationary lights. The structure of our device is shown in Figure 4. In each light, there are two kinds of colors which can transform between them. Our device can be used to make the required training samples. Each training sample contains a single image under simultaneous illumination of the colorful lights and three images of single white illumination. We made sure that objects remain motionless under the colorful lights illumination and white light illumination. Since the positions of the colorful lights are fixed, we can collect new training data when rotating or remodeling the object, a large number of training samples can be obtained.

In this dataset, we collect dozens of real-world observations. The observation consists of more than a hundred varieties of colors and materials with diverse normals viewed under pre-set tri-colorful illuminations and white illumination. Figure 5 shows some sample images of the training data. The dataset looks random at the image-level, but it is useful and plenitudinous at the pixel-level training samples.

B. PSEUDO-COLORED LIGHTS DATASET

There is no image dataset particularly for multispectral photometric stereo. In addition to our dataset, it is difficult to find a dataset containing multiplexing tri-colored lights. DiLiGent [15] is a very significant and widely used dataset in the field of photometric stereo. In order to demonstrate the robustness of our method persuasively, we use the original images in DiLiGent to generate the pseudo-colored light images. The reason for selecting DiLiGent is that it contains 96 directions of lights, better simulating our dataset. In addition, images in DiLiGent are rgb tri-channels rather than single channels.

Pseudo-colored light images can simulate the effect of the tri-colorful lights' images. We select three images of one object with different white illuminations, then we extract the r channel in the first image, g channel in the second image and b channel in the third image. Finally, we combine these

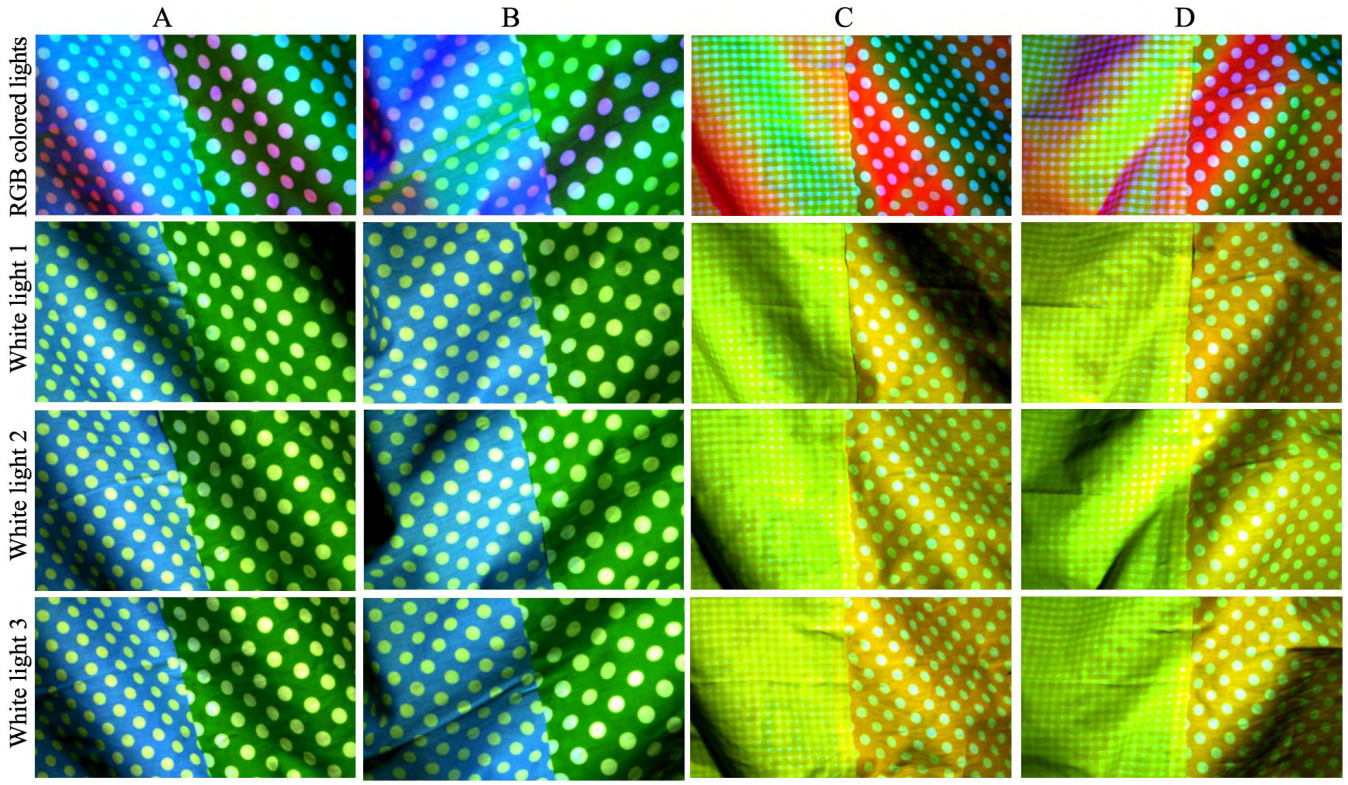

FIGURE 5. Examples of sample images of the training data set. A to D are examples of the dataset. Each training example contains a single image under simultaneous illumination of tri-colored lights and three images of single white illumination. B and D are the remodeling of A and C. The images of different materials and colors under different normal are available through rotation and remodeling.

TABLE 1. The combinatorial process of pseudo-colored lights images.

Pseudo	Extracted r	Extracted g	Extracted b
r	90%	5%	5%
g	5%	90%	5%
b	5%	5%	90%

TABLE 2. SSIM [23] values of 32 examples.

SSIM(min)	SSIM(max)	SSIM(mean)
0.9989	0.9998	0.9992

channels in purpose. In order to simulate the aliasing and multiplexing effects, we combine one main channel and two aliasing channels for each channel in pseudo-colored lights' images. Quantificationally, we ensure the pixel value of each channel in pseudo-colored lights images equals to 90% of the main channel pixel value plus two 5% aliasing channel pixel values, shown in Table 1.

We simulate the aliasing of the colored images with this combination. It is worth noting that our combination simulates the similar effect of multiplexing, but not the same as the real situation, because we do not know the camera's color spectral sensitivities and exact the spectrum of the colored lights.

Therefore, we understand that the inconformity between pseudo-colored lights images and colored lights images is mainly due to the diversity energy distribution between the white illumination and colorful illuminations. We will show experiments below.

Since there is no real colored lights image of DiLiGent, it is difficult to analyze the error between the pseudo-colored lights image and the real colored lights image directly. We use our dataset for verification firstly, comparing pseudo-colored lights image with the real colored light image, as shown in Figure 6 and Table 2.

Figure 6 illustrates three examples, including three kinds of fabric. Each example shows the image, the corresponding normal map and the histogram of the real condition and the pseudo condition respectively. Looking at the first row images, we find that the pseudo-colored light image is greener than the corresponding real tri-colored light image, which means the value of g channel may be large. This phenomenon is caused by the inconsistent spectral energy distribution of the white illumination and the corresponding colorful illumination in the colorful lights. To quantify the extent of this inconsistency, we use Mean Pixel Value Error (MPVE) to calculate the mean deviation of the pixels' value between the pseudo-colored lights image and the real colored light image. The values of MPVE are shown at the bottom of the first row. Also, we show the histogram of each image in the second row. The histograms of the pseudo-colored light image and the real colored light image are similar in general, while there is some difference in a small range. Concretely, we

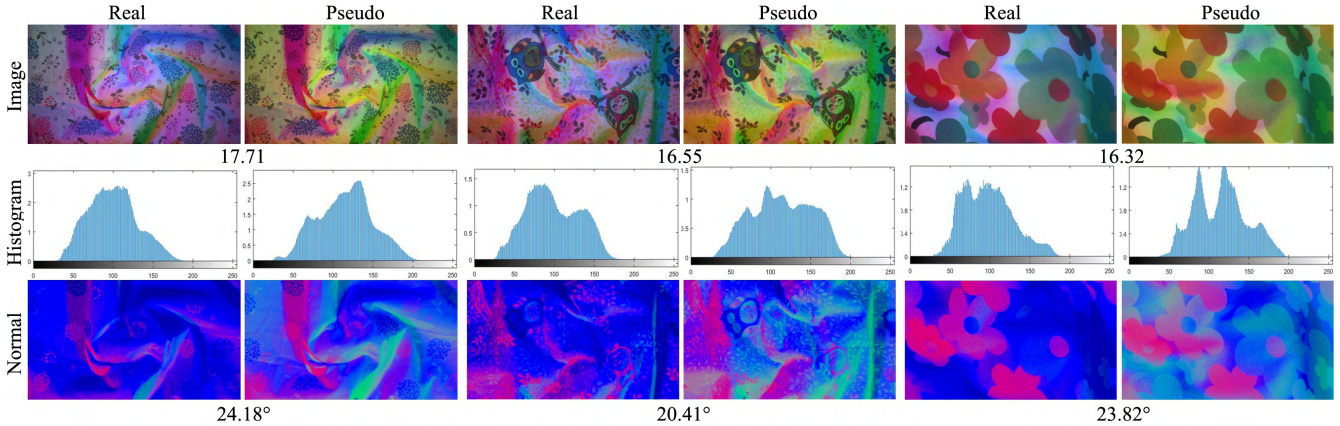

FIGURE 6. Examples of comparison. Here, three out of more than 30 images are shown. Each example includes a real column and pseudo column, which represent real colored lights condition and pseudo-colored lights condition respectively. The first row shows the real colored lights images and pseudo-colored lights images. The numbers at the bottom of the first row are Mean Pixel Value Error (MPVE) between real and pseudo images. The second row shows the histogram of the corresponding first row image. The third row shows the normal maps calculate by photometric stereo using directly three-channels' corresponding image in the first row. The numbers at the bottom of the third row are Mean Angular Error (MAE) in degree of the real and pseudo normal maps.

show the normal maps calculate by photometric stereo using directly the three-channel's corresponding image in the first row. The Mean Angular Error (MAE) of the two normal maps in each example is shown at the bottom of the third row. The values of MAE are 24.18°, 20.41°, 23.82° respectively, which are large. This is because the changes of the pixel values significantly influence the normal in the shape from shading algorithm. The deviation of the surface distortions caused by the inconsistent spectral energy distribution are particularly noticeable on the values of MAE.

However, in our method, we are concerned with the normal error caused by the variety of the reflectance properties. This error is reflected in the discontinuity of the normal compared with the groundtruth. The two kinds of the normal maps in the third row of Figure 6 both show this error. To prove the error is similar between the real colored light image and the pseudo-colored lights image, we introduce Structural Similarity (SSIM) in Table 2. SSIM [23] is a full reference image quality evaluation index which can evaluate the similarity of this error effectively. SSIM is defined as:

$$\mathbf{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{8}$$

where μ_x and μ_y are the means of image x and image y, σ_x and σ_y are the variances of x and y, σ_{xy} is their corresponding covariances, C_1 and C_2 are constants used to keep stability. The SSIM value range is from 0 to 1, the larger the value, the more similar the images. We calculate SSIM values of 32 experimental examples of our dataset, and Table 2 shows the maximum SSIM, minimum SSIM and mean SSIM of all examples. The results show the extent of normal errors caused by the variety of the reflectance properties are almost the same on the normal calculate for the real colored light image and the pseudo-colored light image. Furthermore, this similarity is robust in all the experimental examples.

The above shows the usability of the pseudo-colored lights image. We then use DiLiGent [15] to compose the pseudo-colored lights image dataset. Some examples of the dataset are shown in Figure 7.

VI. EXPERIMENTS

In our experiments, we will firstly verify the ability of our fully connected networks. Furthermore, we evaluate the predicted results of our model with the groundtruth grayscale using mean relative error (MRE) and compare the final normal using MAE both in our dataset and pseudo-colored lights dataset. Several mainstream single frame 3D reconstruction techniques including deep learning methods are selected to deliver the comparisons. We compare the proposed method with multispectral photometric stereo [18] using SLIC super-pixels segmentation [24] which can provide better reconstruction results. To ensure that all the methods do not have additional input information, we use the initial depth obtained by photometric stereo using the direct three-channels of multiplexing tri-colored lights images. In addition, we also compare the proposed method with DPSN [13]. To ensure no additional input information, we reduce the input dimension in DPSN from 96 to 3 representing RGB channels. We also cancel the shadow layer considering only three dimensions of the inputs in the remodeled DPSN.

A. IMPLEMENTATION AND TRAINING SETTINGS

Our model was implemented using TensorFlow, and trained for 15 epochs (our dataset) with the batch size of 100, the leaning rate is to begin with 2×10^{-2} and end with 10^{-5}, performed on an Ubuntu 14.04 machine with Tesla K40c.

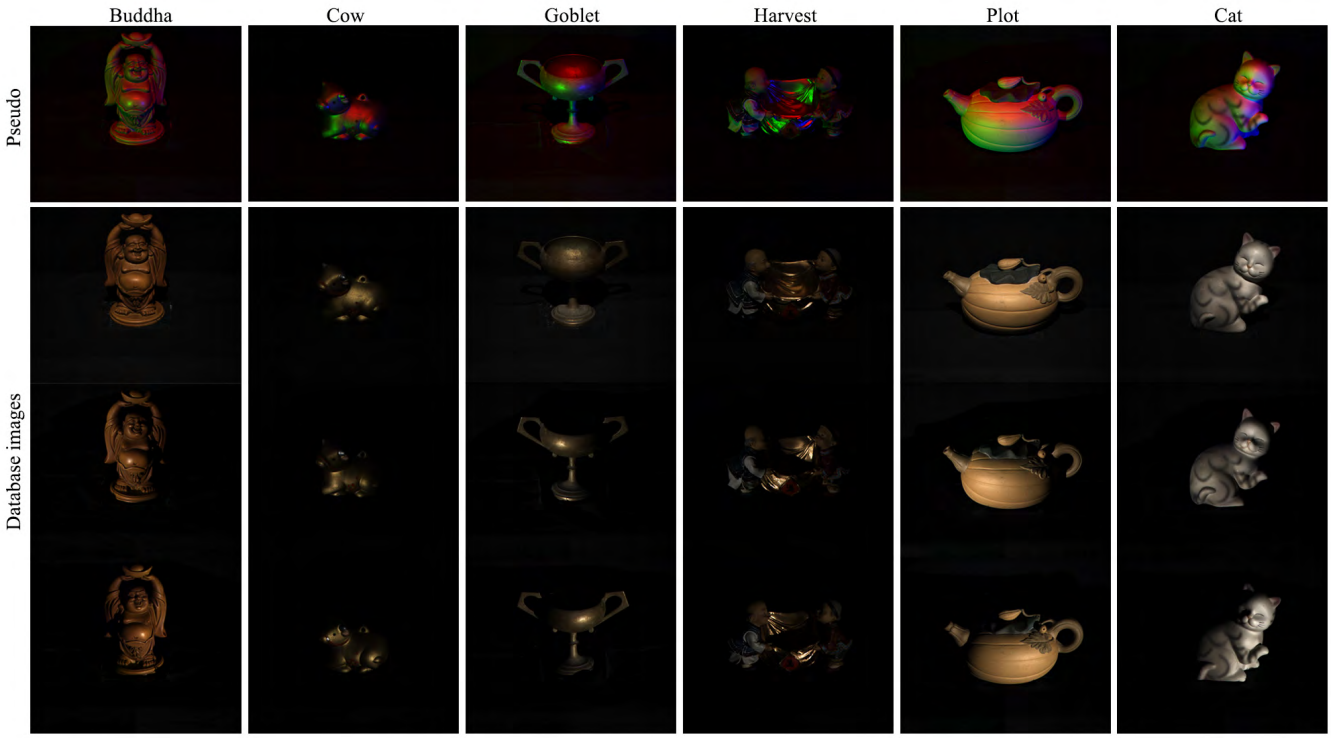

FIGURE 7. Pseudo-colored lights images from DiLiGent. A pseudo-colored lights image is shown on the top of each example.

The model which achieves the highest accuracy for the test data is used for evaluation.

In our dataset, we generated 128×4 images for the training. The resolution of each image is 960×540; then the total number of training pairs in the training set $\{(c, c_w)\}$ becomes about 6.6×10^7.

DiLiGent [15] is a dataset with ten objects, and each object has 96 illumination directions with different intensity. However, for our method we generate only one pseudo-colored image for each object, and the composed white images have the same illumination direction. Due to the limited size of the pseudo-colored lights dataset, we extract approximately 1.6×10^7 useful pixel pairs from ten objects and randomly select 2×10^6 pixel pairs to form the training samples. Accordingly, we increase the epochs to 300 for training the ideal model.

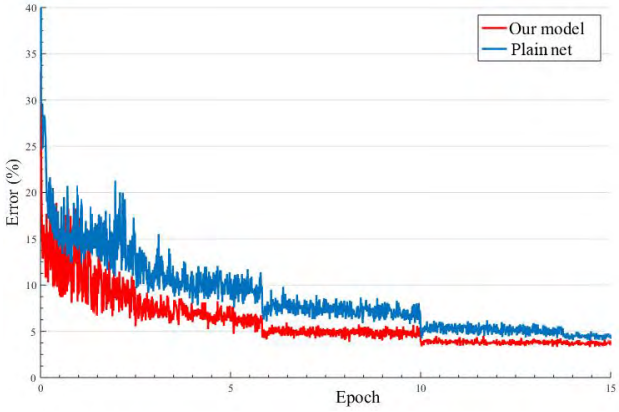

FIGURE 8. Training on our dataset. Red curve denote training error of our model, and blue curve denote training error of corresponding plain net. Two models are trained with same training parameters.

B. NETWORK ARCHITECTURES ANALYSIS

We evaluate our model in comparison with a plain net. The plain net is of a similar architecture to our model. The only difference is that the plain net has no concatenation. The models are trained on our training set with the same training parameters. The results are shown in Figure 8 and Table 3. The training and test errors are calculated by MRE.

We compare the training errors during the training procedure in Figure 8. It shows that our model has a faster convergence rate and a lower error. We can observe that the error curve of the plain net has large fluctuations in the early stage of the training, while our model with concatenation operations has a relatively small and stable error curve at the beginning of the training, and stays ahead until the end of the training. In fact, the differences between $\mathbf{c}$ and $\mathbf{c_w}$ in our dataset are limited. Therefore, the orientation of the concatenating input information will accelerate the convergence, improve the robustness of the system and reduce the integration and entanglement of the output. Compared to the plain net without concatenation, our model reduces the test error by 0.33% as shown in Table 3, resulting from the successfully reduced training error. Concatenating the

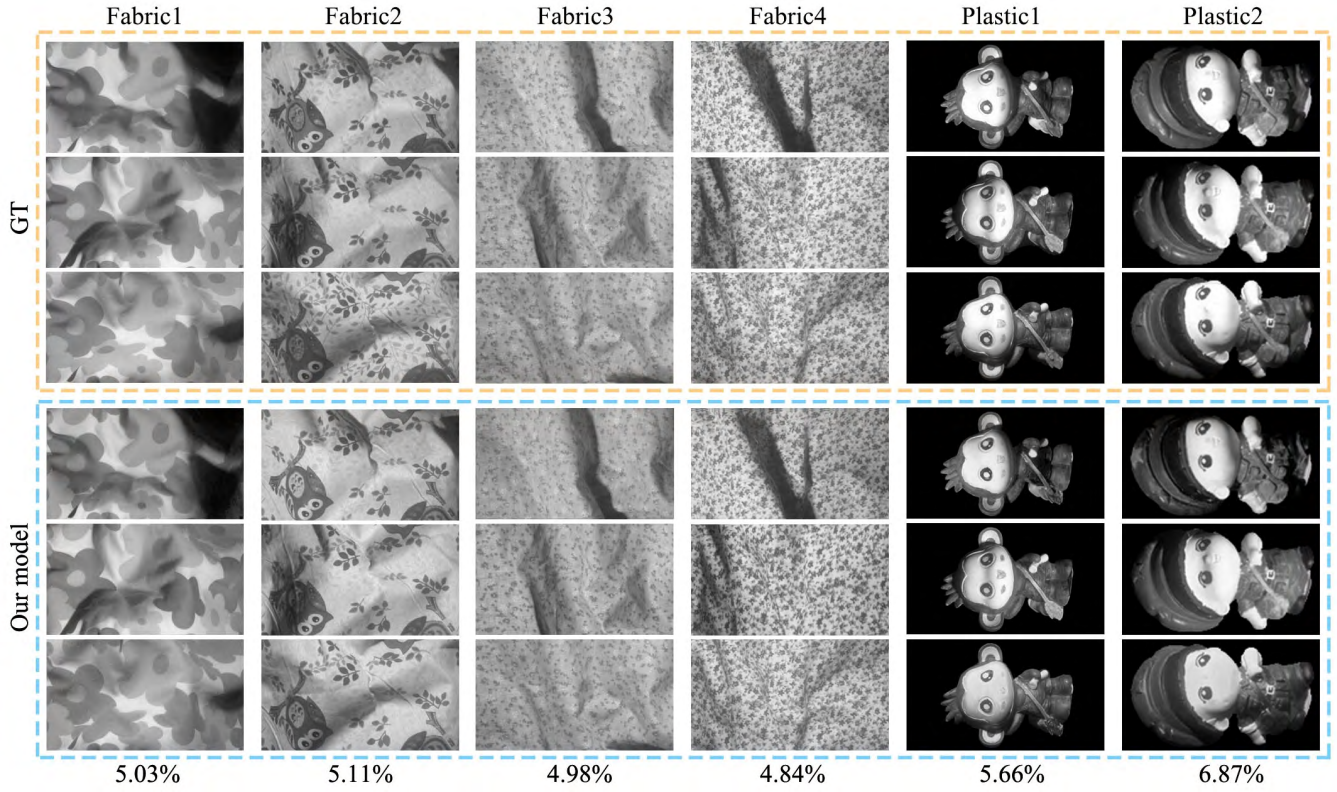

FIGURE 9. Evaluation for our method. The number represents MRE. In each column, the first three images are groundtruth, the three images below the groundtruth are the results of our method. On the top of each column, materials and their names are displayed. The position of the images is adjusted for clear view.

TABLE 3. Test errors of our model (with concatenation) and plain net (without concatenation).

	Our model	Plain net
Error	4.96%	5.29%

input layer improves the robustness of the networks and enhance optimization. We attribute this primarily to the corresponding growth in model capacity. The operation of each layer of the neural networks is equivalent to a nonlinear transformation of the input. With the deeper networks, the complexity of the transformation is also gradually increased. Compared with the plain networks completely dependent on the highest complexity of the last layer, our model can combine and utilize the input information with a smooth output layer and hence achieves better generalization performance. It improves the network optimization ability and anti-overfitting.

We also compare our network with deep convolution neural networks (CNN). We observe that CNN has a positive impact on the scene-level surface normal or depth estimation [14], [25], [26]. However, CNN is not suitable for our task. On the one hand, the feature maps extracted by CNN over-focus on the structure, which is easily influenced by the surface color. It is difficult to distinguish whether the feature maps are activated by the normal or the surface color in fine-level reconstruction. On the other hand, the pixel mapping of the demultiplexing matrix will be only influenced by the pixel itself if inter-reflection is not considered. The surrounding pixel values should not be involved in calculating the mapping for the center point and the demultiplexing matrix is irrelevant to the geometric structure of the object. Thus, using CNN would cause computational waste and even worse results.

C. EVALUATION OF THE PROPOSED METHOD

To the best of our knowledge, this is the first study to introduce the ill-posed demultiplexing task. Thus, we only compare the results produced by our method with groundtruth at the demultiplexing stage. Nevertheless, in the reconstruction experiments, the final output will be the surface normal, and we will compare our surface normal with those produced by several algorithms.

1) OUR DATASET

Firstly, we evaluate the performance of the proposed DNN model using our dataset. Compared with the groundtruth, we show the MRE of the three grayscale images in Figure 9. MRE is the mean ratio of the absolute error to the groundtruth, which better reflects the credibility of the measurement. In this experiment, we do not need to calculate the normal of objects. Therefore, it is reasonable to evaluate the MRE of pixels between groundtruth and our method.

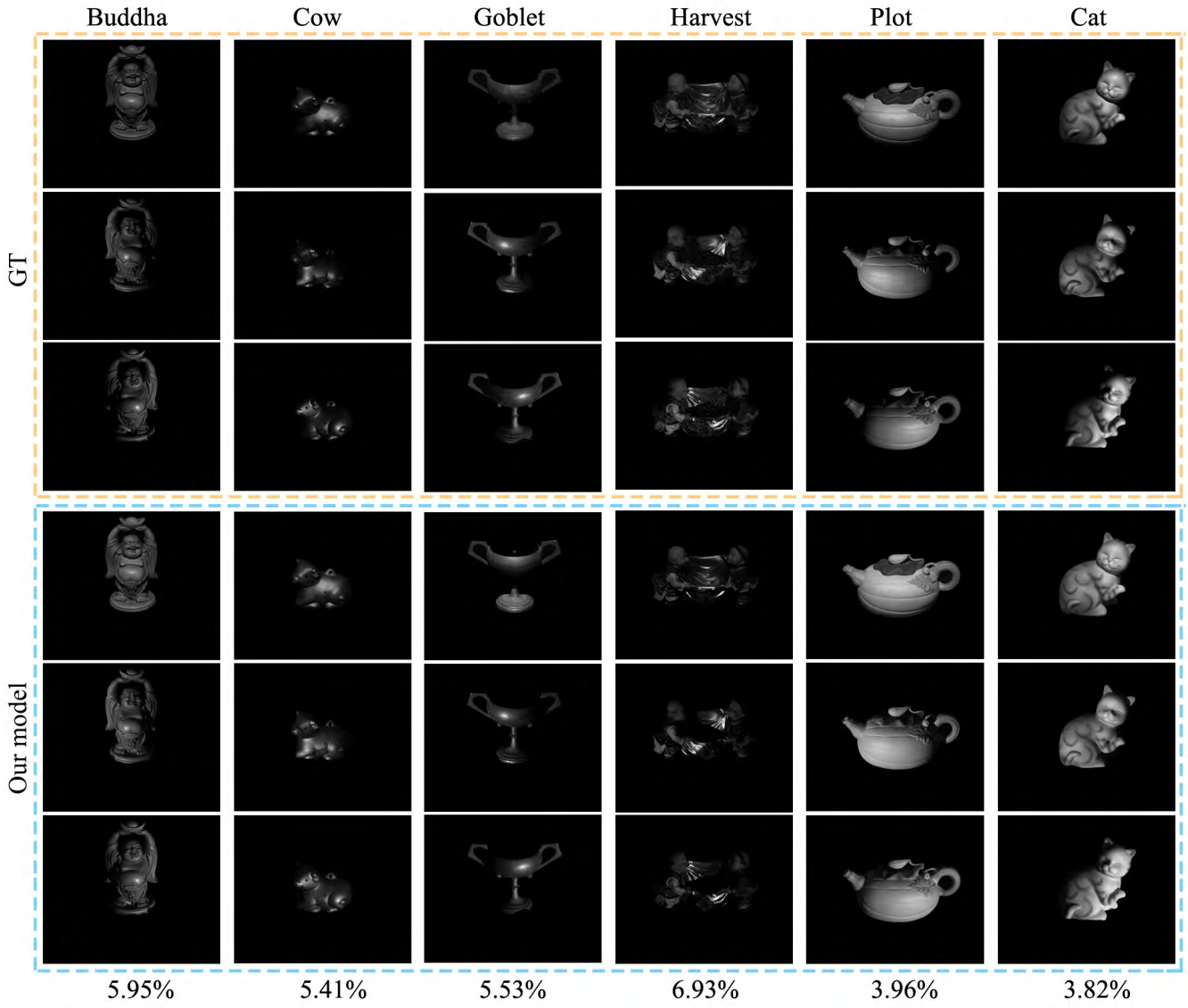

FIGURE 10. Evaluation of our method in the pseudo-colored lights dataset. The number represents MRE. In each column, the first three images are groundtruth, the three images below the groundtruth are the results of our method. On the top of each column, the names of objects are shown.

Figure 9 shows the groundtruth and estimated grayscale images. Here, we show 6 objects out of our test set. Fabric1 and Fabric2 are made of rough clothes which is almost lambertian, while Fabric3 and Fabric4 are made of smooth clothes which may have slight reflection. For Plastic1 and Plastic2, the surfaces are coated with a variety of paints. The MRE of the fabric objects are remains around 5%, while the MRE of plastic objects are larger than the others. It can be seen that the surface color of plastic objects are darker than fabric objects, especially the Plastic2 (the colored images are shown in Figure 11), while the dark samples in our dataset are insufficient. It means that the training pairs in this region are sparse, causing the local under-fitting in our model and larger errors in estimation.

Meanwhile, it is worth noting that there is no plastic object in our training set. This result shows that our model is robust and achieves high accuracy for objects consisting of various materials. Furthermore, we also find the performance is stable even when colorful surface and rich texture are exhibited in the test set. We believe this is because of the robustness of our model. Our per-pixel trained deep fully connected network does not learn the overall features, but demultiplex the image from the pixel point itself.

2) PSEUDO-COLORED LIGHTS DATASET

Secondly, we evaluate the performance of our DNN model using pseudo-colored lights dataset. Figure 10 shows the groundtruth and estimated grayscale images of the pseudo-colored lights dataset.

We show 6 objects out of 10 in the pseudo-colored lights dataset. The materials and the colors of the objects are various including wood, metal, ceramics and plastic (the colored images are shown in Figure 12). The MRE of objects are

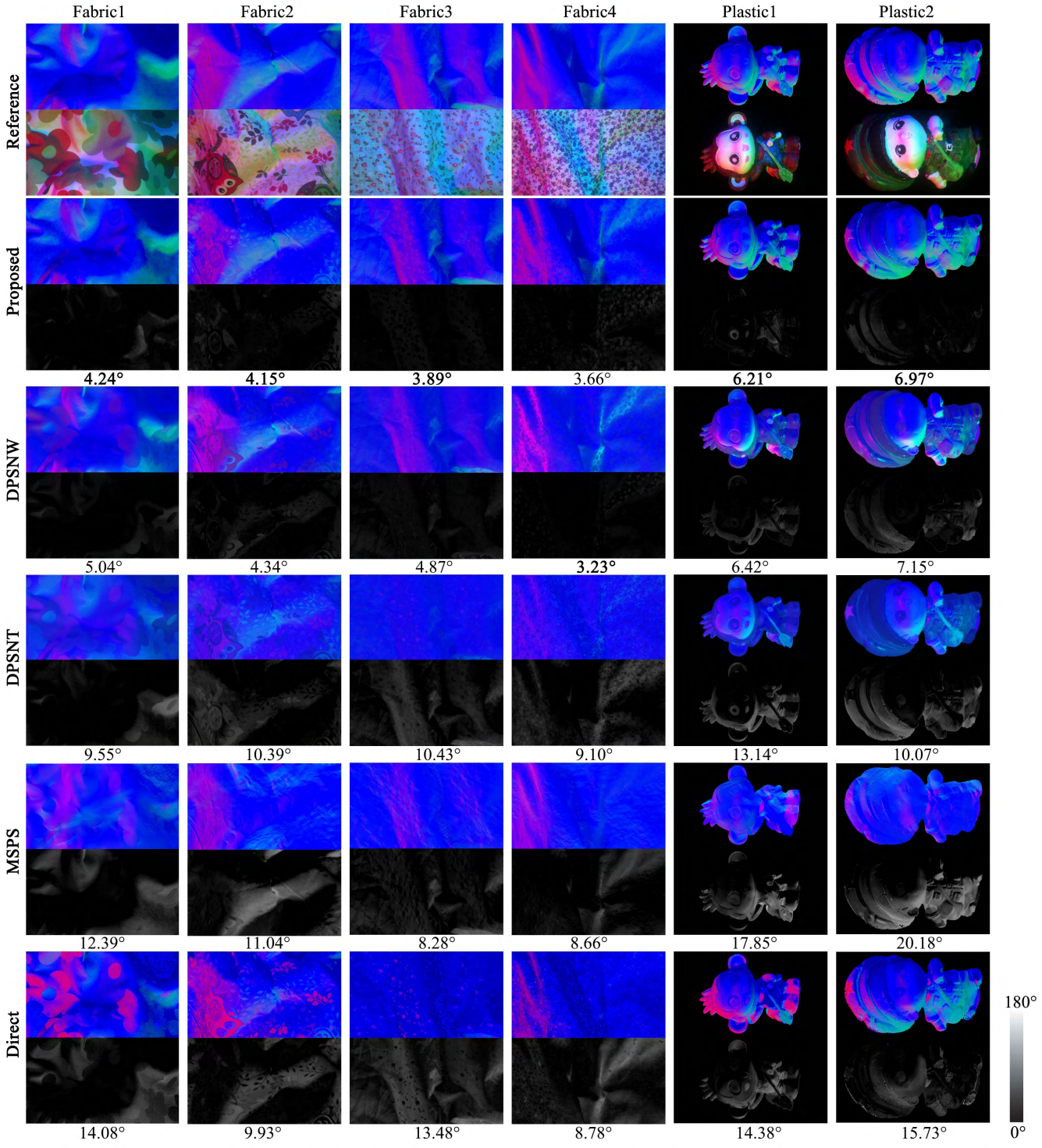

FIGURE 11. Estimation results for our dataset. In each row, normal maps are shown on top of error maps. Reference means classic photometric stereo using three white light illuminated grayscale images and below is the tri-colored lights images of the objects. The number represents Mean Angular Error (MAE) in degree. The position of the images is adjusted for easy viewing.

mainly concentrated in 3% to 6%. The performance may be due to the application of the input layer's concatenation operations in our DNN model. It shows the generalization and fitting ability of our model.

For an object like Harvest, which does not fit most of our training samples, the errors are still below 7%. In this case, the non-Lambertian object surface contains mirror reflection and directional reflection. We note that, though

non-Lambertain does not strictly follow the assumption shown in Eq.6, our method still manages to achieve satisfactory results. For objects like Buddha with complex structures, some pixel values may be affected by inter-reflection. Our model has inevitable deviations in these pixels, lacking the consideration of inter-reflection. We also observe from the experiment that the influence of inter-reflection is negligible in most objects. More comparative experiments will be discussed later.

D. STATE-OF-THE-ART COMPARISONS

Then we evaluate the performance of our method with multispectral photometric stereo (MSPS) reported in [18] using initial depths obtained by photometric stereo of the input image separation and DPSN [13]. For DPSN, we made some changes to make it suitable for our comparison experiments. We reduce the input dimension in DPSN from 96 to 3 representing RGB channels to ensure no additional input information. Furthermore, we cancel the shadow layer considering only three dimensions of the input in the remodeled DPSN. Firstly, in order to demonstrate our method, we input the pixels of three grayscale images $\boldsymbol{c_w}$ under white illumination in DPSN, whereas we input the pixels of tri-colored lights image $\mathbf{c}$ to our model. Secondly, in order to prove the value of our model and single frame reconstruction, we also input the same tri-colored lights image's pixel $\mathbf{c}$ to DPSN.

1) EXPERIMENTAL RESULT OF OUR DATASET

We first train the model using our dataset, and the results are shown in Figure 11. It shows the estimated normal maps and the corresponding error maps for our model. We show the normal obtained by multiplexing tri-colored lights based image separation (Direct).

As shown in Figure 11, the estimated normal maps and the corresponding error map for the proposed, DPSNW, DPSNT, MSPS, Direct are displayed. The objects used are the same as those shown in Figure 9. The values of Mean Angular Error (MAE) are shown below each method. MAE is a quantitative evaluation method in the surface normal analysis, it is often used in the comparison of the normal estimation algorithms.

We observe the normal maps from our proposed method are consistently accurate in all the objects. Compared to DPSNW using white illumination grayscale images, our proposed method achieves better accuracy in most objects only except Fabric4. The comparison with DPSNW can be considered as: The single-frame normal recovery framework proposed by us has outperformed the three images multi-frames DPSN method. This comparison shows the powerful demultiplexing ability of our DNN model.

DPSNT uses the same input as our model, and the results are worse than our method. DPSN achieved excellent results with 96 inputs [13], while its performance may be deteriorated with the reduced input dimensions. When the inputs are the same, the mapping from pixel values to pixel values will be better learned than that from pixel values to normals. We believe it may be difficult for DPSNT to distinguish whether the change of the pixel value is caused by the change of the normal or the change of the surface property. This is why we took the two-step framework instead of an end-to-end strategy like DPSNT.

Furthermore, we also compare our results with traditional multispectral photometric stereo (MSPS). Since the initial normal by tri-colored lights image separation (Direct) is not accurate, the error of MSPS becomes rather large. This experiment also illustrates that the traditional MSPS heavily relies on prior information to solve the aliasing problem, which is not robust. We also note that the performance of all the methods is deteriorated on Plastic1 and Plastic2. The reason is that the surface color is quite dark in these objects, which causes a larger deviation in the mapping because of the sparser samples in the model learning. Moreover, there is no plastic object in our training set. Even so, our proposed method still can yield the best estimation.

However, we also admit that the groundtruth used in the normal experiment on our dataset are ''reference'' which calculated by photometric stereo with three white lights. While the ''reference'' groundtruth may lead to deviations of MAE, the proposed method remains the best results compared with the MAE of other methods. In order to rigorously prove our method, we then performed the same experiment on the Pseudo-colored lights datasets with real groundtruth.

2) EXPERIMENTAL RESULT FOR PSEUDO-COLORED LIGHTS DATASET

We also compare our proposed method with DPSNW, DPSNT, MSPS, Direct on pseudo-colored lights datasets. The evaluation results are summarized in Figure 12. The objects are the same as those illustrated in Figure 10.

It is remarkable that many objects in DiLiGent [15] have non-Lambertian surfaces. Although our method maintains robust accuracy during mapping, the algorithm based on classic photometric stereo still causes large errors. Simultaneously, as an end-to-end learning method, DPSN has the advantage to learn the complex reflection model of non-Lambertian surfaces. For objects like Harvest, which have a strong non-Lambertian property, our method is obviously worse in this case. We note that in our experiments DPSNW is regarded as three images multi-frame normal estimation method. Compared with our single-frame method, their results are better than ours. When we compare our system with DPSNT which uses the same input as our model, our method shows better results in most objects except Harvest.

Furthermore, the ''Direct'' has very large errors caused by problems of spectrum-multiplexed and non-Lambertain surfaces simultaneously on pseudo-colored lights datasets. Therefore, the initial needed MSPS has even larger errors. The results illustrate the lack of robustness of traditional MSPS, which heavily relies on prior information to solve the spectrum-multiplexed problem.

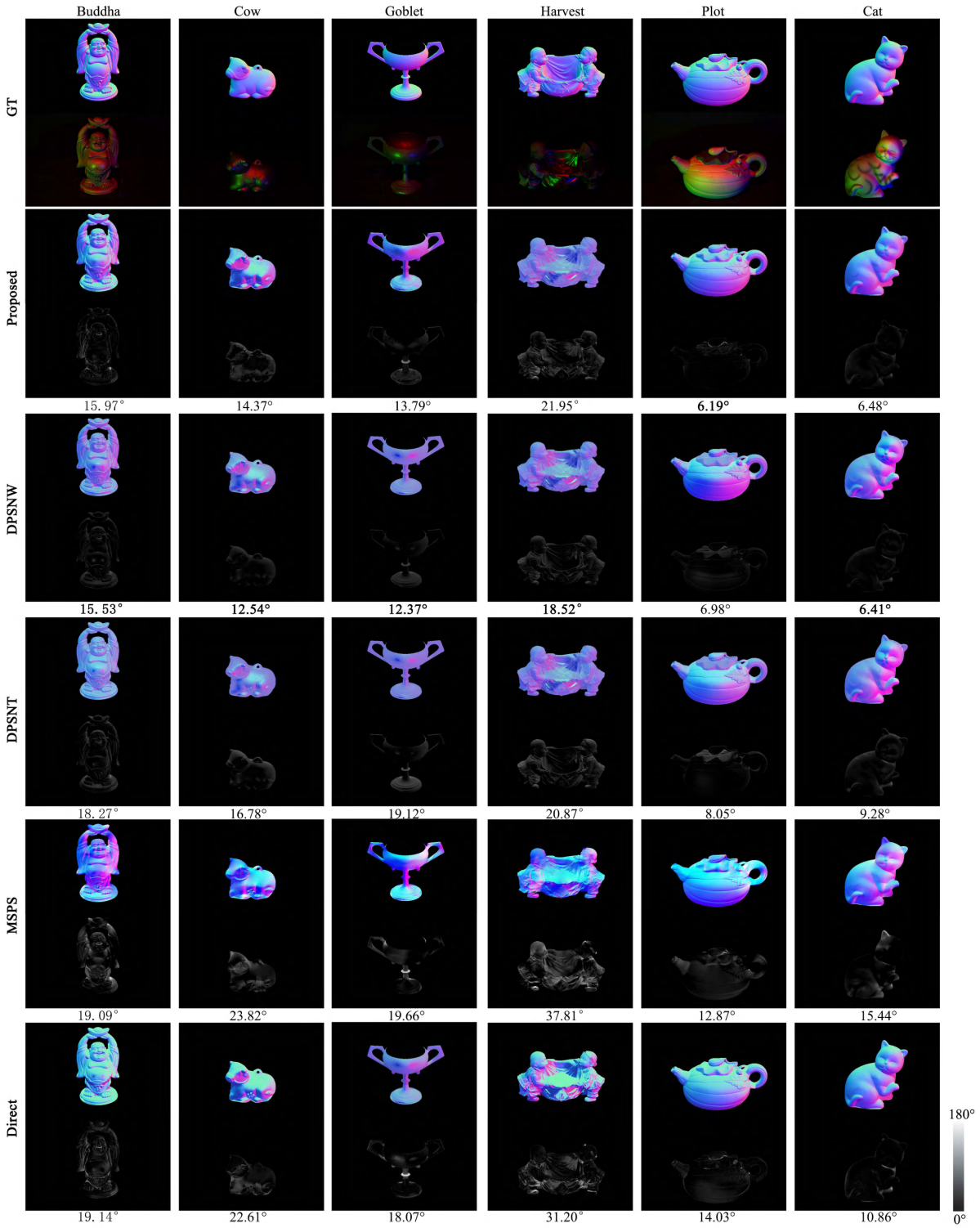

FIGURE 12. Estimation result for Pseudo-colored lights dataset. In each row, normal maps are shown on top of error maps. GT means groundtruth and below Figures are pseudo-colored lights images of objects. The number represents Mean Angular Error (MAE) in degree calculate with groundtruth. The position of images are adjusted for easy viewing.

VII. CONCLUSION

This paper presented a novel method to reconstruct surface shapes, handling non-rigid/moving objects and producing per-pixel dense results. Differing from previous solutions which require calibration or other prior information, we first formulated the problem in a learning framework, which directly seeks the per-pixel mapping of the aliased and spectrum-multiplexed pixel response to the anti-aliased and

demultiplexed counterpart. In this paper, we employed neural networks as the ''demultiplexer''. Experiments using our dataset and a public dataset have verified the effectiveness of our deep fully connected networks based mapping. Our proposed method can better handle the multiplexing in multispectral photometric stereo.

Our future work will try to establish an end-to-end model for estimating surface normal from a single multispectral photometric stereo image. Considering the spectrum-multiplexed and less consideration of inter-reflection, we will introduce both feature cues and photometric cues in our end-to-end model. Moreover, it will solve the limitation of calibrated and fixed light directions in this paper.

REFERENCES

[1] R. J. Woodham, ''Photometric method for determining surface orientation from multiple images,'' *Opt. Eng.*, vol. 19, no. 1, pp. 139–144, 1980.

[2] B. Wagner, P. Stüber, T. Wissel, R. Bruder, A. Schweikard, and F. Ernst, ''Accuracy analysis for triangulation and tracking based on time-multiplexed structured light,'' *Med. Phys.*, vol. 41, no. 8, 2014.

[3] D. Scharstein and R. Szeliski, ''A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,'' *Int. J. Comput. Vis.*, vol. 47, nos. 1–3, pp. 7–42, Apr. 2002.

[4] G. Xu, J. Chen, and X. Li, ''3-D reconstruction of binocular vision using distance objective generated from two pairs of skew projection lines,'' *IEEE Access*, vol. 5, pp. 27272–27280, 2017.

[5] C. Tomasi and T. Kanade, ''Shape and motion from image streams under orthography: A factorization method,'' *Int. J. Comput. Vis.*, vol. 9, no. 2, pp. 137–154, Nov. 1992.

[6] R. Szeliski and S. B. Kang, ''Recovering 3D shape and motion from image streams using nonlinear least squares,'' *J. Vis. Commun. Image Represent.*, vol. 5, no. 1, pp. 10–28, 1994.

[7] G. Wang, ''Robust structure and motion recovery based on augmented factorization,'' *IEEE Access*, vol. 5, pp. 18999–19011, 2017.

[8] L. L. Kontsevich, A. P. Petrov, and I. S. Vergelskaya, ''Reconstruction of shape from shading in color images,'' *J. Opt. Soc. Amer. A*, vol. 11, no. 3, pp. 1047–1052, 1994.

[9] A. Petrov, ''Light, color, and shape,'' in *Proc. Cognit. Processes Simulation*, 1987, pp. 350–358.

[10] C. Hernandez, G. Vogiatzis, G. J. Brostow, B. Stenger, and R. Cipolla, ''Non-rigid photometric stereo with colored lights,'' in *Proc. 11th IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2007, pp. 1–8.

[11] R. Anderson, B. Stenger, and R. Cipolla, ''Augmenting depth camera output using photometric stereo,'' in *Proc. MVA*, 2011, pp. 1–4.

[12] Y. Yoon, G. Choe, N. Kim, J.-Y. Lee, and I. S. Kweon, ''Fine-scale surface normal estimation using a single NIR image,'' in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 486–500.

[13] H. Santo, M. Samejima, Y. Sugano, B. Shi, and Y. Matsushita, ''Deep photometric stereo network,'' in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Oct. 2017, pp. 501–509.

[14] D. Eigen, C. Puhrsch, and R. Fergus, ''Depth map prediction from a single image using a multi-scale deep network,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2366–2374.

[15] B. Shi, Z. Wu, Z. Mo, D. Duan, S.-K. Yeung, and P. Tan, ''A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3707–3716.

[16] H. Kim, B. Wilburn, and M. Ben-Ezra, ''Photometric stereo for dynamic surface orientations,'' in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2010, pp. 59–72.

[17] R. Anderson, B. Stenger, and R. Cipolla, ''Color photometric stereo for multicolored surfaces,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 2182–2189.

[18] H. Jiao, Y. Luo, N. Wang, L. Qi, J. Dong, and H. Lei, ''Underwater multispectral photometric stereo reconstruction from a single RGBD image,'' in *Proc. APSIPA*, Dec. 2016, pp. 1–4.

[19] R. M. Nguyen, D. K. Prasad, and M. S. Brown, ''Training-based spectral reconstruction from a single RGB image,'' in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2014, pp. 186–201.

[20] Y. Jia *et al.*, ''From RGB to spectrum for natural scenes via manifold-based mapping,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4715–4723.

[21] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, ''Densely connected convolutional networks,'' in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jul. 2017, pp. 2261–2269.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, ''Gradient-based learning applied to document recognition,'' *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[23] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, ''Image quality assessment: From error visibility to structural similarity,'' *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[24] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, ''Slic superpixels compared to state-of-the-art superpixel methods,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[25] D. Eigen and R. Fergus, ''Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 2650–2658.

[26] X. Wang, D. Fouhey, and A. Gupta, ''Designing deep networks for surface normal estimation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 539–547.

YAKUN JU received the B.Sc. degree in industrial design engineering from Sichuan University, Chengdu, China, in 2016. He is currently pursuing the Ph.D. degree in computer application technology with the Department of Computer Science and Technology, Ocean University of China, Qingdao, China. His research interests include 3-D reconstruction, machine learning, and image processing.

LIN QI received the B.Sc. and M.Sc. degrees from the Ocean University of China, in 2005 and 2008, respectively, and the Ph.D. degree in computer science from Heriot-Watt University in 2012. He is currently an Associate Professor with the Department of Computer Science and Technology, Ocean University of China. His research interests include computer vision and visual perception.

HUIYU ZHOU received the B.E. degree in radio technology from the Huazhong University of Science and Technology, China, the M.Sc. degree in biomedical engineering from the University of Dundee, U.K., and the Ph.D. degree in computer vision from Heriot-Watt University, Edinburgh, U.K. He is currently a Reader with the Department of Informatics, University of Leicester, U.K. He has authored or co-authored over 150 peer-reviewed papers in the field. He was a recipient of the Computer Vision and Image Understanding 2012 Most Cited Paper Award and the International Conference on Pattern Recognition Applications and Methods 2016 Best Paper Award, and was nominated for ICPRAM 2017 Best Student Paper Award and Medical & Biological Engineering & Computing 2006 Nightingale Prize. Four of his papers recently published by Elsevier were ranked as the ScienceDirect Top 25 Articles. He serves as the Editor-in-Chief of Recent Advances in electrical & electronic engineering and an Associate Editor of the IEEE TRANSACTION ON HUMAN-MACHINE SYSTEMS, and is on the Editorial Boards of five refereed journals. He is one of the Technical Committee of Information Assurance & Intelligent Multimedia-Mobile Communication in IEEE SMC Society, Robotics Task Force and the Biometrics Task Force of the Intelligent Systems Applications Technical Committee, and the IEEE Computational Intelligence Society.

JUNYU DONG (M'09) received the B.Sc. and M.Sc. degrees from the Department of Applied Mathematics, Ocean University of China, Qingdao, China, in 1993 and 1999, respectively, and the Ph.D. degree in image processing from the Department of Computer Science, Heriot-Watt University, U.K., in 2003. He joined the Ocean University of China, in 2004, where he is currently a Professor and the Head of the Department of Computer Science and Technology. His research interests include machine learning, big data, computer vision, and underwater vision.

LIANG LU received the B.Sc. degree in electronic information science and technology and the M.Sc. degree in communication and information system from the Ocean University of China, Qingdao, China, in 2006 and 2009, respectively, where he is currently pursuing the Ph.D. degree in computer application technology. His research interests include image processing, computer vision, and machine learning.