Incorporating Lambertian Priors into Surface Normals Measurement

Yakun Ju, *Graduate Student Member, IEEE,* Muwei Jian, Shaoxiang Guo, Yingyu Wang, Huiyu Zhou, Junyu Dong, *Member, IEEE,*

***Abstract*—The goal of photometric stereo is to measure the precise surface normal of a 3D object from observations with various shading cues. However, non-Lambertian surfaces influence the measurement accuracy due to irregular shading cues. Despite deep neural networks have been employed to simulate the performance of non-Lambertian surfaces, the error in specularities, shadows, and crinkle regions is hard to be reduced. In order to address this challenge, we here propose a photometric stereo network that incorporates Lambertian priors to better measure the surface normal. In this paper, we use the initial normal under the Lambertian assumption as the prior information to refine the normal measurement, instead of solely applying the observed shading cues to deriving the surface normal. Our method utilizes the Lambertian information to reparameterize the network weights and the powerful fitting ability of deep neural networks to correct these errors caused by general reflectance properties. Our explorations include: the Lambertian priors (1) reduce the learning hypothesis space, making our method learn the mapping in the same surface normal space and improving the accuracy of learning, and (2) provides the differential features learning, improving the surfaces reconstruction of details. –Extensive experiments verify the effectiveness of the proposed Lambertian prior photometric stereo network in accurate surface normal measurement, on the challenging benchmark dataset.**

***Index Terms*—Photometric stereo, surface normal measurement, prior fusion, deep neural networks**

I. Introduction

MEASUREMENT of 3D geometry from 2D scenes is a key problem in machine vision and industrial applications [1]–[4]. Unlike multi-view stereo and binocular that use different viewpoints scenes to triangulate sparse 3D points, photometric stereo [5] measures the pixel-wise surface normal from a fixed scene under varying shading cues. The photometric methods prevail in measuring fine details of the surface and dense reconstruction, while it bases on the Lambertian assumption, hardly handling the general reflectance properties existing in real-world objects and deviating from realistic applications. To deal with the limitations, previous research adopted the bidirectional reflectance distribution functions (BRDFs) to model general reflectance [6]–[8] or treated the non-Lambertian regions as anomalies [9]–[11]. However, these traditional methods are accurate for a limited class of surface reflectance and suffer from unstable optimization.

Recently, deep learning-based methods have been introduced to photometric stereo [12], [13]. These methods directly learn the surface normal of objects from the observed images, where the capability of handling diverse BRDFs has been proved to be notable. However, strong non-Lambertian and varying reflectance regions still lead to significant errors, in both the traditional algorithms and deep learning-based methods. We reckon that the failure in these regions is due to the fact that (1) the regions and surfaces rarely appear in the training dataset, (2) the overexposed values in specularity and the dark areas in shadows are difficult to produce useful features, (3) the surface with spatially-varying reflectance causes measurement errors on the normal map. These challenging problems remain to be solved. We show our state-of-the-art performance over these conditions in Fig. 1.

To address these challenges, we propose a Lambertian model guided photometric stereo network to better handle non-Lambertian surfaces. Instead of directly embedding the observed shading cues on the surface normal, we propose a strategy that utilizes the observations to modify the Lambertian priors, in other words, our method focuses on the use of the input images as the patterns to correct the measurement of surface normal from the Lambertian priors. Compared with previous methods mainly focusing on network architectures, our method has the following advantages: First, our method utilizes a mapping in the same surface normal space $\{\mathcal{Y} \rightarrow \mathcal{Y}\}$, where the learning space is reduced, while the previous methods take a mapping from the RGB image space to the surface normal space $\{\mathcal{X} \rightarrow \mathcal{Y}\}$. Second, the Lambertian priors and ground-truth surface normal are theoretically similar in terms of diffuse reflection. In this condition, our method is equivalent to enlarge the proportion of non-Lambertian errors in total errors, where the network always tends to learn the parameters to make the loss drop. In training, the network therefore will be more inclined to optimize this part of errors. *i.e.*, our method learns the differential features, amplifying the non-Lambertian errors, where the details of measurement is improved. In short, our method reduces the learnable hypothesis space and improves the estimation results. The experimental results have demonstrated the effectiveness of the Lambertian priors: Our architecture notably improves the accuracy of surface normal measurement, outperforming several state-of-the-art calibrated

Y. Ju, S. Guo, Y. Wang, and J. Dong are with the Department of Computer Science and Technology, Ocean University of China, Qingdao, China (e-mails: {juyakun, guoshaoxiang, wangyingyu}@stu.ouc.edu.cn, dongjunyu@ouc.edu.cn.)

M. Jian is with the School of Information Science and Engineering, Linyi University, Linyi, China, and the School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan, China (e-mail: jianmuweihk@163.com).

H. Zhou is with the Department of Informatics, University of Leicester, Leicester, UK (e-mail: hz143@leicester.ac.uk).

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

The work was supported by the National Key R & D Program of China under Grant (2018AAA0100602), the National Key Scientific Instrument and Equipment Development Projects of China (41927805), the National Natural Science Foundation of China (61501417, 61976123), Royal Society - K. C. Wong International Fellowship (NIF\R1\180909) and the Taishan Young Scholars Program of Shandong Province.

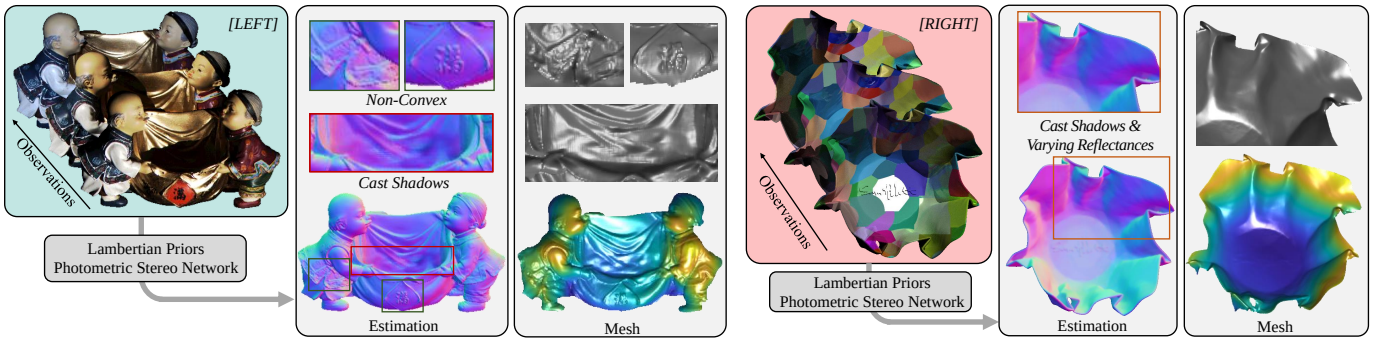

Fig. 1. We present a novel Lambertian prior guided photometric stereo network to measure the surface normal of the target. Our method improves the accuracy of strong non-Lambertian surfaces, non-convex structure, and varying reflectance surface. *[LEFT]*: the results on "Harvest" of the DiLiGenT benchmark dataset [14] (using 96 observations), where the strong cast shadows and non-convex structures exist in the observations. *[RIGHT]*: the results on "Paperbowl (Specular)" of the CyclesPSTest dataset [15] (using 17 observations). The cast shadows and varying reflectance are displayed in the observation. The details of 3D mesh are also shown in each example, which are derived from the measured surface normal by the standard integration method [16].

photometric stereo approaches on the widely used DiLiGenT benchmark dataset [14]. Furthermore, the ablation experiments show the fast convergence and more accurate surface normal measurement by the proposed photometric stereo network.

The main two contributions of this work are summarized as follows:

- We propose a Lambertian prior guided photometric stereo network to better handle non-Lambertian surfaces. Our method reduces the learnable hypothesis space and improves the surface normals measurement.
- To the best of our knowledge, this is the first work that involves embedding the priors into the learning-based photometric stereo network. This strategy can be also used to refine wider non-ideal photometric stereo measurements, by using inaccurate priors.

II. Related Work

A measurement pixel of a real-world object I_j can be modeled by general BRDFs ρ, which is associated with the surface normal $\boldsymbol{n} \in \mathbb{R}^3$, illumination direction $\boldsymbol{l_j} \in \mathbb{R}^3$, and view direction $\boldsymbol{v} \in \mathbb{R}^3$. With the illumination intensity e, the imaging model can be expressed as:

$$I_j = e\rho\left(\boldsymbol{n}, \boldsymbol{l_j}, \boldsymbol{v}\right) \max\left(\bar{\boldsymbol{n}}^\top \boldsymbol{l_j}, 0\right) + \epsilon_j, \tag{1}$$

where $\max\left(\boldsymbol{n}^\top \boldsymbol{l}_j, 0\right)$ revealed the attached shadows, and ϵ_j is an error term where the impacts are hardly represented by the BRDF model, such as cast shadows, imaging noise, and inter-reflections [17]. As an inverse problem, the goal of photometric stereo is to measure the surface orientation from a combination of reflectance and illuminations in multiple images. The literature of photometric stereo is vast, and we here briefly review the mainstream calibrated photometric stereo technologies, including traditional algorithms and deep learning based methods. Comprehensive photometric stereo surveys on hand-craft reflectance models and robust methods can be found in [18], [19].

A. Lambertian Photometric Stereo

In 1980, Woodham firstly proposed the Lambertian photometric stereo algorithm [5]. Under the Lambertian assumption, the error term ϵ_j is ignored and the BRDF has the diffuse property, where the observed intensity is proportional to the cosine of the angle between the illumination direction and the surface normal but irrelevant to the view direction. Therefore, the imaging model can be simplified and easily solved by the least-square approach. Although the Lambertian method failed to estimate most of the real-world non-Lambertian objects, the basic theory is significant that reveals the image formation model can be cast into a linear system of equations and solved, establishing the relationship between two-dimensional images and the object geometry.

B. Non-Lambertian Photometric Stereo

To extend photometric stereo to work with non-Lambertian surfaces in practice, researchers investigated different strategies. Commonly, according to the taxonomy of [14], non-Lambertian photometric stereo technologies can be divided into four categories: robust methods, analytic and empirical reflectance model methods, example-based methods, and deep learning methods.

1) Robust methods: Robust methods treat most regions on the surface as a simple diffuse reflectance model (Lambertian) while treating non-Lambertian phenomena (such as specularity and cast shadows) as outliers. These methods assume specularity and shadows are local and sparse, which can be detected and discarded. In the early work, the surface normal is estimated by selecting the three images with the lowest specularity and the closest Lambertian appearance from multiple images [20]. Afterwards, Wu *et al.* [9] proposed a robust principal component analysis (RPCA) method to decompose images into the minimized-rank Lambertian composition and the non-Lambertian sparse counterpart. Similarly, Ikehata *et al.* [10] employed an improved rank = 3 decomposition instead of rank-minimization, achieving better computational stability. Several other robust techniques were also applied to solve the outliers, such as maximum-likelihood estimation [21], shadow cuts [22], and maximum feasible subsystem [23]. Although robust methods are effective, these approaches generally cannot handle the surface with broad and soft specularity, where non-Lambertian regions are hard to be detected as the outliers. In

addition, these methods usually need a huge number of the observed images to achieve stable computations.

2) Analytic and Empirical Reflectance Model Methods: To handle the non-Lambertian, using analytic or empirical reflectance model to approximate the non-Lambertian BRDFs is a fairly straightforward idea. Along this direction, many models were proposed to fit the nonlinear analytic BRDF, such as the specular spike model [24], the Blinn-Phong model [25], the Torrance-Sparrow model [26], the Ward model and its variations [27], [28], and the microfacet BRDF model [29]. In addition, empirical reflectance models consider the general properties of a BRDF, such as isotropy and monotonicity, to deal with multiple types of surface materials. Some basic derivations for isotropy BRDFs were proposed in [30], [31]. Based on empirical models, some researchers applied isotropic constraints for the measurement of surface orientation [32], [33]. Shi *et al.* [34] and Ikehata *et al.* [7] further approximated the isotropic BRDFs by bivariate functions to deal with the instability of the estimation. However, these hand-crafted analytic and empirical reflectance models are generally useful for limited categories of reflectance as the reflectance properties are significantly changing from materials to materials. Moreover, most of these methods are pixel manners which ignore inter-reflection and cast shadows.

3) Example Based Methods: In addition to the above two strategies, amounts of photometric stereo algorithms can be treated as example based methods. Under the same illumination environment, the calibration object with the known surface normal (usually a sphere) transformed the non-Lambertian photometric stereo to a pixel matching problem. Early work required exactly the same material between the target and the calibrated objects. Hertzmann and Seitz [35] relaxed this restriction by assuming that a small number of basis materials compose the general materials. Hui and Sankaranarayanan [36] used a BRDF dictionary to render virtual spheres instead of putting the real calibrated objects. However, the drawback of the same illumination configuration limits its practical use.

4) Deep Learning Methods: Deep learning techniques have been introduced to solve the problem of non-Lambertian photometric stereo, which have achieved inspiring performance. Santo *et al.* first presented DPSN [12] to address the non-Lambertian photometric stereo. DPSN regresses the per-pixel normal from the fixed number of the observed images, where the training and the testing have to use the same predefined illumination directions. Therefore, a new model has to be retrained if the input number or illumination directions are changed. To relax this constraint and take advantage of the information embedded in the neighborhood, subsequent methods were improved by applying convolutional neural networks (CNNs) and explored flexible input strategies [13], [15], [37]. Ikehata [15] introduced the observation map, which rearranges observation intensities according to light directions, to overcome the fixed inputs problem. The observation map strategy was also adopted in [37], [38], which is effective for inputs with order-agnostic illuminations. Others [13], [39], [40] applied the max-pooling method to aggregate features from a number of inputs. In addition, some works proposed to better handle specified problems. Taniai and Maehara [41] proposed an unsupervised method to better handle the condition of lacking ground-truth surface normals, where the surface normals are estimated by minimizing the reconstruction loss. Ju *et al.* [42] proposed an adaptive attention-weighted loss to improve the performance of complex-structured areas, where the detail-preserving gradient loss can produce clear reconstructions. More recently, Yao *et al.* [43] attempt to introduce GNN for learning-based photometric stereo, named GPS-Net. Ju *et al.* [44] proposed a dual-regression method to recover both surface normal and rendered observations, which provides an additional supervision.

However, these deep learning methods still need to be improved when dealing with strong non-Lambertian surfaces and high-frequency structures. Unlike previous deep learning-based photometric stereo methods, we introduce the Lambertian priors to guide the learning and correct the errors caused by non-ideal solutions, which outperform state-of-the-art methods on challenging benchmark datasets.

III. Proposed Method

In this section, we will present our proposed Lambertian priors photometric stereo network. Our goal is to measure surface normal of non-Lambertian surfaces with complex reflectance such as inter-reflections and cast shadows. Our strategy is to boost the measurement based on the physical Lambertian prior. Before introducing the network details, we will first present the learning objective.

A. What Is Learned in Our Method

Assuming that a surface point on the surface with a unit normal $\boldsymbol{n} \in \mathbb{R}^3$ is illuminated by j^{th} light source with direction $\boldsymbol{l_j} \in \mathbb{R}^3$ in intensity observation $\boldsymbol{i_j} \in \mathbb{R}^3$. We artificially split the observation $\boldsymbol{i_j}$ into two parts: the diffuse $\boldsymbol{i_j^d}$ and the other $\boldsymbol{i_j^o}$, where $\boldsymbol{i_j}$ can be shown by $\boldsymbol{i_j^d}$ + $\boldsymbol{i_j^o}$. The $\boldsymbol{i_j^d}$ represents the ideal Lambertian reflectance, while the $\boldsymbol{i_j^o}$ stands for specularity, cast shadows, inter-reflections, and global illuminations.

If we can exactly extract the diffuse $\boldsymbol{i_j^d}$ from observation $\boldsymbol{i_j}$, then the surface normal $\boldsymbol{n}$ can be expressed, according to ideal Lambertian photometric stereo [5], as follows:

$$\boldsymbol{n} = \frac{\boldsymbol{L}^{-1}\boldsymbol{I^d}}{|\boldsymbol{L}^{-1}\boldsymbol{I^d}|} , \tag{2}$$

where $\boldsymbol{I^d} = [\boldsymbol{i_1^d}, \boldsymbol{i_2^d}, \ldots, \boldsymbol{i_j^d}]'$ is the column vector of the diffuse part in $\{1, 2, \ldots, j\}$ and the matrix $\boldsymbol{L} = [\boldsymbol{l_1}, \boldsymbol{l_2}, \ldots, \boldsymbol{l_j}]'$ is composed of illumination direction in $\{1, 2, \ldots, j\}$.

However, few methods can exactly express the diffuse part from the observation, including outliers methods and illumination models. Rather than using sophisticated models or deep learning to further reduce fitting errors, we, hereafter, rethink the Lambertian model in the non-Lambertian condition. Here, we attend the prior surface normal $\boldsymbol{n'}$, which can also be calculated via the ideal physical model [5]:

$$\boldsymbol{n'} = \frac{\boldsymbol{L}^{-1}\boldsymbol{I}}{|\boldsymbol{L}^{-1}\boldsymbol{I}|} , \tag{3}$$

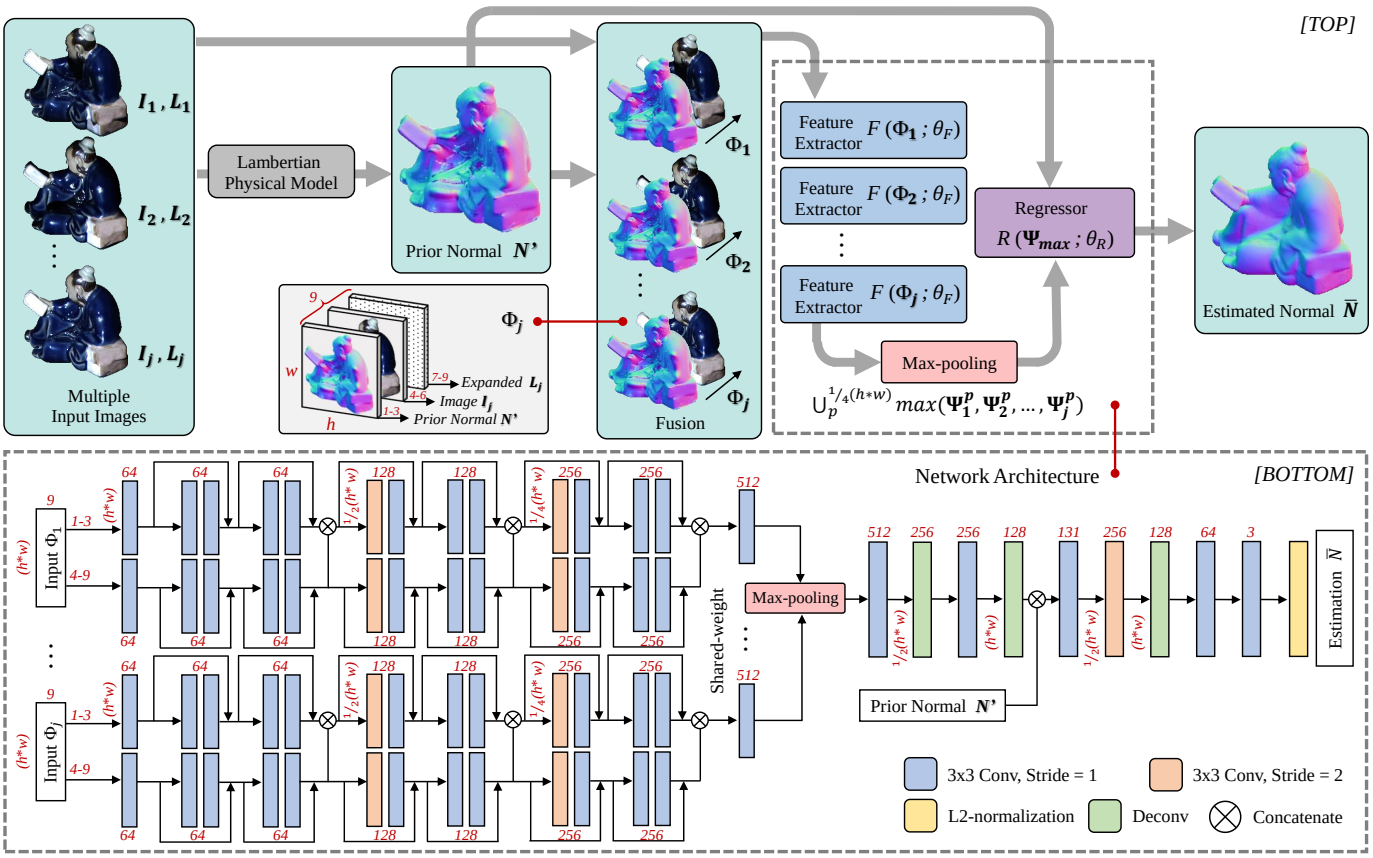

Fig. 2. *[TOP]* Overview. The prior normal $\boldsymbol{N}'$ is calculated by Lambertian physical model using input images $\{\boldsymbol{I_1}, \boldsymbol{I_2}, \ldots, \boldsymbol{I_j}\}$ with illumination directions $\{\boldsymbol{L_1}, \boldsymbol{L_2}, \ldots, \boldsymbol{L_j}\}$. Then, we fuse $\boldsymbol{N}'$ with $\{\boldsymbol{I_j}, \boldsymbol{L_j}\}$ in $\boldsymbol{\Phi_j}$, where the $\boldsymbol{L_j}$ is expanded to have the same spatial resolution as images. The detail of a $\boldsymbol{\Phi_j}$ is shown in the gray box, where the first three dimensions are normal and the last six dimensions are image and the expanded illumination direction. Our method takes the fusion $\{\boldsymbol{\Phi_1}, \boldsymbol{\Phi_2}, \ldots, \boldsymbol{\Phi_j}\}$ including prior normals as inputs and estimates the surface normal $\bar{\boldsymbol{N}}$. *[BOTTOM]* Network architecture. For each $\boldsymbol{\Phi}$, we split it into two parts, the prior normal and the image information. The channel dimension and spatial resolution of layers are highlighted in red.

where $\boldsymbol{I} = [\boldsymbol{i_1^d}+\boldsymbol{i_1^o}, \boldsymbol{i_2^d}+\boldsymbol{i_2^o}, \ldots, \boldsymbol{i_j^d}+\boldsymbol{i_j^o}]'$. It can be seen that the error of $\boldsymbol{n'}$ is caused by the non-Lambertian and other global effects $\boldsymbol{I^o} = [\boldsymbol{i_1^o}, \boldsymbol{i_2^o}, \ldots, \boldsymbol{i_j^o}]'$. Naturally, we wish to learn the nonlinear mapping by the pattern of observation $\boldsymbol{I}$ including $\boldsymbol{I^d}$ and $\boldsymbol{I^o}$ as:

$$Mapping\ (\frac{\boldsymbol{L}^{-1}\boldsymbol{I}}{|\boldsymbol{L}^{-1}\boldsymbol{I}|} \rightarrow \frac{\boldsymbol{L}^{-1}\boldsymbol{I^d}}{|\boldsymbol{L}^{-1}\boldsymbol{I^d}|}\ , with\ \boldsymbol{I^d}+\boldsymbol{I^o}) \quad (4)$$

Unlike previous deep learning-based methods focusing on using different setups to infer the surface normal from solely input images, as $Mapping\ (\boldsymbol{I^d}+\boldsymbol{I^o} \rightarrow \boldsymbol{n})$, our method focuses on the use of the input images $\boldsymbol{I^d}+\boldsymbol{I^o}$ as the patterns to correct the surface normal from Lambertian priors. For deep learning-based photometric stereo, we argue that the above mapping is a better choice for estimating the surface normal. Our analysis is that (1) mapping the errors in estimating surface normals $\{\mathcal{Y} \rightarrow \mathcal{Y}\}$ instead of mapping the surface normals from RGB images $\{\mathcal{X} \rightarrow \mathcal{Y}\}$ reduce the space of solving possible learning functions and improve the estimation results ($\mathcal{Y}$ represents the surface normal space, where $\mathcal{X}$ represent the RGB image space). (2) the prior normals $\boldsymbol{n'}$ are theoretically accurate in terms of diffuse reflection.In this condition, our method is equivalent to enlarge the proportion of non-Lambertian errors in total errors, where the network always tends to learn the parameters to make the loss drop. In training, the network therefore will be more inclined to optimize this part of errors. *i.e.*, our method learns the differential features, amplifying the non-Lambertian errors.

B. Network Architecture

In this section, a novel Lambertian guided photometric stereo network is designed according to the inference reported in Section III-A. The network enhances the measurement on non-Lambertian surfaces, crinkle regions, and varying reflectance edges. The overview and detailed architecture of our Lambertian priors based photometric stereo network is illustrated on the *[TOP]* and *[BOTTOM]* of Fig. 2, respectively.

We first fuse the Lambertian priors with input images. Given j^{th} images with known illumination directions, we take the Lambertian assumption [5] to calculate the prior normals $\boldsymbol{N}'$ by Eq. 3. We then concatenate the prior normal with each observation and illumination as $\boldsymbol{\Phi_j}$. For the illumination of the corresponding image, we expand the 3-vector illumination and expand it to the same spatial resolution as the image, $\boldsymbol{L_j} \in \mathbb{R}^{3\times h\times w}$, and concatenate the illumination $\boldsymbol{L_j}$ with the image. This operation makes the illumination directions fully fuse with the corresponding observation in a pixel-wise manner. Hence, each input $\boldsymbol{\Phi_j}$ has the dimensions of $\mathbb{R}^{9\times h\times w}$, where the concatenation order is the prior normal, the image, and the illumination direction. In fact, the above mapping model

(Eq. 4) is constructed in the per-pixel manner. However, the proposed network predicts the surface normal from the image patch. Inspired by previous works [13], [45], we argue that the embedded features from a neighboring image patch enhances the flexibility of the proposed network to various reflectance.

We separately feed these $\{\mathbf{\Phi_1}, \mathbf{\Phi_2}, \ldots, \mathbf{\Phi_j}\}$ to the network, as shown on the *[BOTTOM]* of Fig. 2. The red numbers represent the dimensions of the feature channels. We apply the Leaky-ReLU as the activation function of each layer. The network includes three stages: feature extraction, max-pooling fusion, and regression.

The first stage of our network can be seen as the j^{th} multi-branch shared-weight feature extraction, as:

$$\mathbf{\Psi_j} = F(\mathbf{\Phi_j}; \theta_F) \ , \tag{5}$$

where $\mathbf{\Psi_j} \in \mathbb{R}^{512\times\frac{1}{4}h\times\frac{1}{4}w}$ is the feature from the feature extractor F with learnable parameters θ_F. The feature extraction stage contains two residual networks [46], handling the surface normal and the image with its associated illumination direction, respectively. It can be seen that each residual network contains 6 residual blocks with two down-sampling convolutional layers. Residual blocks can effectively avoid gradient vanishing [46] in a deep network. Also, we argue that the shortcut fuses previous blocks, which is a combination of features at different levels and scales. In addition, the shortcut structure is equivalent to adding all the information of the previous layer image in each block, which retains more original features. We also compare the results of different network architectures in ablation study (Section IV-B3). The two down-sampling, from $h \times w$ to $\frac{1}{4}h \times \frac{1}{4}w$ are executed by stride = 2 convolutional layer on the third and fifth residual blocks. By concatenating the features of the image residual network to those of the prior normal residual network in different scales, the network increase the receptive field.

A convolutional layer is applicable for multi-feature fusion only when the number of inputs is fixed. Unfortunately, this is not practical for photometric stereo where the number of inputs often change. Therefore, in the second stage, we apply a max-pooling operation [13], [42] for multi-feature fusion, and getting fixed feature that can be backpropagated. Max-pooling extracts the most salient information from all the features. In fact, it is the regions with high intensities or specularities that provide strong clues for surface normal estiamtion. Also, the max-pooling operation can exclude the shadows from multi-illumination directions. In contrast, the average-pooling will smooth out useful features and may be impacted by non-activated features such as shadows. Here, we choose the superscript p to denote the index of the feature tensor as follows:

$$\mathbf{\Psi_{max}} = \bigcup_{p}^{\frac{1}{4}h\times\frac{1}{4}w} max(\mathbf{\Psi_1^p}, \mathbf{\Psi_2^p}, \ldots, \mathbf{\Psi_j^p}) \tag{6}$$

Finally, the normal regression stage R takes $\mathbf{\Psi_{max}}$ as input and regresses the estimation surface normal $\bar{\boldsymbol{N}}$ as:

$$\bar{\boldsymbol{N}} = R(\mathbf{\Psi_{max}}; \theta_R) \ , \tag{7}$$

where θ_R is the learnable parameter of the regressor R. The regressor consists of six convolutional layers, three deconvolutional layers, and an L2-normalization layer. The feature map is up-sampled (by the deconvolutional layer) three times and down-sampled (by the stride = 2 convolutional layer) once to fully utilize the embedded information, resulting in up-sampling of twice. We argue that this design can expand the receptive field and keep spatial information with a small GPU memory. We concatenate the prior normal $\boldsymbol{N}'$ to the feature map after the second deconvolutional layer. We reckon that the fusion in the regression stage will enhance the high-frequency details of the estimated surface normal. The detailed discussion can be found in the ablation experiments (Section IV-B).

The learning of our network is supervised by the error between the measured and the ground-truth surface normals. We optimize the networks parameters θ_F, θ_R by minimizing the cosine similarity loss function as:

$$\mathcal{L}_{\text{normal}} = \frac{1}{hw}\sum_{p}^{hw}\left(1 - \bar{\boldsymbol{N}}^p \cdot \boldsymbol{N}^p\right) \ , \tag{8}$$

where $\bar{\boldsymbol{N}}^p$ and $\boldsymbol{N}^p$ denote the measured normal and the ground-truth, respectively, at pixel p. If the estimated normal $\bar{\boldsymbol{N}}^p$ at pixel p has a similar orientation as the ground-truth $\boldsymbol{N}^p$, then the $\bar{\boldsymbol{N}}^p \cdot \boldsymbol{N}^p$ will be close to 1 and the cosine similarity loss will approach 0.

Our network was implemented using PyTorch and the Adam optimizer is used with the default settings (β_1= 0.9 and β_2=0.999) on a RTX 2080 GPU. The initial learning rate is set to 0.002, and divided by 2 every 5 epochs. We train the model using a batch size of 24 for 40 epochs. The number of observations for training and prior normals is 32. In addition, we set the spatial resolution h, w = 32 in training.

IV. Experiments

In this section, we present datasets, experimental results, and analysis. To verify the quantitative performance of our method, we use some widely used metrics to measure accuracy. We adopt the mean angular error (MAE) in degree to evaluate the performance of estimated surface normal, as follows:

$$MAE = \frac{1}{HW}\sum_{p}^{H\times W}\left(arccos(\bar{\boldsymbol{N}}^p \cdot \boldsymbol{N}^p)\right) \ , \tag{9}$$

where $H \times W$ represents the spatial resolution of the tested surface normal. We also apply $< err_{15^\circ}$ and $< err_{30^\circ}$ to measuring the percentage (%) that pixels with the angular error less than 15° and 30°, respectively.

A. Dataset

1) Training and validation datasets: As a supervised learning method, we adopt two publicly available synthetic datasets from [13], called blobby shape and sculpture shape datasets [47], which are rendered with the MERL dataset [48] by the physically-based raytracer Mitsuba [49]. Blobby and Sculpture datasets provide surfaces with complex structures and rich surface orientations, and the MERL dataset contains 100 different BRDFs of real-world materials. The combinations

provide comprehensive data distribution. The training set contains 85212 samples. For each sample, 64 observation images are rendered by random illumination directions in an upper-hemisphere, with a 128 $\times$ 128 spatial resolution. We randomly crop 32 $\times$ 32 (default training spatial resolution) images patches in each sample for data augmentation.

2) Testing datasets: To evaluate our method, we apply several commonly used datasets, including both synthetic and real datasets. For the synthetic dataset, we first employ the CyclesPSTest dataset [15]. CyclesPSTest is a synthetic dataset of three objects, "Sphere", "Turtle", and "Paperbowl". "Turtle" and "Paperbowlare" are objects with the non-convex surface where specularity and cast shadow extensive appear. We also employ the synthetic object "Dragon". The object "Dragon" was rendered with 100 different BRDFs from the MERL data set [48] under 100 random illumination directions in an upper-hemisphere, for testing the results of our method on different surface reflectances.

For the real dataset, we employ the public DiLiGenT benchmark dataset [14], which contains 10 objects of various shapes with complex materials. Each object provides images with a resolution of 612 $\times$ 512 from 96 different known illumination directions. The DiLiGenT benchmark dataset is challenging for its strong non-Lambertian surfaces and non-convex structures. Besides, we also employ the Light Stage Data Gallery [50], which contains six objects without ground-truth. Each object has 253 images under different illumination directions. Therefore we qualitatively evaluate our method on the Light Stage Data Gallery.

B. Ablation Experiments and Network Analysis

We took quantitative ablation experiments on the validation set of 852 samples. We first evaluated the effectiveness of our Lambertian priors based photometric stereo network (Experiments with IDs 0 & 1). We compared our default network with only using input images, where the residual network for the Lambertian priors branch (channel 1-3 of each input $\mathbf{\Phi}_j$) and concatenation are removed. We then investigated the influence of different prior inputs on our network (Experiments with IDs 0 & 2). For comparisons, we selected rank minimization [9] (a robust photometric stereo method) as the prior input. Furthermore, we evaluated the effectiveness of the selected residual architecture (experiments with IDs 0 & 3). We compared the residual blocks' settings [46] in the feature extraction stage with the plain settings (without shortcut connections). Finally, we evaluated the effectiveness of concatenating the prior normal $\boldsymbol{N}'$ in the regression stage (experiments with IDs 0, 4, & 5). We compared the concatenation with the prior after the second deconvloutional layer (default) with the concatenation with the prior after the third deconvloutional layer, and without the concatenation with the prior. For all the experiments in the ablation study, we train the networks with 32 input images, and reported the average results of the validation set of 852 samples. The results were summarized in Table I. To further evaluate the performance and the generalization of our method, we also report the ablation study on the synthetic object "Dragon" with 100 different materials, as shown in Fig. 3, corresponding to IDs 0, 1, and 3 in Table I.

TABLE I
RESULTS OF SYSTEM ANALYSIS ON THE VALIDATION SET WITH 32 INPUT IMAGES, WHERE THE NUMBERS REPRESENT THE AVERAGE VALUES OF ALL THE SAMPLES OF THE VALIDATION SET. THE LOWER MAE, THE BETTER. FOR $< err_{15^\circ}$ AND $< err_{30^\circ}$, THE HIGHER, THE BETTER.

ID	Variants	MAE	$< err_{15^\circ}$	$< err_{30^\circ}$
0	With prior normal (Ours)	12.30	**84.39%**	94.86%
1	Without prior normal	12.98	82.05%	94.71%
2	With rank minimization [9]	**12.27**	84.26%	**94.91%**
3	Plain network	12.45	83.94%	94.83%
4	Concatenate 3rd deconv	12.32	84.15%	94.65%
5	Without concatenate	12.54	82.85%	94.61%

1) Effectiveness of Lambertian priors based photometric stereo network: Experiments with IDs 0 and 1 demonstrated the performance of adding Lambertian priors to the photometric stereo networks. It can be seen that our method of using prior normal consistently performs better than the traditional mapping strategy over all the metrics, for example, 12.30° for MAE, 84.39% and 94.86% of the pixels have the angular errors of less than 15° and 30°. This is due to the fact that our method learns the mapping in the same normal space while the previous methods learn the mapping over different spaces: from RGB images to normal. Therefore our method can converge faster and achieve accurate estimations. Also, as shown in Fig. 3, It can be seen that on most materials, our method significantly outperformed the network without a normal prior and the baseline Lambertian method [5]. Note that the proposed method performed particularly well on the surfaces that have larger errors in the baseline method, which suggested our method can improve the prediction over strong non-Lambertian reflectance.

We further evaluated the robustness of our method with spatially varying BRDFs on the surface. Fig. 4 futher quantitatively shows two objects from "Paperbowl" and "Sphere" in synthetic CyclesPSTest dataset [15] with only 17 input images. Similar to the results on the synthetic object "Dragon", our method outperformed the counterpart without prior normal. It can be seen that reflectance is rapidly changed on these two object, denote that our method can lead to smoother surface normals compared with the method of using only image input which suffer from wide varieties of real-world materials.

In addition, we reported the convergence error in Fig. 5. As shown in Fig. 5, our method of using Lambertian priors achieved lower convergence error in the training processing. It shows that our network is more effective for feature extraction and regression than the previous mapping method learning surface normal from only input images.

Moreover, the convergence of our method is faster, which means that our method can further benefit from fewer training samples condition. To prove this, we tabulated the performance with fewer training datasets in Table II. The training with fewer dataset causes larger errors on methods whether using Lambertian priors. However, our method with priors performs slight decrease (0.34°) when using Blobby dataset with 25600 samples, while the method without priors reports a larger drop (0.81°). In fact, if our method was trained by very few samples, then the error would be close to least square priors [5]. However, the errors of the method without priors might be unacceptable.

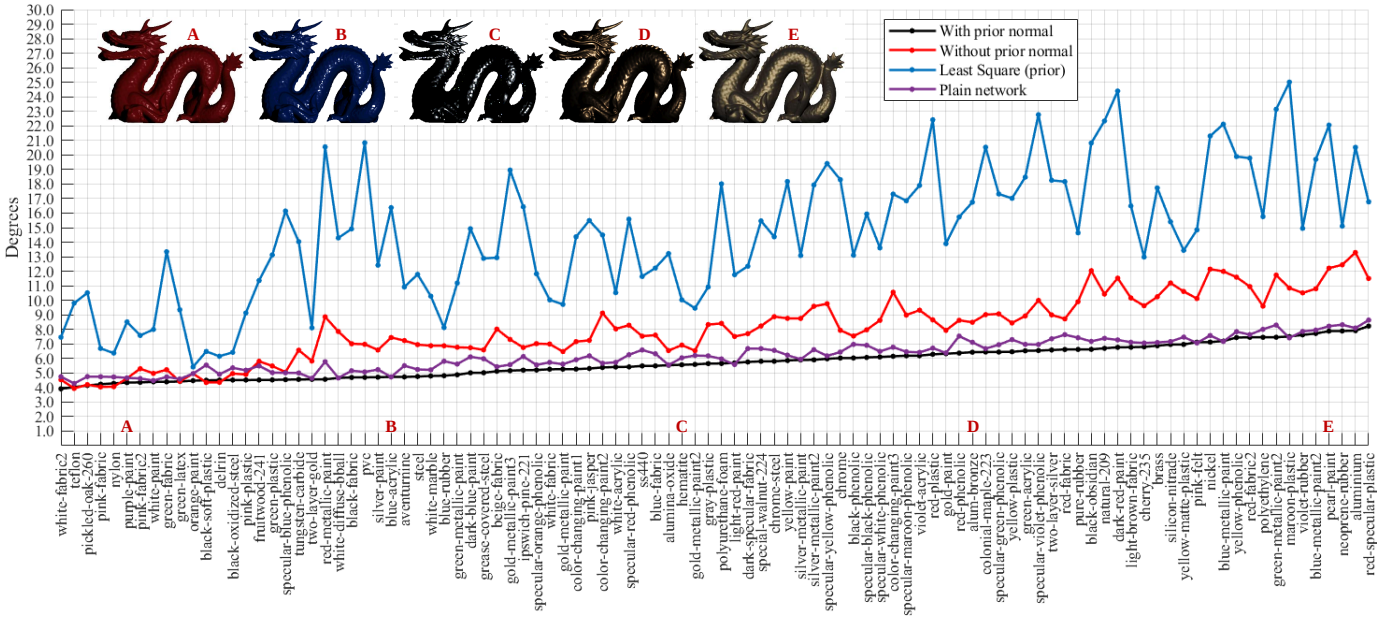

Fig. 3. The MAE in degrees of the estimated surface normals on the samples of "Dragon" with 100 materials from MERL BRDFs [48]. We report the performance of our default settings, without prior normal, plain network (with prior normal), and least square method (the prior normal) [5].

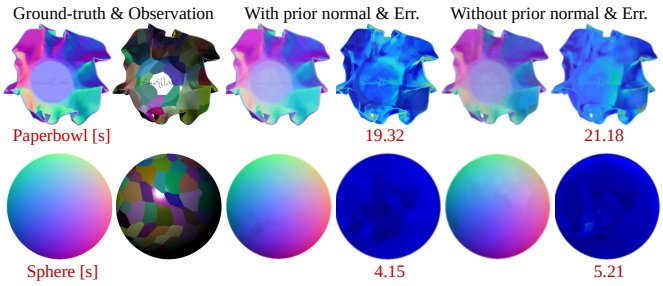

Fig. 4. Visual comparisons between networks with Lambertian priors and without prior normal. We show the results of "Paperbowl" and "Sphere" in synthetic CyclesPSTest dataset [15] with only 17 input images, [s] stands for Specular.

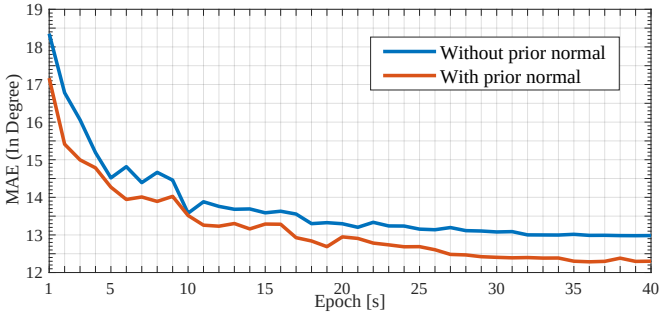

Fig. 5. Convergence comparison on the validation set. The blue curve represents the network without the prior normal while the orange curve is the network using Lambertian priors. For the network without the prior normal, we keep the same architecture but remove the second residual network which handles the prior normal, and all the concatenate operations. Both two networks are trained with the same parameters and 32 images as the input.

TABLE II
NORMAL ESTIMATION RESULTS ON THE VALIDATION SET WITH 32 INPUT IMAGES. BLOBBY DATASET CONTAINS 25660 TRAINING SAMPLES WHILE SCULPTURE GIVES 58700 SAMPLES.

Training datasets	With prior normal	without prior normal
Only Blobby	**12.64**	13.79
Only Sculpture	**12.39**	13.23
Blobby + Sculpture	**12.30**	12.98

We also observed that the MAE of the validation set is larger than that of the DiLiGenT dataset (see in Section IV-C). The reason for this may be because the illumination directions of the validation set are randomly generated in an upper-hemisphere, while the illumination directions of the DiLiGenT dataset are more clustered, which benefits the learning with fewer cast shadows.

2) Effectiveness of different priors: Experiments with IDs 0 and 2 show the influence of different prior inputs on our network. In ID 2, we use the results of the rank minimization method [9] as the prior normal. Referring to Table I, we discover that the prior normal from the rank minimization is slightly better than that from the Lambertian model on MAE and $< err_{30^\circ}$. This is due to the fact that the rank minimization method [9] has better performance than the Lambertian model. However, we still choose the Lambertian surface normal as the priors because of the multi-facet. First, the rank minimization method needs much time for detecting and removing the outliers. Therefore, the time-consumption of the network using rank minimization priors is demanding on both the training (65 hours for training, while Lambertian priors only need 19 hours) and the test stages. Second, the rank minimization priors are limited in varying reflectance surfaces. As shown in Fig .4, the results of using rank minimization as priors are worse than those of Lambertian priors on object "Sphere". The boundary errors can be found on the predicted surface normal, and the MAE of it is larger. This might be because the outlier removal methods such as rank minimization are generally effective for limited categories of reflectance.

3) Effectiveness of residual blocks in the feature extraction stage: Experiments with IDs 0 and 3 show the performance of residual architectures and plain counterparts in the feature extraction stage. Referring to Table I, experiments show that applying residual blocks in the feature regression stage has a lower mean angular error. We also found that the $< err_{15^\circ}$

had a relatively large drop when the plain network was used. Also, as shown in Fig. 3, It can be seen that our method has a slightly small error among most materials and average MAE. The results suggested the residual architecture increased the accuracy of surface normal estimation. The reason might be because the residual blocks can effectively avoid gradient vanishing [46] in a deep network. Also, we argue that the shortcut fuses previous blocks, which is a combination of features at different levels and scales. In addition, the shortcut structure is equivalent to adding all the information of the previous layer image in each block, which retains more original features.

We further compare the different architectures of the residual module. As tabulated in Table III, our default settings (6 residual blocks) achieve better performance compared with using fewer residual blocks. We find that our default settings are slightly worse than 7 residual blocks on MAE. However, the additional residual blocks increase the parameters and training time. For simplifying the complexity of the network, we just remain the 6 blocks eventually. Also, we compare our method with a more simplified single residual branch (directly handle all channels of input, rather than handle the 1-3 channels and 4-9 channels by two residual branches, as shown in Fig. 2 *[BOTTOM]*). However, the performance of using the single residual branch is worse.

TABLE III
RESULTS OF THE DIFFERENT RESIDUAL BLOCKS ARCHITECTURES TESTING ON THE VALIDATION SET.

Architectures	MAE	$< err_{15^\circ}$	$< err_{30^\circ}$
ours (6 residual blocks)	12.30	**84.39%**	**94.86%**
3 residual blocks	12.37	84.35%	94.83%
5 residual blocks	12.32	84.38%	94.85%
7 residual blocks	**12.29**	**84.39%**	94.85%
Single branch	12.52	84.21%	84.77%

4) Effectiveness of concatenating prior normal: Experiments with IDs 0, 4, and 5 shown in Table I reported the effectiveness of concatenating prior normals in the regression stage. For ID 4, we concatenated the prior normal to the third deconvolutional layer in the regression stage. For ID 5, we removed the concatenation operation in the regression stage (without any prior normal). Table I shows that the default settings (concatenating the prior normal to the second deconvolutional layer) helped us to boost the performance. In particular, non-concatenating in the regression stage has a negative impact on the predicted results. This suggests that adding prior normals in deep feature layers will enrich details and increase accuracy. We also noted that the performance was slightly worse, when moving the concatenation back to the 3rd deconvolutional layer. This may be due to the fact that the subsequent up-sampling and down-sampling operations well support the feature fusion in different scales.

C. Evaluation on the DiLiGenT benchmark dataset

We reported the results on the DiLiGenT benchmark dataset [14] with 96 input images in Table IV. In Table IV, we compared our Lambertian model guided photometric stereo network with both traditional algorithms and deep learning methods. For traditional algorithms, we evaluate the Lambertian baseline (our prior normal) [5], robust methods [9], [10] of outlier rejection-based technologies (robust method), and analytic and empirical models [6]–[8], [28], [33], [34]. For deep learning methods, we also compared with several state-of-the-art networks, such as DPSN [12], IRPS [41], PS-FCN [13], CNN-PS [15], Attention-PSN [42], DR-PSN [44], and GPS-Net [43]. Besides, we also evaluated our method against several deep learning methods with sparse inputs (with 10 input images) [15], [37], [38] and flexible inputs methods [5], [10], [13], [34], [44] shown in Table V. Note that LMPS method [37] takes 10 optimal images as inputs, while other sparse methods takes 10 random images as input. In particular, we reported our method of training with 10, 32, and 64 input images, respectively.

As shown in Tables IV and V, our method (default) outperformed the other state-of-the-art methods on both 96 and 10 input images, with higher average MAE. In particular, it can be seen that our prior guided photometric stereo network substantially improves the cases with strong non-Lambertian and non-convex surfaces, such as "Buddha", "Cow", "Harvest", and "Reading". Therefore, we can reveal that providing prior normal has the effectiveness of handling specularity and cast shadows. We compared and showed the results of these objects in Fig. 6. It can be observed that our method recovered accurate surface normals on the regions with specularity, cast shadows, and crinkles, such as the collar of the "Buddha", the wrinkled clothes of the "Reading", and the pocket of the "Harvest". Note that the original CNN-PS [15] discarded the first 20 images of "Bear" (tested with 76 images remaining), achieving a much better MAE of 4.25° for "Bear". The reason for discarding first 20 images of "Bear" is that the intensity values around the stomach region are wrong. However, all the images of "Bear" were used in evaluating all the other methods, except CNN-PS. For a fair comparison, we show the results based on the same test images (all 96 images).

In addition, we furthermore explored the influence of different input images in training. As shown in Tables IV and V, we reported the results of our method trained with 10 and 64 images respectively (64 is the maximum number of the images in the training dataset). It can be seen that the method trained with 64 images achieved even better performance than the one trained with 32 images when using 96 images. On the contrast, the method trained with 10 images outperformed the method when using 10 images. In other words, similar input images between the training and the evaluation will benefit the estimation of the surface normal. The reason is that the prior normal is related to the input images, where the differences of prior normals caused by varying numbers of input images will affect the patterns the method learned to some extent. In order to obtain the best performance, we, therefore, recommend that a similar number of input images be used during the training and testing. Nonetheless, our default setting (trained with 32 images) has achieved the state-of-the-art performances on the measurement with 96 and 10 images (for a fair comparison).

TABLE IV

COMPARISON OF DIFFERENT METHODS ON THE DILIGENT BENCHMARK DATASET. ALL METHODS ARE EVALUATED WITH 96 IMAGES. HERE, WE MEASURE MAE IN DEGREES. OUR METHOD WAS TRAINED WITH 10, 32, 64 IMAGES, RESPECTIVELY. BLACK BOLD TEXTS REPRESENT THE BEST PERFORMANCE, AND UNDERLINED TEXTS REPRESENT THE SECOND BEST.

Method	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Avg.
Baseline (Least squares) [5]	4.10	8.39	14.92	8.41	25.60	18.50	30.62	8.89	14.65	19.80	15.39
Monotonic BRDF [33]	13.58	19.44	18.37	12.34	7.62	17.80	19.30	10.37	9.84	17.17	14.58
Matrix rank = 3 [10]	2.54	7.32	11.11	7.21	25.70	16.25	29.26	7.74	14.09	16.17	13.74
Rank minimization [9]	2.06	6.50	10.91	6.73	25.89	15.70	30.01	7.18	13.12	15.39	13.35
Consensus constraint [8]	3.55	11.48	13.05	8.40	14.95	14.89	21.79	10.85	16.37	16.82	13.22
2D discrete table [6]	2.71	5.96	12.54	6.53	21.48	13.93	30.50	7.23	11.03	14.17	12.61
Multi-Ward models [28]	3.21	6.62	14.85	8.22	9.55	14.22	27.84	8.53	7.90	19.07	12.00
Bivariate BRDF [7]	3.34	7.11	10.47	6.74	13.05	9.71	25.95	6.64	8.77	14.19	10.60
Bi-polynomial [34]	1.74	6.12	10.60	6.12	13.93	10.09	25.44	6.51	8.78	13.63	10.30
SDPS-Net [39]	2.77	6.89	8.97	8.06	8.48	11.91	17.43	8.14	7.50	14.90	9.51
DPSN [12]	2.02	6.31	12.68	6.54	8.01	11.28	16.86	7.05	7.86	15.51	9.41
IRPS [41]	**1.47**	5.79	10.36	5.44	6.32	11.47	22.59	6.09	7.76	11.03	8.83
PS-FCN [13]	2.82	7.55	7.91	6.16	7.33	8.60	15.85	7.13	7.25	13.33	8.39
CNN-PS [15]	2.12	12.30	8.07	**4.38**	7.92	**7.42**	**13.83**	**5.37**	**6.38**	12.12	7.99
Attention-PSN [42]	2.93	**4.86**	7.75	6.14	6.86	8.42	15.44	6.92	6.97	12.90	7.92
DR-PSN [44]	2.27	5.46	7.84	5.42	7.01	8.49	15.40	7.08	7.21	12.74	7.90
GPS-Net [43]	2.92	5.07	7.77	5.42	**6.14**	9.00	15.14	6.04	7.01	13.58	7.81
Ours (Trained with 10 images)	2.57	5.89	8.94	7.10	7.54	8.82	15.48	7.68	7.71	11.53	8.33
Ours (Trained with 64 images)	2.49	5.64	**7.70**	6.45	6.23	8.36	14.67	7.12	7.22	**10.89**	**7.68**
Ours (Trained with 32 images, default)	2.51	5.77	7.88	6.56	6.29	8.40	14.95	7.21	7.40	**11.01**	**7.80**

TABLE V

COMPARISON OF DIFFERENT METHODS ON THE DILIGENT BENCHMARK DATASET. WE NOTE THAT ALL METHODS ARE EVALUATED WITH 10 IMAGES FOR MAE IN DEGREES. OUR METHOD WAS TRAINED WITH 10, 32, 64 IMAGES, RESPECTIVELY. BLACK BOLD TEXTS REPRESENT THE BEST PERFORMANCE, AND UNDERLINED TEXTS REPRESENT THE SECOND BEST.

Method	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Avg.
Baseline (Least squares) [5]	5.09	11.59	16.25	9.66	27.90	19.97	33.41	11.32	18.03	19.86	17.31
Bi-polynomial [34]	5.24	9.39	15.79	9.34	26.08	19.71	30.85	9.76	15.57	20.08	16.18
Matrix rank = 3 [10]	**3.33**	7.62	13.36	8.13	25.01	18.01	29.37	8.73	14.60	16.63	14.48
CNN-PS [15]	9.11	14.08	14.58	11.71	14.04	15.48	19.56	13.23	14.65	16.99	14.34
PS-FCN [13]	4.02	7.18	9.79	8.80	10.51	11.58	18.70	10.14	9.85	15.03	10.51
SPLINE-Net [38]	4.96	**5.99**	10.07	7.52	8.80	10.43	19.05	8.77	11.79	16.13	10.35
LMPS [37]	3.97	8.73	11.36	**6.69**	10.19	10.46	17.33	**7.30**	9.74	14.37	10.02
DR-PSN	3.83	7.52	**9.55**	7.92	9.83	10.38	17.12	9.36	**9.16**	14.75	9.94
Ours (Trained with 10 images)	3.62	7.36	9.61	7.66	**8.42**	**10.17**	**16.70**	9.24	9.38	**14.15**	**9.63**
Ours (Trained with 64 images)	3.94	7.60	9.83	7.94	**8.63**	10.38	**17.07**	9.45	9.73	14.42	9.94
Ours (Trained with 32 images, default)	3.86	7.49	9.69	7.82	**8.55**	**10.31**	**16.94**	9.28	9.54	**14.30**	**9.78**

D. Evaluation on the Light Stage Data Gallery

We further evaluated our method on a more complex dataset with general non-Lambertian materials. Fig. 8 shows the results of our method using the Light Stage Data Gallery [50]. We show the qualitative outcomes for four complex objects "Helmet", "Kneeling", "Standing", and "Plant" in the dataset, due to the absence of ground-truth surface normals. Similarly, our method was trained with 32 images while being evaluated with random 96 observations in all 253 images.

As shown in Fig. 8, the estimated normal keeps the details without blur, such as the screws of the "Helmet", the hair of the "Kneeling", the lumpy-looking clothes of the "Standing", and the succulent plants of the "Plant". Note that the reflectance of plants was not trained in the training process. However, the result of object "Plant" is quite visually accurate, which shows the robustness of our method. We can also see the accurate reconstruction of the cast shadow areas (red boxes in Fig .8). The reconstructed surface normal and the integral mesh convincingly reflect the shapes of the objects, demonstrating the accuracy of our physical prior photometric stereo network. We also observed that the estimated surface normal of the same objects, such as "Standing", are with certain noise. It may be due to the poor quality of the observations of "Standing" with noise, where the high-frequency noise existing in observation may affect the generation of the prior normal.

E. Limitations

We also noticed that the proposed method did not achieve the best performance on some objects of the DiLiGenT dataset [14], such as "Ball" and "Cat", as shown in Fig. 7. We argue that these objects usually have few region with non-Lambertian reflectance (specularities and shadows). In this case, our method does not outperform others on MAE, such as IRPS [41] and DPSN [12]. However, our error map of "Cat" shown in Fig. 7 is more robust than the others: compared with others, our method handles better in the non-convex regions (crinkles). Also, as shown in Fig. 3, our method with prior normal reports a slightly worse MAE than counterpart without priors, on very few materials in MERL BRDFs dataset [48], which show almost diffused properties (such as "teflon", "pink-fabric", and "nylon"). It also illustrates that our method may exist limitation on very slight non-Lambertian samples.

Fig. 6. Quantitative results for strong non-Lambertian real-world scenes from the DiLiGenT benchmark dataset. Err. is short for error map. The numbers under the error maps represent MAE in degrees. The contrast of observations are adjusted for easy view. The red boxes are regions with specularities, the white boxes are the regions with cast shadows, and the orange boxes represent the region with non-convex surfaces (crinkles). Our method produces more accurate estimations in those regions compared with the other methods.

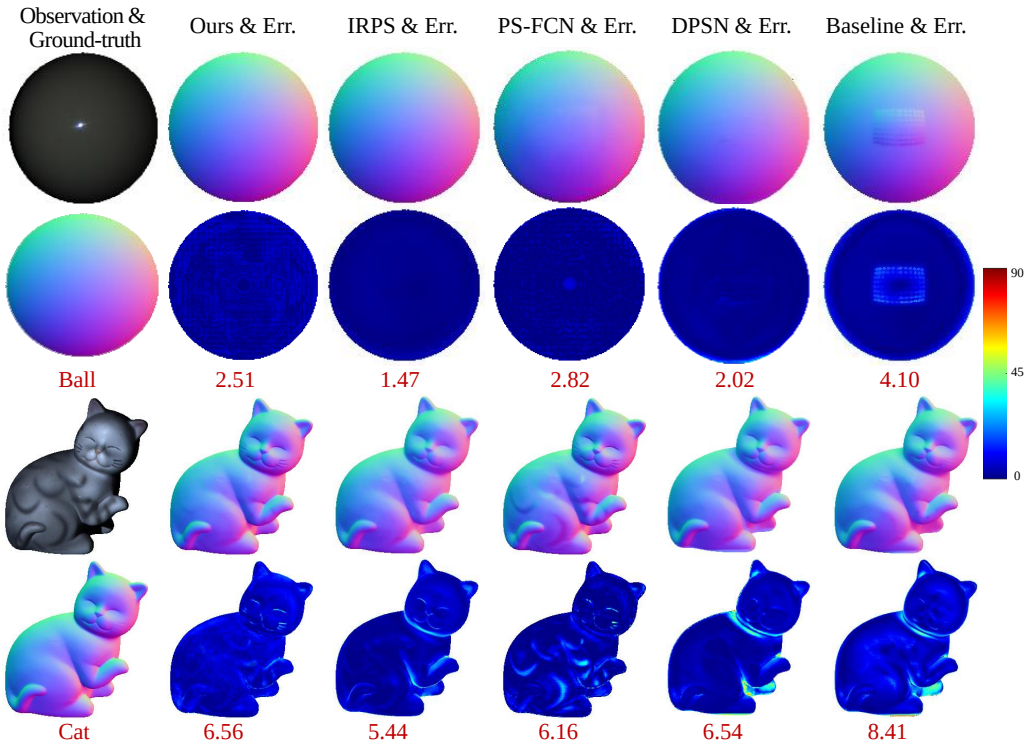

Fig. 7. Quantitative results on"Ball" and "Cat" from the DiLiGenT benchmark dataset. Err. is the abbreviation for the error map. The numbers under the error maps represent MAE in degrees. The contrast of the observations is adjusted for easy view.

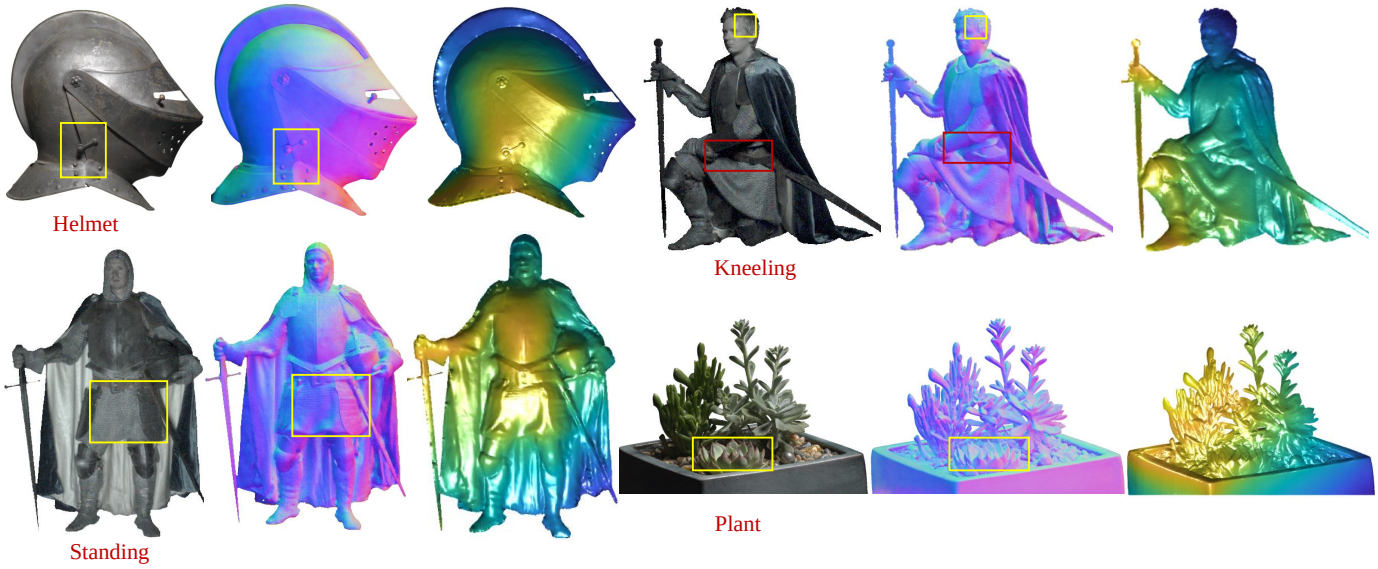

Fig. 8. Qualitative results of our method on objects "Helmet", "Kneeling", "Standing", and "Plant". The contrast of observations is adjusted for easy viewing. For each object, the surface normal estimation and 3D reconstruction using [16] are shown after the observation. The yellow boxes are the regions with complex structures, while the red boxes represent the regions with cast shadows.

V. Conclusions and Future Work

In this paper, we have proposed a Lambertian normal guided photometric stereo network, which utilizes the prior normal to derive accurate surface measurement. Compared with previous deep learning approaches that derives the normal space from the RGB space, our method takes the mapping in the same normal space and pay more attention to the errors in the prior normal. Ablation experiments have illustrated that our method performs more accurate reconstruction. Moreover, the convergence of our method is faster than the traditional methods using the observations to derive the surface normal, which means that our method can be trained with fewer samples. Extensive quantitative and qualitative comparisons on both real (the DiLiGenT benchmark and the Light Stage Data Gallery) and synthetic datasets (the "Dragon" and the CyclesPSTest) have shown that our method outperforms the state-of-the-art methods. The examples have demonstrated that our Lambertian priors photometric stereo network better

handles the surface normal in strong non-Lambertian materials and surfaces with varying reflectance. In the future work, we will explore several alternative normals as the priors. For example, better priors further improve the estimation of the surface normals, such as the high-quality photometric stereo with outlier rejection.

Furthermore, the proposed priors guided photometric stereo network can also support wider applications. For instance, our framework can be used in a non-ideal illuminations environment, where the illuminations are not parallel light or with extra natural illumination. In these tasks, the priors normal can be calculated under the ideal illumination assumption and then refined with the network.

REFERENCES

[1] Mingjun Ren, Lingbao Kong, Lijian Sun, and ChiFai Cheung, "A curve network sampling strategy for measurement of freeform surfaces on coordinate measuring machines," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 11, pp. 3032–3043, 2017.

[2] Jieji Ren, Zhenxiong Jian, Xi Wang, Ren Mingjun, Limin Zhu, and Xiangqian Jiang, "Complex surface reconstruction based on fusion of surface normals and sparse depth measurement," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.

[3] Yakun Ju, Xinghui Dong, Yingyu Wang, Lin Qi, and Junyu Dong, "A dual-cue network for multispectral photometric stereo," *Pattern Recognition*, vol. 100, pp. 107162, 2020.

[4] Muwei Jian, Junyu Dong, Maoguo Gong, Hui Yu, Liqiang Nie, Yilong Yin, and Kin-Man Lam, "Learning the traditional art of chinese calligraphy via three-dimensional reconstruction and assessment," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 970–979, 2019.

[5] R. J Woodham, "Photometric method for determining surface orientation from multiple images," *Optical Engineering*, vol. 19, no. 1, pp. 139–144, 1980.

[6] Neil Alldrin, Todd Zickler, and David Kriegman, "Photometric stereo with non-parametric and spatially-varying reflectance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[7] Satoshi Ikehata and Kiyoharu Aizawa, "Photometric stereo using constrained bivariate regression for general isotropic surfaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2179–2186.

[8] Tomoaki Higo, Yasuyuki Matsushita, and Katsushi Ikeuchi, "Consensus photometric stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1157–1164.

[9] Lun Wu, Arvind Ganesh, Boxin Shi, Yasuyuki Matsushita, Yongtian Wang, and Yi Ma, "Robust photometric stereo via low-rank matrix completion and recovery," in *Proceedings of the Asian Conference on Computer Vision*. Springer, 2010, pp. 703–717.

[10] Satoshi Ikehata, David Wipf, Yasuyuki Matsushita, and Kiyoharu Aizawa, "Robust photometric stereo using sparse regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 318–325.

[11] Daisuke Miyazaki, Kenji Hara, and Katsushi Ikeuchi, "Median photometric stereo as applied to the segonko tumulus and museum objects," *International Journal of Computer Vision*, vol. 86, no. 2-3, pp. 229, 2010.

[12] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita, "Deep photometric stereo network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 501–509.

[13] Guanying Chen, Kai Han, and Kwan-Yee K Wong, "Ps-fcn: A flexible learning framework for photometric stereo," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–18.

[14] B Shi, Z Mo, Z Wu, D Duan, SK Yeung, and P Tan, "A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo.," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 271–284, 2019.

[15] Satoshi Ikehata, "Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 3–18.

[16] Tal Simchony, Rama Chellappa, and Min Shao, "Direct analytical methods for solving poisson equations in computer vision problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 435–446, 1990.

[17] Shree K Nayar, Katsushi Ikeuchi, and Takeo Kanade, "Shape from interreflections," *International Journal of Computer Vision*, vol. 6, no. 3, pp. 173–195, 1991.

[18] Jens Ackermann, Michael Goesele, et al., "A survey of photometric stereo techniques," *Foundations and Trends® in Computer Graphics and Vision*, vol. 9, no. 3-4, pp. 149–254, 2015.

[19] Steffen Herbort and Christian Wöhler, "An introduction to image-based 3d surface reconstruction and a survey of photometric stereo methods," *3D Research*, vol. 2, no. 3, pp. 4, 2011.

[20] Fredric Solomon and Katsushi Ikeuchi, "Extracting the shape and roughness of specular lobe objects using four light photometric stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 4, pp. 449–454, 1996.

[21] Frank Verbiest and Luc Van Gool, "Photometric stereo with coherent outlier handling and confidence estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[22] Manmohan Chandraker, Sameer Agarwal, and David Kriegman, "Shadowcuts: Photometric stereo with shadows," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[23] Chanki Yu, Yongduek Seo, and Sang Wook Lee, "Photometric stereo from maximum feasible lambertian reflections," in *Proceedings of the European Conference on Computer Vision*. Springer, 2010, pp. 115–126.

[24] Tongbo Chen, Michael Goesele, and H-P Seidel, "Mesostructure from specularity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2006, vol. 2, pp. 1825–1832.

[25] Silvia Tozza, Roberto Mecca, Marti Duocastella, and Alessio Del Bue, "Direct differential photometric stereo shape recovery of diffuse and specular surfaces," *Journal of Mathematical Imaging and Vision*, vol. 56, no. 1, pp. 57–76, 2016.

[26] Athinodoros S Georghiades, "Incorporating the torrance and sparrow model of reflectance in uncalibrated photometric stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003, p. 816.

[27] Hin-Shun Chung and Jiaya Jia, "Efficient photometric stereo on glossy surfaces with wide specular lobes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[28] Dan B Goldman, Brian Curless, Aaron Hertzmann, and Steven M Seitz, "Shape and spatially-varying brdfs from photometric stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 6, pp. 1060–1071, 2009.

[29] Lixiong Chen, Yinqiang Zheng, Boxin Shi, Art Subpa-Asa, and Imari Sato, "A microfacet-based reflectance model for photometric stereo with highly specular surfaces," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3162–3170.

[30] Manmohan Chandraker, Jiamin Bai, and Ravi Ramamoorthi, "On differential photometric reconstruction for unknown, isotropic brdfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2941–2955, 2012.

[31] Neil G Alldrin and David J Kriegman, "Toward reconstructing surfaces with arbitrary isotropic reflectance: A stratified photometric stereo approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[32] Si Li and Boxin Shi, "Photometric stereo for general isotropic reflectances by spherical linear interpolation," *Optical Engineering*, vol. 54, no. 8, pp. 083104, 2015.

[33] Boxin Shi, Ping Tan, Yasuyuki Matsushita, and Katsushi Ikeuchi, "Elevation angle from reflectance monotonicity: Photometric stereo for general isotropic reflectances," in *Proceedings of the European Conference on Computer Vision*. Springer, 2012, pp. 455–468.

[34] Boxin Shi, Ping Tan, Yasuyuki Matsushita, and Katsushi Ikeuchi, "Bi-polynomial modeling of low-frequency reflectances," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, , no. 6, pp. 1078–1091, 2014.

[35] Aaron Hertzmann and Steven M Seitz, "Example-based photometric stereo: Shape reconstruction with general, varying brdfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1254–1264, 2005.

[36] Zhuo Hui and Aswin C Sankaranarayanan, "Shape and spatially-varying reflectance estimation from virtual exemplars," *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, vol. 39, no. 10, pp. 2060–2073, 2016.

[37] Junxuan Li, Antonio Robles-Kelly, Shaodi You, and Yasuyuki Matsushita, "Learning to minify photometric stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7568–7576.

[38] Qian Zheng, Yiming Jia, Boxin Shi, Xudong Jiang, Ling-Yu Duan, and Alex C Kot, "Spline-net: Sparse photometric stereo through lighting interpolation and normal estimation networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8549–8558.

[39] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K Wong, "Self-calibrating deep photometric stereo networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8739–8747.

[40] Yakun Ju, Muwei Jian, Junyu Dong, and Kin-Man Lam, "Learning photometric stereo via manifold-based mapping," in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2020, pp. 411–414.

[41] Tatsunori Taniai and Takanori Maehara, "Neural inverse rendering for general reflectance photometric stereo," in *Proceedings of the International Conference on Machine Learning*, 2018, pp. 4857–4866.

[42] Yakun Ju, Kin-Man Lam, Yang Chen, Lin Qi, and Junyu Dong, "Pay attention to devils: A photometric stereo network for better details," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 2020, pp. 694–700.

[43] Zhuokun Yao, Kun Li, Ying Fu, Haofeng Hu, and Boxin Shi, "Gps-net: Graph-based photometric stereo network," in *Proceedings of Advances in Neural Information Processing Systems*, 2020, p. 33.

[44] Yakun Ju, Junyu Dong, and Sheng Chen, "Recovering surface normal and arbitrary images: A dual regression network for photometric stereo," *IEEE Transactions on Image Processing*, vol. 30, pp. 3676–3690, 2021.

[45] Xi Wang, Zhenxiong Jian, and Mingjun Ren, "Non-lambertian photometric stereo network based on inverse reflectance model with collocated light," *IEEE Transactions on Image Processing*, vol. 29, pp. 6032–6042, 2020.

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2016, pp. 770–778.

[47] Micah K Johnson and Edward H Adelson, "Shape estimation in natural illumination," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2011, pp. 2553–2560.

[48] Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan, "A data-driven reflectance model," *ACM Transactions on Graphics*, vol. 22, no. 3, pp. 759–769, 2003.

[49] Wenzel Jakob, "Mitsuba renderer," 2010.

[50] Per Einarsson, Charles-Felix Chabert, Andrew Jones, Wan-Chun Ma, Bruce Lamond, Tim Hawkins, Mark Bolas, Sebastian Sylwan, and Paul Debevec, "Relighting human locomotion with flowed reflectance fields," in *Proceedings of the 17th Eurographics conference on Rendering Techniques*, 2006, pp. 183–194.

Yakun Ju received the B.Sc degree 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, supervised by professor Junyu Dong. His research interests include 3D reconstruction, deep learning, and image processing.

Muwei Jian received the PhD degree from the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, in October 2014. He was a Lecturer with the Department of Computer Science and Technology, Ocean University of China, from 2015 to 2017. Currently, Dr. Jian is a Professor and Ph.D Supervisor at the School of Computer Science and Technology, Shandong University of Finance and Economics. His current research interests include human face recognition, image and video processing, machine learning and computer vision. Prof. Jian holds 3 granted national patents and has published over 50 papers in refereed international leading journals / conferences

Shaoxiang Guo received his M.Sc degree from the Ocean University of China, Shandong, China, in 2020. He is currently pursuing his Ph.D. degree in the vision lab of Ocean University of China, Qingdao, China, supervised by professor Junyu Dong. His research interests include computer vision, human pose estimation, 3d hand shape estimation, and deep learning technology.

Yingyu Wang received the B.Sc. degree from Chengdu University of Technology in 2017 and received the M.Sc degree from Ocean University of China in 2020. Now he will pursue the Ph.D. degree with University of Technology Sydney. His research interests include computer vision, robotics and SLAM.

Huiyu Zhou received a Bachelor of Engineering degree in Radio Technology from Huazhong University of Science and Technology of China, and a Master of Science degree in Biomedical Engineering from University of Dundee of United Kingdom, respectively. He was awarded a Doctor of Philosophy degree in Computer Vision from Heriot-Watt University, Edinburgh, United Kingdom. Dr. Zhou currently is a full Professor at School of Informatics, University of Leicester, United Kingdom. He has published over 350 peer-reviewed papers in the field. His research work has been or is being supported by UK EPSRC, ESRC, AHRC, MRC, EU, Royal Society, Leverhulme Trust, Puffin Trust, Invest NI and industry.

Junyu Dong 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 Ocean University of China in 2004. He is currently a Professor and the Vice-Dean of the College of Information Science and Engineering, Ocean University of China. His research interests include computer vision, underwater image processing, and machine learning, with more than ten research projects supported by the NSFC, MOST, and other funding agencies.