Recovering Surface Normal and Arbitrary Images: A Dual Regression Network for Photometric Stereo

Yakun Ju, *Graduate Student Member, IEEE*, Junyu Dong, *Member, IEEE*, and Sheng Chen, *Fellow, IEEE*

***Abstract*— Photometric stereo recovers three-dimensional (3D) object surface normal from multiple images under different illumination directions. Traditional photometric stereo methods suffer from the problem of non-Lambertian surfaces with general reflectance. By leveraging deep neural networks, learning-based methods are capable of improving the surface normal estimation under general non-Lambertian surfaces. These state-of-the-art learning-based methods however do not associate surface normal with reconstructed images and, therefore, they cannot explore the beneficial effect of such association on the estimation of the surface normal. In this paper, we specifically exploit the positive impact of this association and propose a novel dual regression network for both fine surface normals and arbitrary reconstructed images in calibrated photometric stereo. Our work unifies the 3D reconstruction and rendering tasks in a deep learning framework, with the explorations including: 1. generating specified reconstructed images under arbitrary illumination directions, which provides more intuitive perception of the reflectance and is extremely useful for visual applications, such as virtual reality, and 2. our dual regression scheme introduces an additional constraint on observed images and reconstructed images, which forms a closed-loop to provide additional supervision. Experiments show that our proposed method achieves accurate reconstructed images under arbitrarily specified illumination directions and it significantly outperforms the state-of-the-art learning-based single regression methods in calibrated photometric stereo.**

***Index Terms*— Photometric stereo, surface normal estimation, 3D reconstruction, deep neural networks, dual regression.**

I. Introduction

THREE dimensional (3D) shape recovery from images is a fundamental problem in computer vision and graphics. In 1980, Woodham [1] proposed the photometric stereo algorithm to recover surface normal of 3D objects from varying light directions, which started a new research direction in this field. Unfortunately, the early algorithms were limited to the Lambertian reflectance model. Over the past four decades, many methods were developed to address this problem by applying outlier rejection technologies [2]–[4] or modeling sophisticated reflectance [5]–[7].

Recently, inspired by the success of the deep learning framework for various computer vision tasks, e.g., image retrieval [8] and objects recognition [9], researchers have applied the deep learning approach to photometric stereo [10]–[14]. The latest works [15], [16] developed methods to learn the surface normal under near-field illuminations. These learning-based methods attempt to address the non-Lambertian reflectance problem by utilizing the powerful learning ability of deep neural networks. Despite improving the accuracy of the surface normal, these existing learning-based methods mainly suffer from two types of limitations. First, the existing deep learning-based methods only focus on the surface normal of 3D objects but ignoring the reconstructed images. However, accurate reconstructed images under different illumination directions intuitively show the texture and anisotropic reflectance properties of the surface, which is useful in visual applications, e.g., virtual reality, where the texture and material of the object are as important as the 3D shape. Second, the previous learning-based methods focus on the single surface normal constraint without other supervision, while blindly increasing the complexity of the learning-based model can hardly improve the accuracy of non-Lambertian estimation, particularly for the errors of the regions associated with cast shadows, specularities, and non-convex structure. To the authors' best knowledge, no deep-learning work to date has explored how to regress to two dimensional (2D) reconstructed images in order to further promote the accuracy of recovering the surface normal. Indeed, in sophisticated traditional model-based photometric stereo methods, the reconstruction loss is precisely used as the objective function. To some extent, our approach combines the objectives of learning and traditional algorithms, learning to approximate the real rendering process instead of explicitly resorting to an image formation model.

To overcome the aforementioned limitations as well as to unify the 3D reconstruction and rendering tasks, we propose a novel dual regression network for calibrated photometric stereo, called DR-PSN for short, which combines the surface normal constraint and reconstructed images constraint. In particular, we use the dual regression scheme to introduce an additional constraint on reconstructed images to reduce the potential space of the surface normal. Specifically, as shown

Manuscript received May 29, 2020; revised October 28, 2020 and December 18, 2020; accepted February 25, 2021. Date of publication March 11, 2021; date of current version March 17, 2021. This work was supported in part by the National Key Research and Development Programme of China under Grant 2018AAA0100602, in part by the National Key Scientific Instrument and Equipment Development Projects of China under Grant 41927805, in part by the International Science and Technology Cooperation Programme under Grant2014DFA10410, and in part by the National Natural Science Foundation of China under Grant 61501417 and Grant 61976123. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ioannis Kompatsiaris. *(Corresponding author: Junyu Dong.)*

Yakun Ju and Junyu Dong are with the Department of Computer Science and Technology, Ocean University of China, Qingdao 266100, China (e-mail: juyakun@stu.ouc.edu.cn; dongjunyu@ouc.edu.cn).

Sheng Chen is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: sqc@ecs.soton.ac.uk).

Digital Object Identifier 10.1109/TIP.2021.3064230

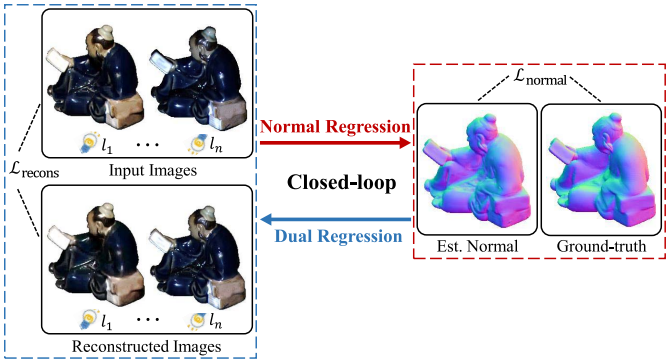


Fig. 1. DR-PSN: The dual regression scheme, which contains a normal regression stage for estimating the surface normal and a dual regression stage to project the surface normal back to reconstructed images.

in Fig. 1, DR-PSN adopts a dual architecture, forming a closed-loop to provide additional supervision. Our method applies an effective channel max-pooling technique [13] in the surface normal task to ensure an arbitrary number of input images. Then in the dual regression task, we design a shortcut to concatenate the shallower fused feature in the normal regression task, which provides the materials information. We also combine encoded lighting with features to output arbitrary specified reconstructed images under different illumination directions. We argue that the combined surface normal, reflectance information, and illumination directions learn the imaging model, which is the inverse task of surface normal estimation. Experimental results show that the estimated surface normal under the proposed additional reconstruction constraint and the closed-loop framework is significantly more accurate than those obtained by traditional photometric stereo algorithms and state-of-the-art learning-based approaches. Moreover, our DR-PSN can generate reconstructed images under arbitrarily specified illumination directions.

II. Related Work

A. Photometric Stereo

Photometric stereo [1] aims at recovering the surface normal of a 3D object from a set of images, captured under different lighting directions with a fixed camera. To solve the limitation on the general reflectance [17], [18], the approaches that researchers developed to address this problem can be categorized into traditional algorithms and learning-based methods.

1) Traditional Methods: Traditional photometric stereo techniques are non-learning-based methods, and they aim to solve the surface normal under unknown reflectance with specularities and cast shadow. Following the survey paper [19], we divide the traditional methods into three categories as outlier rejection-based techniques, sophisticated reflectance model-based techniques, and exemplar-based techniques.

Outlier rejection methods assume that the non-Lambertian surface can be seen as local and sparse (specularity, shadow), which can be discarded as outliers. Several outlier rejection based photometric stereo algorithms have been proposed, including the maximum-likelihood estimation [20], the low rank scheme [21], [22], RANSAC (random sample consensus) [4], the shadow cuts [23], and robust variational method [24]. However, outlier rejection techniques can hardly handle the surface with broad and soft specularities, i.e., the non-Lambertian outliers that are dense and hard to distinguish.

Sophisticated reflectance model-based methods model and approximate non-Lambertian reflectance. Rather than rejecting the specularities and shadow regions as outliers, sophisticated models were developed to approximate all observed pixels. These reflectance models employ sophisticated polynomial functions to approximate the real-world materials, such as the bivariate functions [25]–[28], Ward reflectance model [5], [29], the specular spike reflectance model [7], [30], Blinn-Phong reflectance model [31], Torrance-Sparrow reflectance model [32], and the microfacet models [33]. However, these hand-crafted analytic models are useful only for limited classes of non-Lambertian surfaces because the reflectance models vary dramatically from material to material.

Exemplar-based methods benefit from the additional calibration objects in the same images. The calibration object with the known surface normal transforms the non-Lambertian photometric stereo into a pixel matching problem. For example, Hertzmann and Seitz [34] used a reference sphere to matching the 3D object. However, the material of the reference object has to be the same as the target, which limits the applications of this class of exemplar-based methods.

2) Learning-Based Method: Inspired by the powerful learning ability of deep neural networks, deep learning methods have been introduced to solving the non-Lambertian photometric stereo problem [10], [35]. DPSN [10] was the first proposed method applying deep neural networks to the non-Lambertian photometric stereo. This approach employs a seven-layers fully-connected network to regress the surface normal of the 3D object and adopts the dropout technology [36] to simulate the cast shadow on the images. However, DPSN estimates a normal vector based solely on the single pixel, and the number-fixed and order-fixed manner limit its practical use.

For better estimating the non-Lambertian objects and taking full advantage of the information embedded in the neighborhood, subsequent methods were improved by applying the convolutional neural networks (CNN) [11]–[13], [37], [38]. The works [13] and [39] proposed a fully-convolutional network (FCN) to regress the surface normal, and a channel max-pooling operation was adopted to ensure the arbitrary number of input images. Chen *et al.* [11] also proposed an SDPS-Net to estimate both the surface normal and illumination direction, for uncalibrated photometric stereo. Ikehata [12] proposed another approach, called CNN-PS, which employs the observation map to overcome the fixed inputs problem and to range observation intensities according to light directions. The observation map was also adopted in [37], [38] for inputs with order-agnostic illuminations. Furthermore, Taniai and Maehara [40] proposed an unsupervised learning framework to estimate both the surface normal and reflectance map by minimizing the reconstruction loss. Their method introduced the single constraint and used the physical model to approximately render the re-rendered images, at the cost of expensive computation. Similarly, the works [41]–[43]

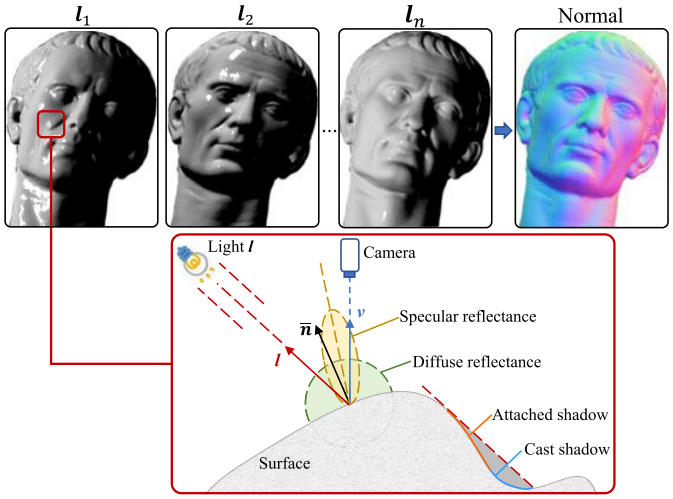


Fig. 2. An example of input images with different illumination directions. In the red box, we show a situation where a pixel on non-Lambertian surface with a normal vector $\bar{\boldsymbol{n}}$ is illuminated by a parallel light source $\boldsymbol{l}$, photoed by a camera in a view direction $\boldsymbol{v}$. The non-Lambertian surface causes both the specularities and diffuse reflections. Furthermore, attached shadow occurs at surface where $\bar{\boldsymbol{n}}^\top \boldsymbol{l} < 0$, while cast shadow occurs at the illumination is occluded by objects.

also utilized reconstruction loss (so-called appearance loss) to inverse the rendering process. Inverse rendering is the problem of estimating illuminations, reflectance properties, and shape from observed appearance. However, these methods all follow the assumption of Lambertian surface, which is generally not the case for real-world reflectance objects. By contrast, our method approximates an accurate imaging process, which can benefit the forward surface normals estimation task.

Unlike all the above methods which only apply the single constraint, our DR-PSN combines both the surface normal and reconstructed image constraints by a closed-loop dual regression architecture. Rather than blindly increasing the complexity of network, we use the normal loss associated with reconstruction loss to address the non-Lambertian photometric stereo, significantly outperforming these previous learning-based methods. Simultaneously, DR-PSN meets the needs of texture visualization, and generates reconstructed images under the arbitrary specified illumination directions.

B. Dual Learning for Enhancement

Dual learning networks [44], [45] contain a primal stage model and a dual stage model to learn two opposite mappings. In the dual learning scheme, the primal task maps the space $\mathcal{X}$ onto the space $\mathcal{Y}$, while the dual task takes the samples from the space $\mathcal{Y}$ and maps them back to the space $\mathcal{X}$. The two opposite mappings enhance the performance of the both tasks, simultaneously. The two tasks are jointly learned and their structural relationship is exploited to improve the learning effectiveness. Therefore, the dual learning scheme outperforms the traditional single regression scheme. Recently, this framework has also been applied to performing image translation without paired training data, such as Dual-GAN [46] and Cycle-GAN [47].

However, our proposed DR-PSN is different from the above unpaired dual learning methods. In our DR-PSN, we introduce the encoded illumination information in the dual regression stage, controlling the reconstructed images with arbitrary specified illumination directions in testing, and thus our reconstructed images can be different from the input images. Hence, our DR-PSN can generate the specified images needed even without the ground-truth of the input images.

III. Notations and Preliminaries

The following standard notations are adopted throughout. Boldface capital letters stand for matrices and tensors, e.g., $\bar{\boldsymbol{N}}$ for the estimated surface normal of an object, while boldface small letters denote vectors, e.g., light direction $\boldsymbol{l}$. We use the subscript $i \in \{1, 2, \cdots, n\}$ to represent the specified observation index, e.g., the i-th illumination direction $\boldsymbol{l}_i$ and the i-th observation image $\boldsymbol{I}_i$. Similarly, we choose the subscript p to denote the index of a pixel in tensors or matrices, e.g., $\bar{\boldsymbol{n}}_p$ is an estimated normal vector at the p-th pixel. Furthermore, we use the $C \times H \times W$ to represent the dimensionality of image, normal and feature map, in which C is some channel number and $H \times W$ represents the spatial resolution, e.g., $\boldsymbol{I}_i \in \mathbb{R}^{3\times H\times W}$, where the first number 3 represents the RGB channels.

Next following the common notations of [11], [40], we recap the fundamental formulation in non-Lambertian photometric stereo. As shown in Fig. 2 [40], given n images under different illumination directions $\boldsymbol{l}_i \in \mathbb{R}^3$, $i \in \{1, 2, \cdots, n\}$, a photometric stereo algorithm calculates the surface normal $\bar{\boldsymbol{n}} \in \mathbb{R}^3$ [40]. Due to the non-Lambertian surface reflectance, the real situation is illustrated in the red box of Fig. 2. Specifically, consider that a pixel on the non-Lambertian surface with the unit normal $\bar{\boldsymbol{n}}$ is illuminated by the parallel light source $\boldsymbol{l}$ with intensity $\boldsymbol{e} \in \mathbb{R}^3$. When this surface is photoed by a linear-response camera in a view direction $\boldsymbol{v} \in \mathbb{R}^3$, the imaging model can be approximated as follows:

$$I = s \cdot \rho\left(\boldsymbol{e}, \bar{\boldsymbol{n}}, \boldsymbol{l}, \boldsymbol{v}\right) \cdot \max\left\{\bar{\boldsymbol{n}}^\top \boldsymbol{l}, 0\right\} + \epsilon, \tag{1}$$

where I denotes the measured intensity of the pixel, s is a binary function for judging cast shadow ($s = 0$ for cast shadow; otherwise $s = 1$), $\rho\left(\boldsymbol{e}, \bar{\boldsymbol{n}}, \boldsymbol{l}, \boldsymbol{v}\right)$ is a bidirectional reflectance distribution function (BRDF), and $\max\left\{\bar{\boldsymbol{n}}^\top \boldsymbol{l}, 0\right\}$ accounts for the attached shadows and shading, while ϵ represents the noise and global illumination effect. In this situation, the BRDF of a non-Lambertian surface exhibits anisotropic characteristics, which makes the numerical resolution more difficult.

Many works design learning frameworks based on Eq. (1) to estimate the surface normal under the problem of non-Lambertian photometric stereo [10]–[13], [37], [38]. However, researchers have seldom explored using reconstructing images (inverse processing) to improve the surface normal estimation in a closed-loop framework.

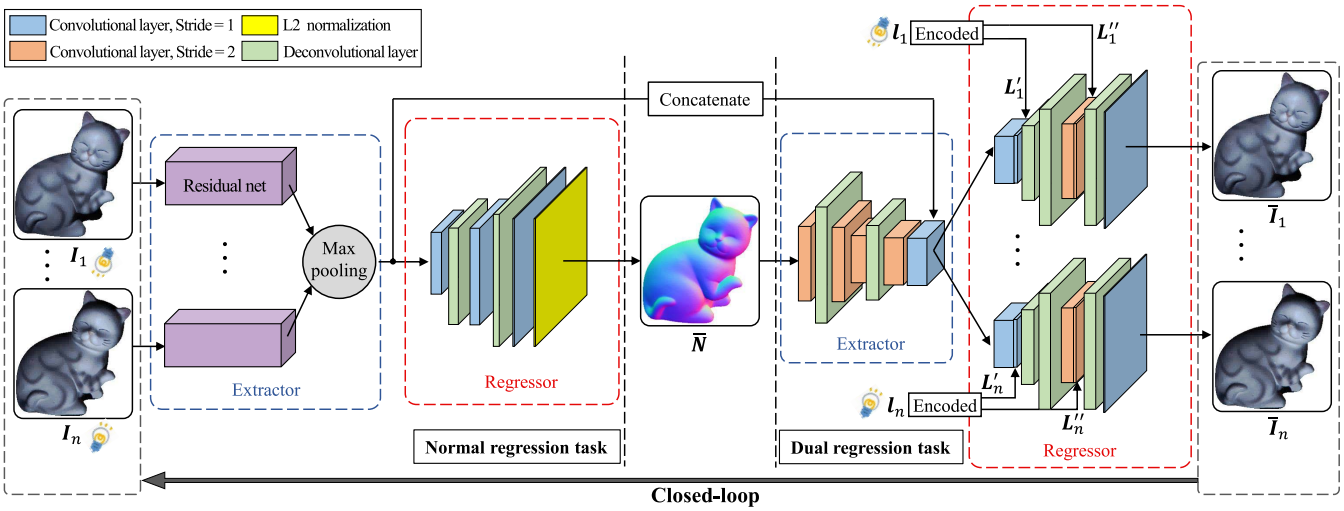


Fig. 3. The details of DR-PSN. The normal regression task estimates the surface normal, while the dual regression task synthesizes each of the reconstructed images $\{\bar{\boldsymbol{I}}_1, \bar{\boldsymbol{I}}_2, \cdots, \bar{\boldsymbol{I}}_n\}$ from the estimated surface normal $\bar{\boldsymbol{N}}$ of the normal regression network. We concatenate the high-dimensional feature from the normal regression task to the dual regression task with a shortcut. To better regress images under specified illumination directions, we also fuse the encoded illumination directions $\boldsymbol{L}'_i$, $\boldsymbol{L}''_i$ with the regressor twice for a corresponding reconstructed image.

IV. Dual Regression Network

We propose a dual regression network to estimate both surface normal and arbitrary reconstructed images. The architecture of our proposed DR-PSN is depicted in Fig. 3.

A. Normal Regression Task

The normal regression task estimates the surface normal of a 3D object. For a 3D object photoed under n distinct illumination directions, we first expand each illumination direction from $\boldsymbol{l}_i \in \mathbb{R}^3$ to a 3-channel illumination tensor $\boldsymbol{L}_i \in \mathbb{R}^{3\times H\times W}$, having the same spatial resolution as the observation image $\boldsymbol{I}_i$. Then, we concatenate all the expanded illumination directions and observation images to a tensor $\boldsymbol{M} \in \mathbb{R}^{nC\times H\times W}$, where $C = 6$ is composed of the three RGB channels and the three illumination direction channels.

In the normal regression task, we seek to find a mapping $\boldsymbol{F}_N : \boldsymbol{M} \rightarrow \bar{\boldsymbol{N}}$, such that the estimation $\boldsymbol{F}_N(\boldsymbol{M})$ is similar to the corresponding real surface normal $\boldsymbol{N}$. More specifically, the normal regression task contains three parts: the extractor $\boldsymbol{F}_{Ne}$, the max-pooling fusion, and the regressor $\boldsymbol{F}_{Nr}$. First, the extractor can be expressed as follows:

$$\boldsymbol{\Phi} = \boldsymbol{F}_{Ne}(\boldsymbol{M}; \boldsymbol{\theta}_{Ne}), \tag{2}$$

where $\boldsymbol{F}_{Ne}(\cdot; \boldsymbol{\theta}_{Ne})$ is the mapping function of the 6 residual blocks with two down-sampling convolutional layers [48] and the learnable parameters $\boldsymbol{\theta}_{Ne}$. We actually compared the architectures of VGG [49]. Among the architectures tested, the residual network [48] is slightly better and, therefore, it was chosen. Inspired by PS-FCN [13], we apply a channel max-pooling operation to handle arbitrary number of inputs. By max-pooling, we obtain a fixed feature map $\boldsymbol{\Phi}' \in \mathbb{R}^{C_1\times H'\times W'}$ from the multi-fusion feature map $\boldsymbol{\Phi} \in \mathbb{R}^{nC_1\times H'\times W'}$, where C_1 is 256, $H' = \frac{1}{4}H$, and $W' = \frac{1}{4}W$. Then the regressor $\boldsymbol{F}_{Nr}$ outputs the surface normal $\bar{\boldsymbol{N}}$, giving $\boldsymbol{\Phi}'$ as:

$$\bar{\boldsymbol{N}} = \boldsymbol{F}_{Nr}(\boldsymbol{\Phi}'; \boldsymbol{\theta}_{Nr}), \tag{3}$$

where $\boldsymbol{F}_{Nr}(\cdot; \boldsymbol{\theta}_{Nr})$ is the regressor with three 3×3 convolutional layers and two 3×3 deconvolutional layers, ending with an $L2$ normalization that makes each pixel's normal $\bar{\boldsymbol{N}}_p$ a unit vector, while $\boldsymbol{\theta}_{Nr}$ are the parameters of the regressor.

B. Dual Regression Task

After the normal regression task, we explore a dual regression task for learning the reconstructed images. The dual regression task aims to learn a function $\boldsymbol{F}_D : \bar{\boldsymbol{N}} \rightarrow \bar{\boldsymbol{I}}_i, \forall i \in \{1, 2, \cdots, n\}$, where $\bar{\boldsymbol{I}}_i$ are expected to approximate the real observation images $\boldsymbol{I}_i$. The architecture of this dual regression task consists of an extractor and a regressor, as can be seen in Fig. 3.

Given the estimated surface normal $\bar{\boldsymbol{N}}$, the extractor $\boldsymbol{F}_{De}$ learns the feature map $\boldsymbol{\Psi} \in \mathbb{R}^{C_1\times H'\times W'}$ as follow:

$$\boldsymbol{\Psi} = \boldsymbol{F}_{De}(\bar{\boldsymbol{N}}; \boldsymbol{\theta}_{De}), \tag{4}$$

where $\boldsymbol{F}_{De}(\cdot; \boldsymbol{\theta}_{De})$ is a network having five 3×3 convolutional layers (four stride $= 2$, and one stride $= 1$) and two 3×3 deconvolutional layers, as shown in Fig. 3, with the learnable parameters $\boldsymbol{\theta}_{De}$. We believe that the fused feature map $\boldsymbol{\Phi}'$ in the normal regression task represents the reflectance information. In order to recover more details in the reconstructed images, therefore, we concatenate $\boldsymbol{\Phi}'$ to $\boldsymbol{\Psi}$ to yield the mixed feature $\boldsymbol{\Psi}' \in \mathbb{R}^{512\times H'\times W'}$ before the regressor $\boldsymbol{F}_{Dr}$. Furthermore, we encode the illumination direction from vector $\boldsymbol{l}_i$ to $\boldsymbol{L}'_i \in \mathbb{R}^{3\times H'\times W'}$ and $\boldsymbol{L}''_i \in \mathbb{R}^{3\times H''\times W''}$, respectively, where $H'' = \frac{1}{2}H$ and $W'' = \frac{1}{2}W$. The encoding is similar to encoding $\boldsymbol{l}_i$ to $\boldsymbol{L}_i$, which expands the vector to include the corresponding spatial resolution. We concatenate these two encoded illuminations to the two convolutional layers in

the regressor which has the structure of two 3×3 convolutional layers and four 3×3 deconvolutional layers. In this way, we can control each reconstructed image $\bar{\boldsymbol{I}}_i$. With the combined reflectance feature $\boldsymbol{\Phi}'$, illumination direction $\boldsymbol{l}_i$, and surface normal feature $\boldsymbol{\Psi}$, the dual regression network simulate the imaging model Eq. (1) with powerful deep learning ability. This dual regression task can be expressed as follows:

$$\bar{\boldsymbol{I}}_i = \boldsymbol{F}_{Dr}(\boldsymbol{\Psi}', \boldsymbol{L}_i', \boldsymbol{L}_i''; \boldsymbol{\theta}_{Dr}), \tag{5}$$

where $\boldsymbol{\theta}_{Dr}$ are the learnable parameters.

C. Dual Regression Loss Function and Training Procedure

We optimize the networks' parameters, $\boldsymbol{\theta}_{Ne}$, $\boldsymbol{\theta}_{Nr}$, $\boldsymbol{\theta}_{De}$ and $\boldsymbol{\theta}_{Dr}$, by minimizing the joint loss function of the above two regression tasks. The joint training loss can be written as:

$$\mathcal{L} = \mathcal{L}_{\text{normal}}(\bar{\boldsymbol{N}}, \boldsymbol{N}) + \lambda_t \mathcal{L}_{\text{recons}}(\bar{\boldsymbol{I}}_i, \boldsymbol{I}_i, \forall i), \tag{6}$$

where λ_t is the weighting for the reconstruction loss.

The first part of the joint loss, $\mathcal{L}_{\text{normal}}$, denotes the normal loss between the estimated surface normal $\bar{\boldsymbol{N}}$ and the ground-truth $\boldsymbol{N}$ in the normal regression task, which is given by

$$\mathcal{L}_{\text{normal}}(\bar{\boldsymbol{N}}, \boldsymbol{N}) = \frac{1}{HW} \sum_{p}^{HW} \left(1 - \bar{\boldsymbol{N}}_p \odot \boldsymbol{N}_p\right), \tag{7}$$

where $\odot$ represents the dot-product operation. If the estimated surface normal $\bar{\boldsymbol{N}}_p$ at pixel p has a similar orientation as the ground-truth $\boldsymbol{N}_p$, $\bar{\boldsymbol{N}}_p \odot \boldsymbol{N}_p$ will be close to 1 and the associated loss will approach 0.

The second part of the joint loss, $\mathcal{L}_{\text{recons}}(\bar{\boldsymbol{I}}_i, \boldsymbol{I}_i, \forall i)$, denotes the reconstruction loss between the reconstructed images $\{\bar{\boldsymbol{I}}_1, \bar{\boldsymbol{I}}_2, \cdots, \bar{\boldsymbol{I}}_n\}$ and the real observation images $\{\boldsymbol{I}_1, \boldsymbol{I}_2, \ldots, \boldsymbol{I}_n\}$ in the dual regression task, and we define $\mathcal{L}_{\text{recons}}(\bar{\boldsymbol{I}}_i, \boldsymbol{I}_i, \forall i)$ as

$$\mathcal{L}_{\text{recons}}(\bar{\boldsymbol{I}}_i, \boldsymbol{I}_i, \forall i) = \frac{1}{n} \sum_{i=1}^{n} \|\bar{\boldsymbol{I}}_i - \boldsymbol{I}_i\|_2^2. \tag{8}$$

We adopt a varying λ_t, rather than a fixed value λ, to control the weight of the reconstruction loss. Specifically, λ_t changes during training epochs. We set $\lambda_t = 0$ for the first training epoch, and increase λ_t by 0.02 after each training epoch. This varying λ_t design is based on the fact that the dual regression task learns the imaging model Eq. (1) with the predicted normals, materials (from the shallower shortcut), and illumination directions. Therefore, the learning processing of the dual regression needs accurate surface normal, and we can realize an accurate estimate by gradually increasing the weight of reconstruction loss, while maintaining a stable training. We also limit the maximum value of λ_t (protection threshold) to 0.8. This prevents the weight of the reconstruction loss to become too big which would make the normal regression task ineffective. Experiments have shown that this varying λ_t strategy is capable of providing powerful supervision for both the surface normal estimation and reconstruction images. Detailed experimental investigation and analysis for the impact of λ_t on the achievable performance are given in Subsection V-B.

The architecture and parameters of our DR-PSN are detailed in Appendix. Our network is implemented in PyTorch [50] and Adam optimizer [51] is used with the default settings of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is initially set to 0.001 and is divided by 2 every 5 epochs. We train the model using a batch size of 32 for 50 epochs and choosing the fixed $n = 32$ as the default number of input images. We set the spatial resolution to $H = W = 32$ for the patch size. It takes about 19 hours for training, using a single RTX 2080 GPU.

V. Experimental Results

To verify the quantitative performance of our model, we use some common metrics to measure the accuracy. For the estimated surface normal, we adopt the widely used mean angular error (MAE) in degree, calculated as

$$\text{MAE} = \frac{1}{T} \sum_{p=1}^{T} \arccos\left(\bar{\boldsymbol{N}}_p \odot \boldsymbol{N}_p\right) \text{ [degree]}, \tag{9}$$

where T is the total number of pixels in an evaluated image. We also measure the percentage (%) that the pixels with angular error less than 15°, which is denoted by $< err_{15^\circ}$. For the reconstructed images, we adopt the commonly used average relative error (REL), which is defined by

$$\text{REL} = \frac{1}{nT} \sum_{i=1}^{n} \sum_{p=1}^{T} \frac{\left|\bar{\boldsymbol{I}}_{i,p} - \boldsymbol{I}_{i,p}\right|}{\boldsymbol{I}_{i,p}}, \tag{10}$$

and the structural similarity index (SSIM) [52]. We evaluate the SSIM with the minimum spatial size mask. We divide the reconstructed images into two categories: belong to the illumination directions of the input images (BI) and not belong to the illumination directions of the input images (NBI). Clearly, for the MAE and REL metrics, the smaller the better, which is indicated by $\downarrow$ after the metrics, while $\uparrow$ after the $< err_{15^\circ}$ and SSIM metrics indicates that the larger the better.

A. Datasets

For training the network, we adopt two 3D datasets that provide the surface normal, namely, the blobby shape dataset [53] and the sculpture shape dataset [54]. Then, we employ the MERL dataset [55] to render the 3D model from the blobby and sculpture datasets, where the MERL dataset contains 100 different BRDFs of real-world materials. Following the rendering settings in [11], we eventually obtain 85212 samples of the training data. For each sample, 64 observation images are rendered by random illumination directions in a half-sphere. We split these samples into a ratio of 99 : 1, for training (84360) and validation (852). With the 99:1 ratio of training to validation, we have large number of training data (84360), which is necessary for training a deep-network based model. The validation set of 852 samples are also sufficiently large to include all surface materials (100 kinds of reflectance in the MERL dataset are all included) as well as to comprehensively represent different types of objects in the training set (simple and complex objects are all available). The rendering processing of training data is shown in Fig. 4. We use the known

TABLE I

PERFORMANCE COMPARISON OF THE SINGLE NORMAL REGRESSION NETWORK (SINGLE) AND THE PROPOSED DUAL REGRESSION NETWORK (DUAL) AS WELL AS THE IMPACT OF DIFFERENT WEIGHTING STRATEGIES FOR DUAL REGRESSION TASK, IN TERMS OF FOUR METRICS AVERAGED OVER THE VALIDATION DATASET. Δ STANDS FOR THE RATE-INCREASING, AND PT REPRESENTS THE PROTECTION THRESHOLD. FOR $< err_{15^\circ}$ AND SSIM, THE HIGHER, THE BETTER. FOR MAE (°) AND REL, THE LOWER, THE BETTER

IDs	Variants	Surface normal		Reconstructed images	
		MAE (°) ↓	$< err_{15^\circ}$ ↑	SSIM ↑	REL ↓
0	Dual, proposed linear λ_t ($\Delta = 0.02$, PT= 0.8)	**11.47**	84.99%	0.947	0.171
1	Single $\lambda = 0$	12.53	81.55%	-	-
2	Dual, fixed $\lambda = 0.1$	11.64	84.61%	0.895	0.235
3	Dual, fixed $\lambda = 0.5$	11.88	82.94%	0.939	0.182
4	Dual, fixed $\lambda = 1$	12.50	81.79%	**0.963**	**0.166**
5	Dual, linear λ_t ($\Delta = 0.02$, PT= 0.6)	11.57	**85.01%**	0.926	0.197
6	Dual, linear λ_t ($\Delta = 0.02$, PT= 1)	11.80	83.33%	0.951	0.169
7	Dual, linear λ_t ($\Delta = 0.01$, PT= 0.8∗)	11.58	84.52%	0.914	0.209
8	Dual, linear λ_t ($\Delta = 0.04$, PT= 0.8)	11.55	84.39%	0.929	0.175
9	Dual, quadratic λ_t ($\Delta = 0.001$, PT= 0.8)	11.49	84.95%	0.916	0.197
10	Dual, quadratic λ_t ($\Delta = 0.0005$, PT= 0.8)	11.58	84.78%	0.934	0.188

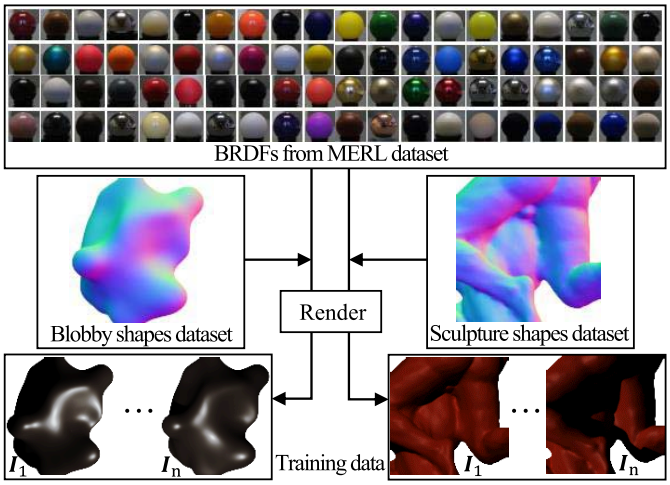


Fig. 4. The rendering processing of training data. We use the MERL BRDFs dataset [55] to render the blobby shape dataset [53] and the sculpture shape dataset [54]. For each training sample, we render 64 observation images, i.e., $n = 64$, in our settings. The intensity of observation images is adjusted for easy viewing.

image intensity to normalize the input image and reconstruct the original image intensity after regression.

Note that the rendered training set CyclesPS employed in [12], which uses the Disney's principled BSDFs [56] to render the training set, is inappropriate for training many deep-learning-based methods, including our DR-PSN, SDPS-Net [11], PS-FCN [13], IRPS [40], etc. The reason is due to the scale of CyclesPS, which only contains 45 samples (images). This dataset is suitable for training the methods, CNN-PS [12] and SPLINE-Net [38], which have the characteristic of 'using per-pixel observation as an observation map'. For these per-pixel methods, every pixel of the image can be seen as a training sample. Therefore, 45 samples are sufficient. However, CyclesPS dataset is far too small for training the most deep-learning-based methods, which use patches or whole images as inputs.

To evaluate the performance of our network, we employ two real photometric stereo datasets for testing, namely, the DiLiGenT benchmark dataset [19] and the Light Stage Data Gallery [57]. The DiLiGenT dataset is the most widely used real-world dataset for evaluating photometric stereo. It contains ten real-world scenes of photometric stereo, which is challenging for its strong non-Lambertian surfaces and complex structures. Each object has 96 observation images illuminated under different directions. The Light Stage Data Gallery contains six objects without ground-truth. Therefore, We quantitatively evaluate our method on the DiLiGenT dataset while qualitatively evaluate on the Light Stage Data Gallery. We change $\boldsymbol{I}_i$ in the dual regression task to produce arbitrary reconstructed images under specified illumination directions in the test.

B. Effectiveness of Dual Regression Network

We first evaluate the effectiveness of our dual regression network by comparing it with the single normal regression network without the dual regression task. We keep the settings and architecture of the single normal regression network the same with the normal regression task of our dual regression network. Moreover, we explore the influence of the weight of the reconstruction loss in the dual regression task. We validate the effectiveness of our strategy of linearly changing λ_t with training epochs by comparing it with different fixed weight values λ (0.1, 0.5, and 1). Note that the single normal regression network can be regarded as a special case of our dual regression network with the fixed weight $\lambda = 0$. We also evaluate the impact of different rate-increasing (Δ) settings and different protection threshold (PT) values on the achievable performance of our linear λ_t strategy. Furthermore, we compare this linearly varying λ_t with the quadratically changing λ_t. For all these experiments, the performance are evaluated on the validation set with 64 input images for each sample, where all the reconstructed images belong to the BI class. The results are summarized in Table I.

1) Comparison With Fixed λ: Experiments with IDs 0 and 1 demonstrate that the dual regression network with the proposed linearly increasing weight λ_t achieves better performance in surface normal estimation than the single regression network ($\lambda = 0$). Specifically, for the dual regression network, the MAE is 11.47° and the $< err_{15^\circ}$ metric is 84.99%, while for the single regression network, the MAE is 12.53° and the

$< err_{15^\circ}$ metric is 81.55%. This confirms that the dual regression task (learning reconstructed images) enhances surface normal learning. The reason is that the dual regression introduces an additional constraint on reconstructed images to provide additional supervision, effectively reducing the potential space of surface normal learning. In other words, it reduces the learning difficulty of surface normal.

Compared with the dual regression networks with fixed weight values λ (experiments with IDs 2, 3, and 4), the dual regression network adopting the proposed strategy of linearly changing λ_t (experiment with ID 0) achieves the best accuracy of surface normal estimation and the second best accuracy of reconstructed images, which is very close to the case of adopting the fixed $\lambda = 1$. Although the dual regression network with $\lambda = 1$ achieves the best accuracy of reconstructed images, it hardly improves the performance of surface normal estimation compared with the single regression network (experiment with ID 1). This suggests that fixing the weight of the dual regression task to $\lambda = 1$ may be too big for producing sufficient beneficial effect on the normal regression task. Note that all the dual regression networks with the fixed nonzero weights λ outperform the single normal regression network with $\lambda = 0$, in terms of the accuracy of surface normal estimation. This is because the dual regression task provides an additional constraint on surface normal estimation. The effectiveness of the proposed strategy of changing λ_t with training epochs can be explained as follows. As explained previously, the dual regression task learns the imaging processing with the predicted normals, materials, and illumination directions, and therefore the learning process of the dual regression needs accurate surface normal. We set $\lambda_t = 0$ in the first epoch to ensure that the primal normal regression task is well learned first. Then, we gradually increase λ_t after each epoch, gradually introducing the additional supervision for optimizing the surface normal estimation.

2) Impact of Rate-Increasing/Protection Threshold: Given $\Delta = 0.02$, experiments with IDs 0, 5, and 6 evaluate the impact of protection threshold. It can be seen that choosing PT $= 0.8$ outperforms the other two settings, in terms of the metrics measuring the both tasks. Specifically, with PT $= 0.6$, the performance of surface normal estimation is almost equal to that of PT $= 0.8$ (slightly worse in MAE and slightly better in $< err_{15^\circ}$). However, the performance of reconstructed images is degraded in this case. This shows that a too small PT may not provide sufficient supervision to the dual regression task. By contrast, with PT $= 1$, the accuracy of the predicted surface normal becomes considerably worse than the case of PT $= 0.8$, although the results of reconstructed images are slightly better. We also find that the loss of the dual regression network cannot decrease any further after 40 epochs when PT $= 1$. The results thus indicate that PT $= 1$ can hardly produce sufficient beneficial effect on the normal regression task. The empirical results therefore suggest that limiting the maximum value of λ_t to 0.8 protects the priority of the normal regression task, which in turn provides the best overall performance.

Experiments with IDs 0, 7, and 8 compare the different rate-increasing values on the achievable performance, given PT $=$ 0.8. Note that the weight λ_t in experiment with ID 7 cannot reach PT (Due to $\Delta = 0.01$, the maximum weight reached is 0.5), and this is marked with * in Table I. It can be seen that our choice of $\Delta = 0.02$ achieves better performance than the other two settings. The worse performance of experiment with ID 7 is due to insufficient supervision, while with $\Delta = 0.04$, the rate increasing may be too large for a stable learning (the weight reaches the maximum value 0.8 after only 20 epochs), resulting a worse performance.

3) Comparison With Nonlinearly Varying λ_t: Experiments with IDs 0, 9, and 10 compare the proposed linearly increasing weight λ_t with the nonlinearly (quadratically) increasing weight λ_t. It can be seen that the proposed linearly varying weight strategy clearly outperforms the quadratically varying weight. The reason for worse performance of the quadratically increasing weight λ_t is due to too slow or too fast increasing rate of the nonlinear weight at the beginning or ending epochs.

Therefore, we choose the linearly increasing weight strategy with $\Delta = 0.02$ and PT $= 0.8$ as the default settings.

C. Evaluation on the DiLiGenT Benchmark Dataset

The test results of various methods on the DiLiGenT benchmark dataset with 96 input images are listed in Table II, where we compare our DR-PSN with both traditional and learning-based methods in terms of achievable MAE (°). For traditional methods, we evaluate the low rank method [22] of outlier rejection-based techniques, and bivariate functions methods [26], [28] of sophisticated reflectance model-based techniques. For learning-based methods, we choose SDPS-Net [11], DPSN [10], IRPS [40], PS-FCN [13], and CNN-PS [12].

In practical applications, where sparse input images are common, it is difficult to obtain 96 densely input images. Therefore, we also compare our DR-PSN with both traditional methods and deep learning-based methods with only 10 input images in Table III to test the robustness of the DR-PSN under fewer input images. For traditional methods, we keep the three same methods [22], [26], [28] as in the case of 96 inputs. For deep learning-based methods, we keep the methods [12], [13] which can flexibly change the input images. Furthermore, we also add the two new methods, SPLINE-Net [38] and LMPS [37], which are designed for sparse inputs condition.

For the traditional methods, Matrix rank $= 3$ [22], Bivariate BRDF [26] and Bi-polynomial [28], we report the results from the original references. The deep learning-based methods, DPSN [10], SDPS-Net [11], PS-FCN [13] and LMPS [37], were all trained with the same MERL dataset described in Subsection V-A in their respective original papers. Therefore, we run these models on DiLiGenT benchmark test dataset, following the authors' original settings and implementations without any change. IRPS [40] is an unsupervised method and was trained using the un-rendered real image dataset without ground-truth normal. Therefore, for IRPS [40], we use the results reported in the original paper. For SPLINE-Net [38], we run the original model, which was trained by CyclePS

TABLE II

TEST PERFORMANCE COMPARISON OF VARIOUS METHODS ON DILIGENT BENCHMARK DATASET. ALL METHODS ARE EVALUATED WITH 96 IMAGES AND PERFORMANCE IS MEASURED BY MAE (°). SDPS-NET TAKES 96 IMAGES WITHOUT ILLUMINATION DIRECTIONS (UNCALIBRATED PHOTOMETRIC STEREO). CNN-PS* IS TRAINED WITH OUR MERL DATASET, WHILE CNN-PS IS THE ORIGINAL CNN-PS [12] TRAINED WITH DISNEY DATASET

Method	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Avg.
Baseline (Least squares) [1]	4.10	8.39	14.92	8.41	25.60	18.50	30.62	8.89	14.65	19.80	15.39
Matrix rank = 3 [22]	2.54	7.32	11.11	7.21	25.70	16.25	29.26	7.74	14.09	16.17	13.74
Bivariate BRDF [26]	3.34	7.11	10.47	6.74	13.05	9.71	25.95	6.64	8.77	14.19	10.60
Bi-polynomial [28]	1.74	6.12	10.60	6.12	13.93	10.09	25.44	6.51	8.78	13.63	10.30
SDPS-Net [11]	2.77	6.89	8.97	8.06	8.48	11.91	17.43	8.14	7.50	14.90	9.51
DPSN [10]	2.02	6.31	12.68	6.54	8.01	11.28	16.86	7.05	7.86	15.51	9.41
IRPS [40]	**1.47**	5.79	10.36	5.44	**6.32**	11.47	22.59	6.09	7.76	**11.03**	8.83
CNN-PS* [12]	2.23	8.29	8.53	5.75	9.74	8.66	17.75	5.91	8.16	11.61	8.66
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 [12]	2.12	12.30	8.07	**4.38**	7.92	**7.42**	**13.83**	**5.37**	**6.38**	12.12	7.99
DR-PSN (Ours)	2.27	**5.46**	**7.84**	5.42	7.01	8.49	15.40	7.08	7.21	12.74	**7.90**

TABLE III

TEST PERFORMANCE COMPARISON OF VARIOUS METHODS ON DILIGENT BENCHMARK DATASET. ALL METHODS ARE EVALUATED WITH 10 IMAGES AND PERFORMANCE IS MEASURED BY MAE (°). CNN-PS* IS TRAINED WITH OUR MERL DATASET, WHILE CNN-PS IS THE ORIGINAL CNN-PS [12] TRAINED WITH DISNEY DATASET

Method	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Avg.
Bivariate BRDF [26]	12.94	16.40	20.63	15.53	18.08	18.73	32.50	**6.28**	14.31	24.99	19.04
Baseline (Least squares) [1]	5.09	11.59	16.25	9.66	27.90	19.97	33.41	11.32	18.03	19.86	17.31
Bi-polynomial [28]	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 [22]	**3.33**	7.62	13.36	8.13	25.01	18.01	29.37	8.73	14.60	16.63	14.48
CNN-PS [12]	9.11	14.08	14.58	11.71	14.04	15.48	19.56	13.23	14.65	16.99	14.34
CNN-PS* [12]	6.39	14.51	15.08	10.96	15.26	14.40	19.73	11.35	13.58	16.67	13.79
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 (Ours)	3.83	7.52	**9.55**	7.92	9.83	**10.38**	**17.12**	9.36	**9.16**	14.75	**9.94**

TABLE IV

TEST METRICS OF OUR DR-PSN METHOD ON DILIGENT BENCHMARK DATASET WITH 96 AND 10 INPUT IMAGES, RESPECTIVELY. FOR THE CASE OF 96 INPUT IMAGES, ALL IMAGES ARE USED FOR EVALUATING THE SURFACE NORMAL, WHERE ALL RECONSTRUCTED IMAGES BELONG TO THE BI CLASS

Test images	Metrics	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	Avg.
96	MAE (°) ↓	2.27	5.46	7.84	5.42	7.01	8.49	15.55	7.08	7.21	12.74	7.91
	$< err_{15^\circ}$ ↑	100%	98.26%	90.05%	97.11%	95.83%	88.50%	66.18%	92.79%	93.74%	83.02%	90.55%
	SSIM(BI) ↑	0.939	0.944	0.957	0.958	0.937	0.952	0.938	0.970	0.960	0.924	0.948
	SSIM(NBI) ↑	-	-	-	-	-	-	-	-	-	-	-
	REL(BI) ↓	0.070	0.244	0.133	0.067	0.255	0.154	0.256	0.141	0.167	0.185	0.167
	REL(NBI) ↓	-	-	-	-	-	-	-	-	-	-	-
10	MAE (°) ↓	3.83	7.52	9.55	7.92	9.83	10.38	17.12	9.36	9.16	14.75	9.94
	$< err_{15^\circ}$ ↑	100%	96.32%	85.16%	91.61%	83.72%	82.60%	62.65%	86.54%	88.87%	66.81%	88.43%
	SSIM(BI) ↑	0.943	0.939	0.951	0.974	0.945	0.948	0.930	0.962	0.931	0.917	0.944
	SSIM(NBI) ↑	0.960	0.933	0.964	0.966	0.940	0.955	0.931	0.969	0.936	0.914	0.947
	REL(BI) ↓	0.064	0.227	0.143	0.064	0.257	0.155	0.271	0.129	0.164	0.186	0.166
	REL(NBI) ↓	0.061	0.241	0.140	0.063	0.262	0.158	0.266	0.131	0.155	0.174	0.165

dataset with Disney's principled BSDFs [56], on the test dataset.[1]

Note that CNN-PS [12] was trained by CyclePS dataset with Disney's principled BSDFs [56]. The other deep learning-based methods are all trained with our MERL dataset, except for SPLINE-Net [38]. For a fair comparison, in addition to keep the original CNN-PS [12] trained by Disney dataset, we also train the CNN-PS with our dataset (the blobby shape dataset [53] and sculpture dataset [54] with MERL reflectance dataset [55]), marked it as CNN-PS*. We only set one epoch to train CNN-PS* because the number of training samples (84360) in our dataset is much larger than the original dataset (45 samples, default epochs 10). Due to the large number of samples in our dataset, it takes approximately 220 hours for training one epoch on an NVIDIA TITAN XP GPU.

Next Table IV lists all the test metrics of our model on the DiLiGenT benchmark dataset with 96 and 10 input images, respectively.

We also compare the visual results of our method with those of several state-of-the-art learning-based methods in Figs. 5 and 6 for the 96 and 10 input images, respectively. Then, we depict the visual results of our method using 96, 48, and 10 input images, respectively, in Fig. 7. Furthermore, Fig. 8 investigates the performance differences between BI

[1]We also attempted to train the SPLINE-Net model with our MERL dataset using the authors' code for a fair comparison with other methods. However, the code appears to exist some errors and we could not run it.

Fig. 5. Qualitative results for real-world scenes from DiLiGenT benchmark dataset with 96 input images. The black numbers under the error maps represent MAE(°). The reconstruction errors are shown with intensity ×10 for better details. The contrast of the observation images and reconstructed images are also adjusted for easy viewing in the same parameter (50% enhancement). The red boxes in the error maps are the regions with complex structures, while the orange boxes in the error maps are the regions with strong shadow and inter-reflections. Our method produces more robust estimations in these regions.

reconstructed images and NBI reconstructed images obtained by our DR-PSN and tested with 48 images.

1) Discussion on Estimated Surface Normal (With 96 Input Images): Tables II compares the surface normal estimation results of our DR-PSN and several existing state-of-the-art photometric stereo methods on the DiLiGenT benchmark, with 96 inputs. It can be seen that our DR-PSN ranks the top with an average MAE of 7.90° (test with 96 images). The CNN-PS

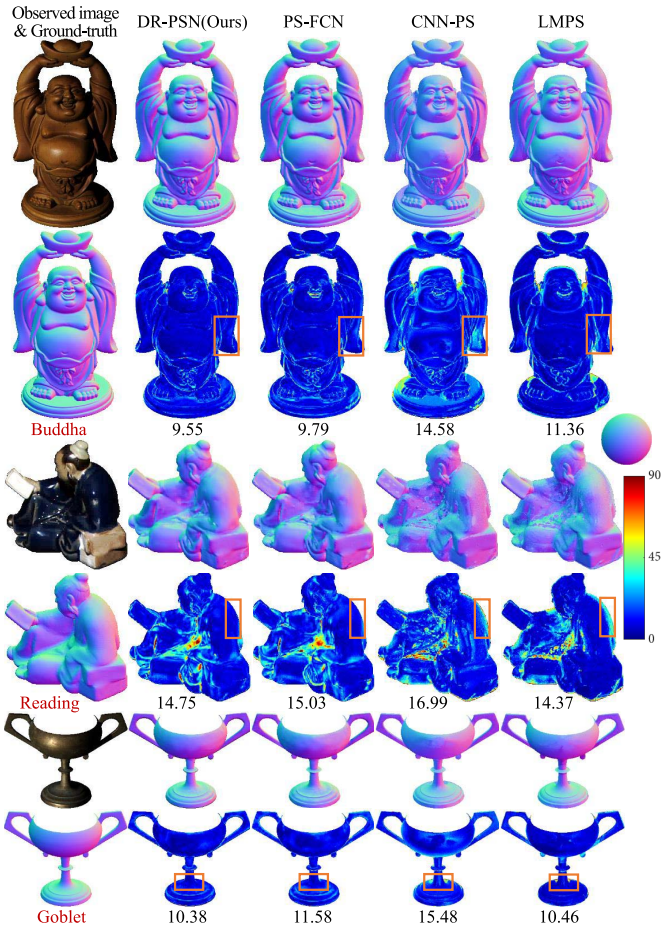


Fig. 6. Qualitative results for real-world scenes from the DiLiGenT benchmark dataset with 10 sparse input images. The black numbers under the error maps represent MAE(°). The contrast of the observation images are adjusted for easy viewing in the same parameter (50% enhancement). The orange boxes in the error maps are the regions with cast shadows. Our method produces more robust estimations in these regions.

[12], trained by Disney dataset, ranks the second best on average and its performance is very close to our DR-PSN, which is trained by the MERL dataset. Also, the performance of CNN-PS is better than that of CNN-PS* because the advanced training set rendered with Disney's principled BSDFs [56]. We note that the author of CNN-PS [12] discarded the first 20 images of "Bear" in the paper to record a much better MAE of 4.25° for "Bear". The reason for discarding some test images given by the author [12] is that the intensity values around the stomach region of the "Bear" are wrong in the first 20 images. For fair comparison, we however input all the 96 images of "Bear" when evaluating all the methods on "Bear". It can be seen that the MAE of CNN-PS [12] on "Bear" dramatically degrades to 12.30° when inputting all the 96 images. It suggests that CNN-PS uses a per-pixel manner for the observation map, which may be not robust to errors in input images. In fact, we note that the results of the other methods for "Bear" reported in Table II with all 96 input images are only slightly worse than the corresponding results by discarding the first 20 images. For space-saving purpose, we omit the results of discarding first 20 images.

We show the visual comparison of several state-of-the-art methods on "Buddha", "Bear", "Pot2" and "Harvest" under 96 input images in Fig. 5, where we mark some boxes. The red boxes represent the complex-structured regions, such as the mouth of "Buddha" and the flower of "Pot2". It can be seen that the error maps of our method show a lower angular error in these regions. Compared with other state-of-the-art methods, our DR-PSN also produces more details in the areas with complex structures. This indicates that our DR-PSN is more robust and accurate in these challenging surfaces. The reason is that our DR-PSN forms a closed-loop architecture to provide additional supervision on surface normal estimation. The extra dual regression network learns the inverse process to reduce the difficulty in the normal regression task.

2) Discussion on Robustness With Fewer Input Images: Many practical applications involves sparse photometric stereo. We therefore evaluate our DR-PSN with 10 sparse inputs. We emphasize that the DR-PSN applies the channel max-pooling [13] in the normal regression task. The max-pooling operation can extract the arbitrary number of features from images captured under different illumination directions and, moreover, it naturally aggregates the strongest response of features while ignoring non-activated shadows. Therefore, our DR-PSN can be flexibly used for an arbitrary number of input images.

In Table III, we compare the surface normal results of our DR-PSN and other state-of-the-art methods on the DiLiGenT benchmark with 10 input images, where it can be clearly seen that our method achieves the best performance with LMPS [37] as the close second best. Furthermore, we show the visualized results in Fig. 6. The orange boxes reveal the regions with cast shadows, such as the cuff of "Buddha", the back of "Reading", and the base of "Goblet". In these areas with cast shadows, the proposed DR-PSN outperforms other methods, as shown in the corresponding error maps.

Table IV lists all the metrics obtained by our DR-PSN for the DiLiGenT benchmark dataset with 96 and 10 input images, respectively. Observe that the average SSIM and REL attained are 0.948 and 0.167 when testing with 96 images, while the average SSIM and REL attained are 0.944 and 0.166 when testing with 10 images. It can be seen that the accuracy of reconstructed images by the DR-PSN is fairly robust with fewer input images. Furthermore, Fig. 7 shows the visual comparison for three examples, "Cat", "Reading" and "Cow", obtained by our method when testing with 96 images, 48 images and 10 images, respectively. Observe that the reconstructed images hardly show any visual differences when testing with different numbers of input images. The yellow boxes in Fig. 7 again represent regions with varying surface colors and detailed structures. Our method produces robust reconstructions of these areas.

3) Discussion on Performance of Arbitrary Reconstructed Images: In Table IV, BI represents the reconstructed images belonging to the illumination directions of input images, while NBI represents the reconstructed images not belonging to the illumination directions of input images. Hence the experiment with 10 input images shows that the accuracy of reconstructed

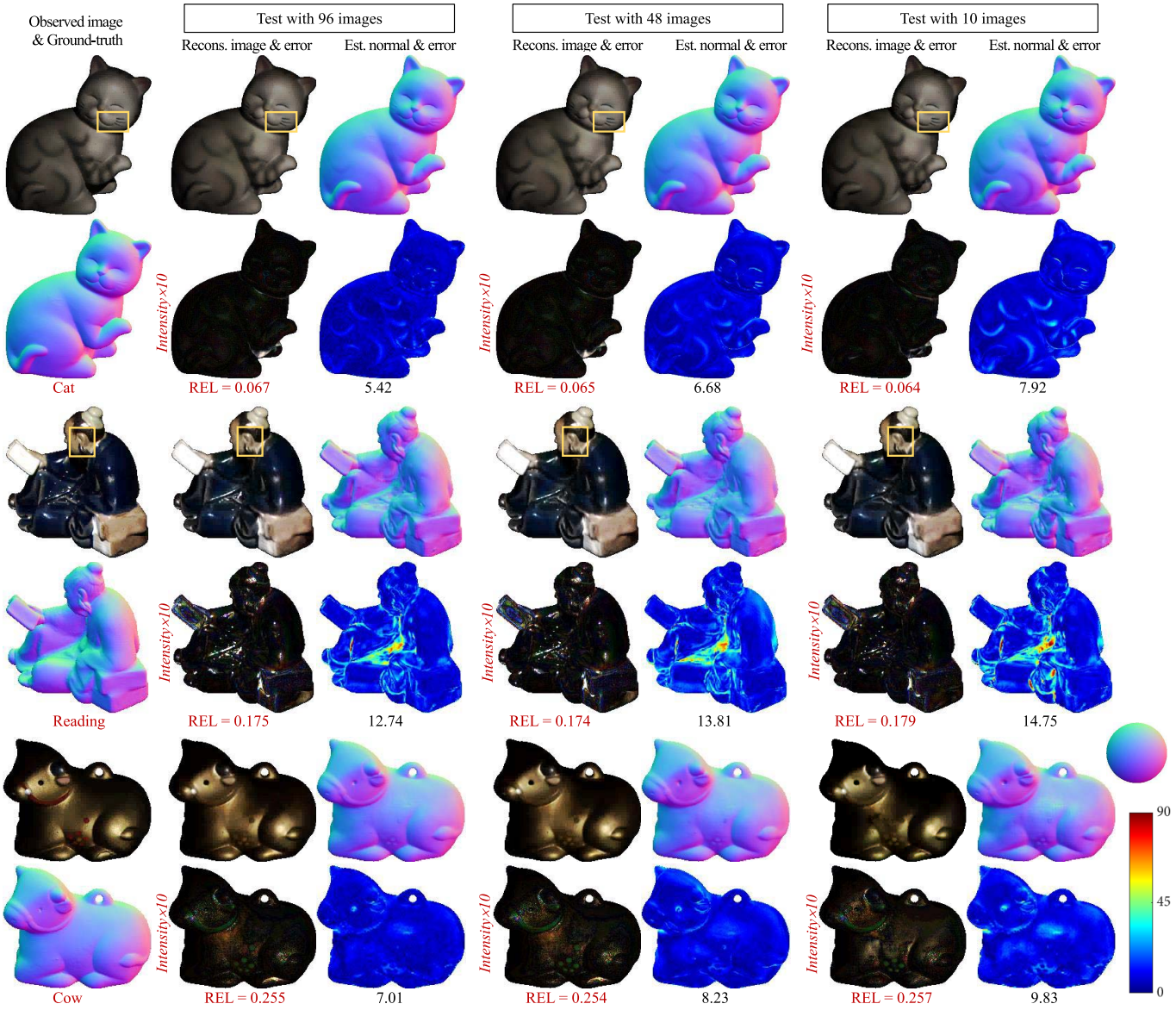


Fig. 7. Qualitative test results of our DR-PSN for real-world scenes from the DiLiGenT benchmark dataset with 96, 48, and 10 input images, respectively. The black numbers under the error maps represent MAE(°). The reconstruction errors are shown with intensity ×10 for better details. The contrast of the observation images and reconstructed images are also adjusted for easy viewing in the same parameter (50% enhancement). The yellow boxes are the regions with varying surface colors and detailed structures. Our DR-PSN achieves accurate reconstructions.

images are almost the same, whether they belong to the illumination directions of input images or not.

Fig. 8 gives a visual example on object "Goblet", which are tested with 48 input images. Specifically, in the test, we choose the odd IDs (BI group) of 96 images as the inputs. Thus, the even IDs (NBI group), $2, 4, \cdots, 96$, are not in inputs. We show some visual results of the BI group and the NBI group, respectively, both groups having the same number of images depicted. The positions of specularities and shadow are accurately estimated in both the BI group and the NBI group. This illustrates that the encoded illumination information is well-utilized and in the dual regression task. Therefore, our DR-PSN can accurately generate specified reconstructed images under arbitrary illumination directions.

4) Discussion on Limitations of Proposed Method: From Table II, although our method attains the best performance on average, it only obtains the second best performance on five objects ("Cat", "Cow", "Goblet", "Harvest" and "Pot2"), evaluated with 96 input images. We infer that the max-pooling [13] adopted by our method discards a large amount of the features from the inputs and only remains the maximum response value. Therefore, the utilization of our method reduces when the input images increase, which may impact on its performance to some extent. Also our method only achieves the fifth and third best performance on "Ball" and "Reading", respectively, among the 10 methods. First, for object "Ball" with a particularly simple structure and almost Lambertian surface, our method does not outperform the traditional and single-supervision learning-based methods. In this extreme

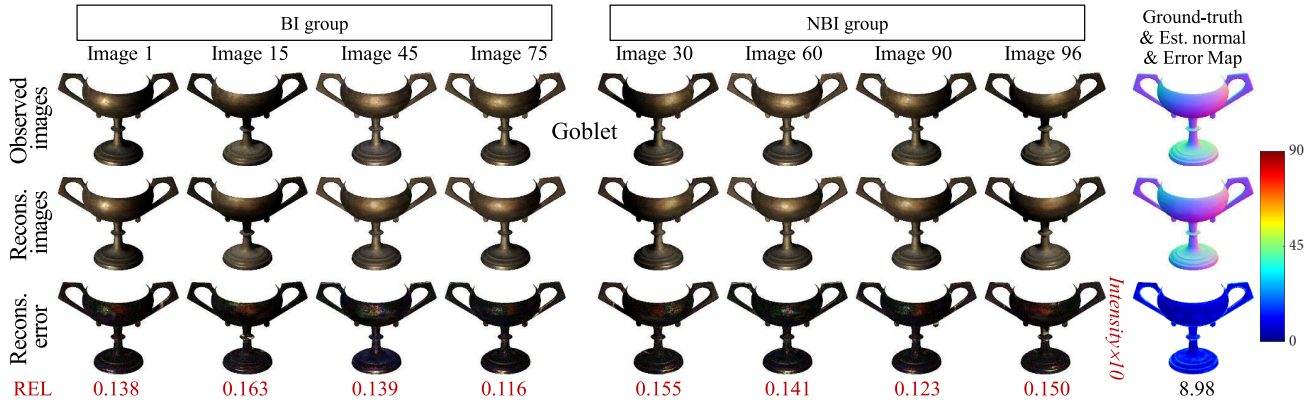


Fig. 8. Visual results of the object "Goblet" by our DR-PSN. We test it with 48 input images, where we choose the odd ID of images as inputs. Therefore, the reconstructed images 1, 15, 45, 75 belong to the illumination directions of input images (BI group), while the reconstructed images 30, 60, 90, 96 do not belong to the illumination directions of input images (NBI group). The black numbers under error maps represent MAE in degree. The reconstruction errors are shown with intensity $\times 10$ for better details. The contrast of the observation images and reconstructed images are also adjusted for easy viewing in the same parameter (50% enhancement).

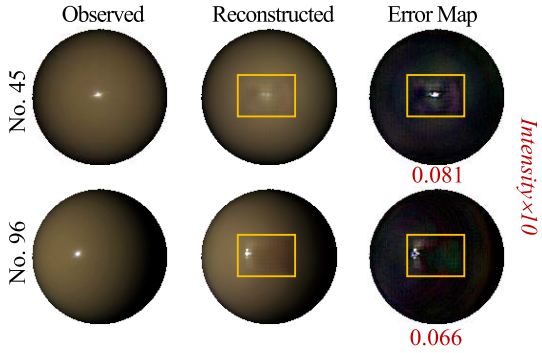


Fig. 9. Reconstruction results of the object "Ball" by our DR-PSN. We test it with 96 input images. The red numbers under error maps represent REL. The reconstruction error maps are shown with intensity $\times 10$ for better details. The contrast of the observation images and reconstructed images are also adjusted for easy viewing in the same parameter (50% enhancement). The yellow boxes stand for the regions with aggregated specularities, and errors, respectively.

case, the normal loss provides a strong penalty, while our dual regression task (reconstruction loss) may weaken this constraint. Second, for object "Reading", which has strong specularities, our DR-PSN also performs poorer than IRPS and CNN-PS. In this case, we can see that the reconstructed images have major errors in the specularities regions. The error on the dual regression task may impact on the accuracy of the normal regression task, as it is a closed-loop process.

We also notice that the errors of reconstructed images mainly exist in specularities regions, such as the crinkles of object "Harvest" and the middle of object "Reading" (see Figs. 5 and 7, respectively). The reason may be the use of max-pooling operation in the normal regression task. In fact, max-pooling is used for handling arbitrary number of inputs and aggregating features from multiple inputs, which extracts the most salient information from all the features [13]. However, the most salient information always includes specularities. In our dual regression task, the shortcut after max-pooling provides the reflectance feature for imaging the reconstructed images. Unfortunately, it also brings specularities information aggregated from all inputs, which may cause errors in the reconstructed images. We show an obvious example of "Ball", which has evenly distributed specularities under the different input images, in Fig. 9. It can be seen that almost all the specularities existed in the inputs are aggregated in the reconstructed images (yellow boxes) due to the max-pooling fusion operation, causing errors.

D. Evaluation on the Light Stage Data Gallery

We also evaluate our DR-PSN on the Light Stage Data Gallery [57]. The resolutions of images in the Light Stage Data Gallery are much larger than those in the DiLiGenT benchmark dataset. Owing to the memory limit of GPU, we test the Light Stage Data Gallery with 72 input images. Fig. 10 shows the results obtained. Due to the absence of ground-truth of the surface normal in this dataset, we qualitatively show the normal estimation. Note that the reconstructed images still have the ground-truth, and hence we can quantitatively evaluate reconstructed images.

1) Discussion on Estimated Surface Normal: As shown in Fig. 10, the estimated surface normal can accurately report the shapes of the objects. The red boxes show the fiber skirt and hands of "Fighting", the belt of "Standing" as well as the branches and leaves of "Plant". It can be seen that the estimated normals reveal the details in these regions without blur. The shape of fingers can be distinguished in "Fighting", and even the rugged texture of the clothes can be observed in the estimated surface normal. These examples illustrate the effectiveness of our DR-PSN. We also observe that some noise exists on the surface normal in some place where it should be smooth, such as the armor of "Standing". This may be caused by the noise in the observed images, because of the low quality of photoed images.

2) Discussion on Reconstructed Images: Fig. 10 also reports the results of reconstructed images. We observe that the REL values of the reconstructed images are worse than those obtained for the DiLiGenT benchmark dataset. It may be due to the high-frequency noise in the observed images

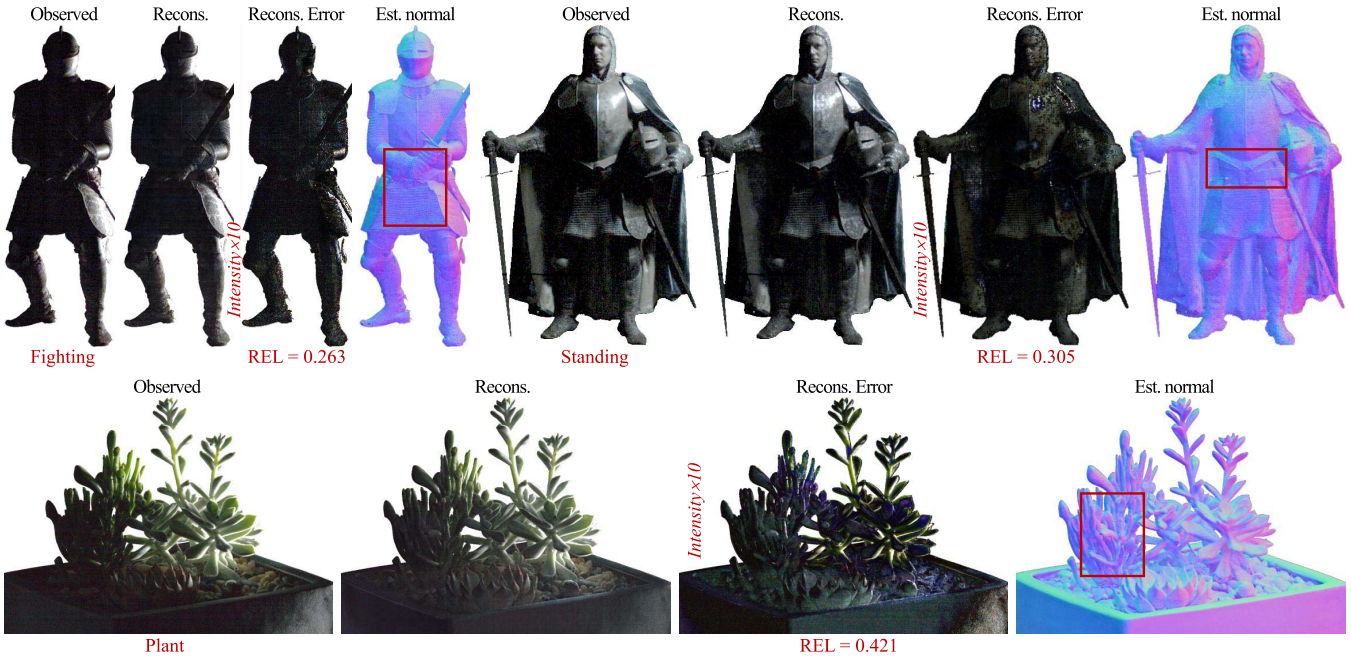


Fig. 10. Evaluation on the Light Stage Data Gallery with 72 input images using our DR-PSN. We qualitatively show the estimated surface normal and quantitatively evaluate the reconstructed image. The red boxes report the details of the estimated surface normal. The reconstruction errors are shown with intensity ×10 for better details. The contrast of observation images and reconstructed images are also adjusted for easy viewing in the same parameter (50% enhancement for "Fighting" and "Standing", 30% for "Plant").

TABLE V

DETAILED ARCHITECTURE AND PARAMETERS OF DR-PSN: 'S' AFTER "CONV" REPRESENTS THE STRIDE OF THE CONVOLUTIONAL LAYER, 'L_RELU' IS THE SHORT FORM FOR 'LEAKY_RELU', AND 'MAX-P' DENOTES THE MAX-POOLING OPERATION. WE SEPARATE THE REGRESSOR AND EXTRACTOR WITH A LINE IN EACH TASK. ALSO, THE RED ARROW REVEALS THE FUSION OF REFLECTANCE FEATURE AND NORMAL FEATURE

Details of the Normal regression task

Act	Output Shape
3×3 conv, S = 1 L_Relu	H×W×64
3×3 conv, S = 1 L_Relu	H×W×64
3×3 conv, S = 1 L_Relu	H×W×64
Skip Connection	H×W×64
3×3 conv, S = 1 L_Relu	H×W×64
3×3 conv, S = 1 L_Relu	H×W×64
Skip Connection	H×W×64
3×3 conv, S = 2 L_Relu	1/2H×1/2W×128
3×3 conv, S = 1 L_Relu	1/2H×1/2W×128
Skip Connection	1/2H×1/2W×128
3×3 conv, S = 1 L_Relu	1/2H×1/2W×128
3×3 conv, S = 1 L_Relu	1/2H×1/2W×128
Skip Connection	1/2H×1/2W×128
3×3 conv, S = 2 L_Relu	1/4H×1/4W×256
3×3 conv, S = 1 L_Relu	1/4H×1/4W×256
Skip Connection	1/4H×1/4W×256
3×3 conv, S = 1 L_Relu	1/4H×1/4W×256
3×3 conv, S = 1 L_Relu	1/4H×1/4W×256
Skip Connection	1/4H×1/4W×256
Max-p	1/4H×1/4W×256
3×3 conv, S = 1 L_Relu	1/4H×1/4W×256
3×3 deconv L_Relu	1/2H×1/2W×128
3×3 conv, S = 1 L_Relu	1/2H×1/2W×128
3×3 deconv L_Relu	H×W×64
3×3 conv, S = 1 L_Relu	H×W×3
L2_ Norm	H×W×3

Details of the Dual regression task

Act	Output Shape
3×3 conv, S = 2 L_Relu	1/2H×1/2W×128
3×3 deconv L_Relu	H×W×64
3×3 conv, S = 2 L_Relu	1/2H×1/2W×128
3×3 conv, S = 2 L_Relu	1/4H×1/4W×256
3×3 deconv L_Relu	1/2H×1/2W×128
3×3 conv, S = 2 L_Relu	1/4H×1/4W×256
3×3 conv, S = 1 L_Relu	1/4H×1/4W×256
Concatenation (from Max-p)	1/4H×1/4W×512 (256+256)
Concatenation (from L'_n)	1/4H×1/4W×515 (512+3)
3×3 deconv L_Relu	1/2H×1/2W×256
3×3 deconv L_Relu	H×W×128
3×3 conv, S = 2 L_Relu	1/2H×1/2W×256
Concatenation (from L''_n)	1/2H×1/2W×259 (256+3)
3×3 deconv L_Relu	H×W×128
3×3 conv, S = 1 L_Relu	H×W×3

impacting on the performance. Clearly, the noisy input images affect the accuracy of reconstructed images. Nevertheless, our method accurately generates the positions of specularities and shadow, e.g., in "Fighting" and "Standing". We also note that the performance of "Plant" is worse than others. Our analysis indicates that this is because the reflectance of real plants (leaves) are rarely observed in our training dataset.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a dual regression network, called DR-PSN, to reconstruct both surface normal and image. Our main contribution has been exploring reconstructed images to further promote the accuracy of recovering the surface normal. This has been achieved by introducing additional constraints on observed images and reconstructed images to form a closed-loop for providing additional supervision. Moreover, our method can generate accurate reconstructed images under arbitrary illumination directions to intuitively show the texture information and anisotropic reflectance properties of the surface. Extensive quantitative comparisons on the most widely used DiLiGenT benchmark dataset have shown that our DR-PSN outperforms state-of-the-art calibrated photometric stereo methods, including traditional algorithms and learning-based methods. Specifically, the experimental results have demonstrated that the estimated surface normal obtained by our DR-PSN is significantly more accurate than those obtained by traditional photometric stereo algorithms and state-of-the-art learning-based approaches. In particular, our method has been shown to better handle the complex-structured and strong shadow regions, and to be capable of generating accurate reconstructed images under arbitrarily specified illumination directions. Additional qualitative experiment on the Light Stage Data Gallery has further confirmed the effectiveness of our proposed dual regression network.

Our method can benefit the photometric stereo community in the following two ways: 1. our work generates the specified images in addition to surface normals, which provides the example of combining multiple tasks in 3D recovery, and 2. the extra task, reconstructing the images, is proved to be beneficial to the estimation of surface normals, which inspires future work to find more auxiliary supervisions to further improve the accuracy of surface normal estimation, rather than blindly increasing the complexity of network architecture. In fact, our work unifies the 3D reconstruction and rendering tasks in one, which has potential for a wider range of applications.

Despite of offering the state-of-the-art performance, our method can be further improved. The training set, which is rendered using the MERL dataset [55], hardly spans the whole set of materials existing in nature. To further improve the accuracy, employing a larger material dataset to cover the tremendous real-world materials is necessary, such as Disney's principled BSDFs dataset [56].

There are several promising ways to extend our work. First, we will explore the reconstruction of arbitrary material properties as the objects. It can be seen that the regression task approximately learns the imaging model, and in this way, the material can be used to render another object while keeping the same material properties of the original object. This can be used to render meshes with realistic appearances. Second, we will extend our DR-PSN to the uncalibrated photometric stereo, which will benefit wider practical applications. We will design an illumination direction prediction network, which has already been investigated in some deep learning-based uncalibrated photometric stereo methods [11], [58], to estimate the lights from the input image, instead of inputting the calibrated illumination directions.

ACKNOWLEDGMENT

The authors would like to thank Guanying Chen for help in code and Hiroaki Santo for help in providing comparison results. The authors' gratitude also goes to the anonymous reviewers for their careful suggestions and enlightenment that have helped the authors to improve this article substantially.

APPENDIX
DETAILED ARCHITECTURE AND PARAMETERS OF DR-PSN

Table V details the architecture and parameters of our DR-PSN. The left part of Table provides the detailed architecture and parameters of the normal regression task, and the right part of Table provides the detailed architecture and parameters of the dual regression task. It can be seen that the architecture of regressors (below the lines in Table V) experience interleaved down-sampling and up-sampling. Note that the down-sampling and the up-sampling are implemented by deconvolution layer and stride $= 2$ convolution layer, respectively. This structure can increase the receptive field and preserve spatial information with a smaller memory consumption [13].

REFERENCES

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

[2] D. Miyazaki, K. Hara, and K. Ikeuchi, "Median photometric stereo as applied to the segonko tumulus and museum objects," *Int. J. Comput. Vis.*, vol. 86, nos. 2–3, pp. 229–242, Jan. 2010.

[3] T.-P. Wu and C.-K. Tang, "Photometric stereo via expectation maximization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 546–560, Mar. 2010.

[4] K. Sunkavalli, T. Zickler, and H. Pfister, "Visibility subspaces: Uncalibrated photometric stereo with shadows," in *Proc. ECCV*, Heraklion, Greece, Sep. 2010, pp. 251–264.

[5] D. B. Goldman, B. Curless, A. Hertzmann, and S. M. Seitz, "Shape and spatially-varying BRDFs from photometric stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 6, pp. 1060–1071, Jun. 2010.

[6] L. Chen, Y. Zheng, B. Shi, A. Subpa-Asa, and I. Sato, "A microfacet-based reflectance model for photometric stereo with highly specular surfaces," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 3162–3170.

[7] S.-K. Yeung, T.-P. Wu, C.-K. Tang, T. F. Chan, and S. J. Osher, "Normal estimation of a transparent object using a video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 890–897, Apr. 2015.

[8] Y. Wang, J. Liang, D. Cao, and Z. Sun, "Local semantic-aware deep hashing with Hamming-isometric quantization," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2665–2679, Jun. 2019.

[9] K. Wei, M. Yang, H. Wang, C. Deng, and X. Liu, "Adversarial fine-grained composition learning for unseen attribute-object recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 3741–3749.

[10] H. Santo, M. Samejima, Y. Sugano, B. Shi, and Y. Matsushita, "Deep photometric stereo network," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Venice, Italy, Oct. 2017, pp. 501–509.

[11] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y.-K. Wong, "Self-calibrating deep photometric stereo networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 8739–8747.

[12] S. Ikehata, "CNN-PS: CNN-based photometric stereo for general non-convex surfaces," in *Proc. ECCV*, Munich, Germany, Sep. 2018, pp. 3–19.

[13] G. Chen, K. Han, and K.-Y. K. Wong, "PS-FCN: A flexible learning framework for photometric stereo," in *Proc. ECCV*, Munich, Germany, Sep. 2018, pp. 3–19.

[14] Y. Ju, K.-M. Lam, Y. Chen, L. Qi, and J. Dong, "Pay attention to devils: A photometric stereo network for better details," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 694–700.

[15] F. Logothetis, I. Budvytis, R. Mecca, and R. Cipolla, "A CNN based approach for the near-field photometric stereo problem," in *Proc. BMVC*, Sep. 2020, pp. 1–12.

[16] H. Santo, M. Waechter, and Y. Matsushita, "Deep near-light photometric stereo for spatially varying reflectances," in *Proc. ECCV*, Aug. 2020, pp. 1–16.

[17] F. Solomon and K. Ikeuchi, "Extracting the shape and roughness of specular lobe objects using four light photometric stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 4, pp. 449–454, Apr. 1996.

[18] S. Barsky and M. Petrou, "The 4-source photometric stereo technique for three-dimensional surfaces in the presence of highlights and shadows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1239–1252, Oct. 2003.

[19] B. Shi, Z. Mo, Z. Wu, D. Duan, S.-K. Yeung, and P. Tan, "A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 271–284, Feb. 2019.

[20] F. Verbiest and L. Van Gool, "Photometric stereo with coherent outlier handling and confidence estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, Jun. 2008, pp. 1–8.

[21] H. Fan, Y. Luo, L. Qi, N. Wang, J. Dong, and H. Yu, "Robust photometric stereo in a scattering medium via low-rank matrix completion and recovery," in *Proc. 9th Int. Conf. Hum. Syst. Interact. (HSI)*, Queenstown, New Zealand, Jul. 2016, pp. 703–717.

[22] S. Ikehata, D. Wipf, Y. Matsushita, and K. Aizawa, "Robust photometric stereo using sparse regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 318–325.

[23] M. Chandraker, S. Agarwal, and D. Kriegman, "ShadowCuts: Photometric stereo with shadows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.

[24] Y. Queau, T. Wu, F. Lauze, J.-D. Durou, and D. Cremers, "A non-convex variational approach to photometric stereo under inaccurate lighting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 99–108.

[25] N. Alldrin, T. Zickler, and D. Kriegman, "Photometric stereo with non-parametric and spatially-varying reflectance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, Jun. 2008, pp. 1–8.

[26] S. Ikehata and K. Aizawa, "Photometric stereo using constrained bivariate regression for general isotropic surfaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 2179–2186.

[27] T. Higo, Y. Matsushita, and K. Ikeuchi, "Consensus photometric stereo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 1157–1164.

[28] B. Shi, P. Tan, Y. Matsushita, and K. Ikeuchi, "Bi-polynomial modeling of low-frequency reflectances," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1078–1091, Jun. 2014.

[29] H.-S. Chung and J. Jia, "Efficient photometric stereo on glossy surfaces with wide specular lobes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Anchorage, AK, USA, Jun. 2008, pp. 1–8.

[30] T. Chen, M. Goesele, and H.-P. Seidel, "Mesostructure from specularity," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2006, pp. 1825–1832.

[31] S. Tozza, R. Mecca, M. Duocastella, and A. Del Bue, "Direct differential photometric stereo shape recovery of diffuse and specular surfaces," *J. Math. Imag. Vis.*, vol. 56, no. 1, pp. 57–76, Sep. 2016.

[32] A. S. Georghiades, "Incorporating the Torrance and Sparrow model of reflectance in uncalibrated photometric stereo," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Nice, France, Oct. 2003, pp. 1–8.

[33] L. Chen, Y. Zheng, B. Shi, A. Subpa-Asa, and I. Sato, "A microfacet-based model for photometric stereo with general isotropic reflectance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 48–61, Jan. 2021.

[34] A. Hertzmann and S. M. Seitz, "Example-based photometric stereo: Shape reconstruction with general, varying BRDFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1254–1264, Aug. 2005.

[35] F. Logothetis, I. Budvytis, R. Mecca, and R. Cipolla, "PX-NET: Simple, efficient pixel-wise training of photometric stereo networks," 2020, *arXiv:2008.04933*. [Online]. Available: http://arxiv.org/abs/2008.04933

[36] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[37] J. Li, A. Robles-Kelly, S. You, and Y. Matsushita, "Learning to minify photometric stereo," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 7568–7576.

[38] Q. Zheng, Y. Jia, B. Shi, X. Jiang, L. Duan, and A. Kot, "SPLINE-Net: Sparse photometric stereo through lighting interpolation and normal estimation networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 8549–8558.

[39] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y.-K. Wong, "Deep photometric stereo for non-Lambertian surfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 29, 2020, doi: 10.1109/TPAMI.2020.3005397.

[40] T. Taniai and T. Maehara, "Neural inverse rendering for general reflectance photometric stereo," in *Proc. ICML*, Stockholm, Sweden, Jul. 2018, pp. 4857–4866.

[41] Y. Yu and W. A. P. Smith, "InverseRenderNet: Learning single image inverse rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 3155–3164.

[42] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs, "SfSNet: Learning shape, reflectance and illuminance of faces 'in the wild,'" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 6296–6305.

[43] S. Wu, C. Rupprecht, and A. Vedaldi, "Unsupervised learning of probably symmetric deformable 3D objects from images in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1–10.

[44] Y. Xia *et al.*, "Dual supervised learning," in *Proc. ICML*, Sydney, NSW, Australia, Aug. 2017, pp. 3789–3798.

[45] Y. Xia *et al.*, "Model-level dual learning," in *Proc. ICML*, Stockholm, Sweden, Jul. 2018, pp. 5383–5392.

[46] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2849–2857.

[47] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 2223–2232.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, San Diego, CA, USA, May 2015, pp. 1–14.

[50] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NIPS*, Vancouver, BC, Canada, Dec. 2019, pp. 1–12.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, San Diego, CA, USA, May 2015, pp. 1–15.

[52] 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.

[53] M. K. Johnson and E. H. Adelson, "Shape estimation in natural illumination," in *Proc. CVPR*, Providence, RI, USA, Jun. 2011, pp. 2553–2560.

[54] O. Wiles and A. Zisserman, "SilNet: Single- and multi-view reconstruction by learning from silhouettes," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, London, U.K., Sep. 2017, pp. 1–13.

[55] W. Matusik, H. Pfister, M. Brand, and L. Mcmillan, "A data-driven reflectance model," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 759–769, Jul. 2003.

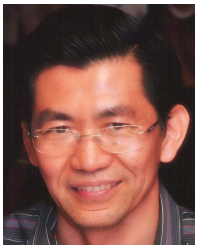
[56] B. Brent, "Physically-based shading at Disney," in *Proc. SIGGRAPH Course, Practical Physically Based Shading Film Game Prod.*, 2012, pp. 1–7.

[57] C.-F. Chabert *et al.*, "Relighting human locomotion with flowed reflectance fields," in *Proc. ACM SIGGRAPH Sketches SIGGRAPH*, Nicosia, Cyprus, Jun. 2006, pp. 183–194.

[58] G. Chen *et al.*, "What is learned in deep uncalibrated photometric stereo?" in *Proc. ECCV*, Aug. 2020, pp. 1–17.

Yakun Ju (Graduate Student Member, IEEE) 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 Prof. J. Dong. His research interests include 3D reconstruction, deep learning, and image processing.

Junyu Dong (Member, IEEE) 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. In 2004, he joined the Ocean University of China, where he is currently a Professor and the Vice-Dean of the College of Information Science and Engineering. 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.



Sheng Chen (Fellow, IEEE) received the B.Eng. degree in control engineering from the East China Petroleum Institute, Dongying, China, in 1982, the Ph.D. degree in control engineering from the City, University of London, U.K., in 1986, and the D.Sc. degree from the University of Southampton, Southampton, U.K., in 2005. From 1986 to 1999, he held research and academic appointments at the universities of Sheffield, Edinburgh, and Portsmouth, U.K. Since 1999, he has been with the School of Electronics and Computer Science, University of Southampton, U.K., where he is currently a Professor in intelligent systems and signal processing. He is also a Chief Scientist at the Center on Artificial Intelligence, Ocean University of China. His research interests include neural network and machine learning, adaptive signal processing, and wireless communications, and nonlinear system modeling. He has published over 700 research articles. He has more than 15 400 Web of Science citations with an H-index 54, and more than 31 200 Google Scholar citations with an H-index 75. He is a Fellow of the U.K. Royal Academy of Engineering and the IET, and an original ISI Highly Cited Researcher in engineering (March 2004).