Deep near-light photometric stereo for spatially varying reflectances

Hiroaki Santo[0000−0003−2891−5993], Michael Waechter[0000−0002−1403−1209], and Yasuyuki Matsushita[0000−0002−1935−4752]

Graduate School of Information Science and Technology,
Osaka University, Osaka, Japan
{santo.hiroaki, waechter.michael, yasumat}@ist.osaka-u.ac.jp

Abstract. This paper presents a near-light photometric stereo method for spatially varying reflectances. Recent studies in photometric stereo proposed learning-based approaches to handle diverse real-world reflectances and achieve high accuracy compared to conventional methods. However, they assume distant (*i.e.*, parallel) lights, which can in practical settings only be approximately realized, and they fail in near-light conditions. Near-light photometric stereo methods address near-light conditions but previous works are limited to over-simplified reflectances, such as Lambertian reflectance. The proposed method takes a hybrid approach of distant- and near-light models, where the surface normal of a small area (corresponding to a pixel) is computed locally with a distant light assumption, and the reconstruction error is assessed based on a near-light image formation model. This paper is the first work to solve unknown, spatially varying, diverse reflectances in near-light photometric stereo.

Keywords: Near-light photometric stereo

1 Introduction

Photometric stereo estimates surface normals of a scene from multiple images captured by a fixed camera under varying light conditions. The basic idea of photometric stereo was introduced in 1980 by Woodham [34] assuming Lambertian reflectance under distant light. In practice, these assumptions typically do not hold; therefore, a photometric stereo method that can deal with *diverse and spatially varying reflectances* in a *nearby light* setting is wanted.

Recent studies have shown that a deep learning-based approach [26, 6, 11] can effectively deal with diverse and spatially varying reflectances by establishing a mapping from observed images to a surface normal map. These methods assume a distant light setting for ease of learning. In a different thread, nearby light photometric stereo has been studied [12, 17, 22] to explicitly eliminate the distant light assumption. These works have shown to be effective for Lambertian or simple parametric reflectances, but still suffer from diverse and spatially varying reflectances. Since these studies for relaxing the Lambertian reflectance

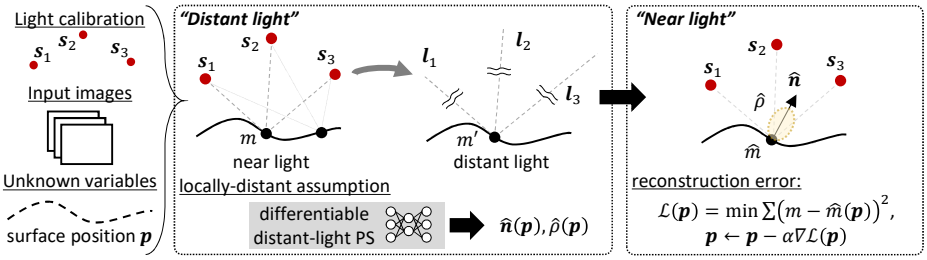


Fig. 1. Overview of the proposed method. The inputs are the given light calibration and observations m, and the unknown is the surface position $\mathbf{p}$. Based on assuming light to be distant (*i.e.*, parallel) locally, we employ a near-light effect cancellation (Eqs. (1) and (4)) to create a pseudo (distant-light) observation m', and compute the surface normal $\hat{\mathbf{n}}$ and reflectance $\hat{\rho}$ using distant-light photometric stereo. We then assess the reconstruction of the observation $\hat{m}$ based on a near-light image formation model.

and distant light assumptions have been developed rather independently, it is still unclear how these two distinct studies can benefit from each other.

In this work, we present a hybrid approach of distant- and near-light models for simultaneously removing the assumptions of both Lambertian reflectance and distant lighting. Specifically, we assume that a single pixel covers a small surface area within which incoming light emitted from a nearby light source can be modeled as distant (*i.e.*, parallel), although different pixels may be illuminated by different light directions and strengths. Based on this locally-distant assumption, our method predicts a surface normal per pixel using a deep learning-based distant-light photometric stereo method that can deal with spatially varying reflectances. Based on the surface normal estimates, we assess the reconstruction error by re-rendering based on a near-light image formation model that explicitly considers the light fall-off effect. The whole procedure is designed in a differentiable manner with respect to the surface positions so that our method can benefit from a gradient-based method to efficiently predict the surface positions.

To sum up, our paper offers the following contributions:

- We propose a near-light photometric stereo method that can deal with spatially varying reflectances.
- Compared to previous near-light photometric stereo methods, the proposed method does not depend on a simplified parametric reflectance model.
- Compared to existing deep learning based photometric stereo methods, the proposed method explicitly takes nearby light conditions into account.

As a result, the proposed method can handle scenes with diverse materials in contrast to existing near-light photometric stereo methods. At the same time, in contrast to most deep learning-based methods, the proposed method can handle near-light conditions, which should always be considered in a practical setting.

2 Related work

In this section, we describe previous works of photometric stereo on both distant- and near-light assumptions. For distant-light photometric stereo, we mainly discuss the recent deep learning-based methods.

Deep learning-based photometric stereo Early works of photometric stereo [34, 30] assume Lambertian reflectance and many extended works study the use of more flexible parametric models, such as the Torrance-Sparrow model [9], microfacet-based models [32, 7], and bi-polynomial models [29]. Although these methods have greater flexibility in representing reflectances, they still cannot represent real-world reflectances well enough, introducing large estimation errors.

Unlike conventional photometric stereo based on parametric reflectance models, deep learning-based photometric stereo does not explicitly assume a specific reflectance model, but learns it from a synthesized training dataset. Santo *et al*. [26] proposed a fully-connected photometric stereo network, called DPSN, which directly learns the mapping from observations to the corresponding surface normal direction. While DPSN assumes pre-defined light conditions for testing, the newer methods CNN-PS [11] and PS-FCN [6] relax this limitation by handling an arbitrary number of lights and their directions in an order-agnostic way. CNN-PS proposed a new representation for a photometric observation, called an observation map, which represents single-pixel observations under an arbitrary number of light sources by a fixed-shape map representation. In PS-FCN, to handle an arbitrary number of input images, Chen *et al*. used a feature fusion technique with max-pooling to extract a fixed-shape feature map. These methods use a synthesized dataset rendered with realistic bidirectional reflectance distribution functions (BRDFs), such as the MERL BRDF database [16] and the Disney principled BRDF [5] for training.

Unlike these methods, Taniai and Maehara [31] proposed an unsupervised approach. Specifically, they use two networks, a photometric stereo network that outputs a prediction of surface normals and an image reconstruction network that estimates reflectances and outputs re-rendered images, and train the networks by re-rendering loss, which is defined as the difference between input and re-rendered images. We use a similar approach of [31] and minimize the re-rendering loss, but our setting explicitly assumes a near-light setting that cannot be directly addressed by Taniai's work.

Near-light photometric stereo While early works of photometric stereo [34, 30] assume ideal distant-light sources, explicit treatment of a near-light setting in photometric stereo began with the work of Iwahori *et al*. [12]. They consider the effects of spatially varying light directions and light fall-off that occur in near-light settings. These effects pose a challenge in photometric stereo because the image formation model becomes non-linear even with a Lambertian assumption.

The non-linear image formation w.r.t. surface normal results is a non-convex optimization problem. One line of approaches to this difficulty is based on iterative optimization [2, 4, 8, 10, 20]. These methods alternatingly estimate the

scene's shape and albedo based on the image formation model using the prediction of the previous step. Although each step of the optimization can be made convex, the whole objective is non-convex; therefore, a good initial guess is needed for these methods to work well.

Another class of approaches is based on a variational method, yielding nonlinear partial differential equations (PDEs) [19, 18, 17]. For example, Mecca *et al*. [19] consider the intensity ratios of two images and formulate the problem as a quasi-linear PDE. They extend their work in [18, 17] to relax the reflectance model from the Lambertian to the Blinn–Phong model [3]. More recently, Quéau *et al*. [22] reviewed iterative and PDE-based methods. To solve the problems that (1) the convergence of iterative methods is not established and (2) PDE-based methods are sensitive to the initialization, they proposed a provably convergent alternating reweighted least-squares scheme for solving the near-light photometric stereo problem. Although they assume Lambertianness, they show that their method can deal with non-Lambertian observations, such as shadows and specularities, by a robust variational approach [24].

To sum up, a major limitation of existing near-light photometric stereo methods is their dependency on simplified reflectance models. Specifically, most of them rely on the Lambertian model, which limits their applicability in practice. Our proposed method eliminates this restriction and works with diverse, spatially varying BRDFs.

3 Image formation model

We first explain our forward model of how images are formed given a scene's parameters with arbitrary BRDFs and nearby light. The actual (and in fact often ill-posed) task of photometric stereo is then the reverse problem, *i.e*., inferring scene parameters from given images. In Sec. 4 we explain how our algorithm does this. Throughout this paper, function $\boldsymbol{u}_1(\cdot) : \mathbb{R}^3 \to \mathcal{S}^2$ ($\subset \mathbb{R}^3$) represents vector normalization, *i.e*., $\boldsymbol{u}_1(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$.

We denote the 3D position of the j^{th} light source by $\mathbf{s}_j \in \mathbb{R}^3$ and the surface position and surface normal corresponding to the i^{th} pixel by $\mathbf{p}_i \in \mathbb{R}^3$ and $\mathbf{n}_i \in \mathcal{S}^2$. Let us use $\mathbf{l}_{ij}$ to represent the light direction from the i^{th} scene point to the j^{th} light source, *i.e*.,

$$\mathbf{l}_{ij} = \boldsymbol{u}_1(\mathbf{s}_j - \mathbf{p}_i)\,. \tag{1}$$

The observed intensity $m_{ij} \in \mathbb{R}$ at the i^{th} pixel under the j^{th} light without global illumination effects (cast shadows, inter-reflection, *etc*.) can be written as [22]

$$m_{ij} = \Phi_{ij} \frac{1}{\|\mathbf{s}_j - \mathbf{p}_i\|_2^2} \max\left(0, \mathbf{l}_{ij}^\top \mathbf{n}_i\right) \rho_{ij}, \tag{2}$$

where Φ_{ij} is the radiant intensity of the j^{th} light at the surface point corresponding to the i^{th} pixel. ρ_{ij} is the reflectance at the i^{th} point under the j^{th} light, expressed as a function $\rho_{ij} : \mathcal{S}^2 \times \mathcal{S}^2 \to \mathbb{R}$ taking surface normal $\mathbf{n}_i$ and

incoming light direction $\mathbf{l}_{ij}$ as input. The term $\frac{1}{\|\mathbf{s}_j - \mathbf{p}_i\|_2^2}$ accounts for light fall-off and the $\max(\cdot)$ operator accounts for attached shadows.

Let $[u_i, v_i]^\top \in \mathbb{R}^2$ be a pixel position in image coordinates. Its corresponding 3D surface point $\mathbf{p}_i = [x_i, y_i, z_i]^\top$ in world coordinates is

$$\mathbf{p}_i = [u_i, v_i, z_i]^\top$$

under orthographic camera projection, and

$$\mathbf{p}_i = \left[\frac{z_i}{f} u_i, \frac{z_i}{f} v_i, z_i\right]^\top$$

under perspective camera projection, where z_i is the depth, and f is the camera's focal length, which can be obtained through camera calibration.

The surface normal $\mathbf{n}_i$ at surface point $\mathbf{p}_i$ is $\mathbf{n}_i = \boldsymbol{u}_1(\partial_x \mathbf{p}_i \times \partial_y \mathbf{p}_i)$, in which $\times$ is the cross product, and ∂_* represents partial gradient with respect to $*$. Therefore, in orthographic and perspective projection models, the surface normal $\mathbf{n}_i$ can be respectively written as

$$\mathbf{n}_i = \boldsymbol{u}_1([\partial_u z_i, \partial_v z_i, -1])^\top,$$

and

$$\mathbf{n}_i = \boldsymbol{u}_1([f\partial_u z_i, f\partial_v z_i, -z - u_i \partial_u z_i - v_i \partial_v z_i])^\top.$$

In this paper, we use $\mathbf{n}_i = \boldsymbol{\nu}(\mathbf{p}_i)$ to represent conversion from a surface point $\mathbf{p}_i$ to its surface normal $\mathbf{n}_i$ for representing either projection model.

We model the light source's radiant intensity Φ_{ij} as anisotropic point light, which is a common assumption in existing near-light photometric stereo methods [19, 22]. It can be written as

$$\Phi_{ij} = \psi_j \left[\mathbf{l}_{ij}^\top \boldsymbol{\omega}_j\right]^{\mu_j}, \tag{3}$$

where $\psi_j \in \mathbb{R}$ is the light source intensity, $\boldsymbol{\omega}_j \in \mathcal{S}^2$ is the principal direction of a light source, and $\mu_j \in \mathbb{R}$ is an anisotropy parameter. In our setting, we assume these parameters as well as the light source positions $\mathbf{s}_j$ are known from a light calibration method [1, 27, 21, 15].

4 Proposed method

Our goal is to determine surface positions $\mathbf{p}_i$, corresponding surface normal $\mathbf{n}_i$ and reflectances ρ_{ij} from a set of observations m_{ij}, given light source positions $\mathbf{s}_j$ and their radiant intensity parameters $(\psi_j, \mu_j, \boldsymbol{\omega}_j)$ in Φ_{ij}. To alleviate the difficulty of the nearby light setting, we cast the problem into a *per-point distant light* setting, where individual surface points receive different strengths of light from different directions. It allows us to use pre-trained learning-based photometric stereo networks, that are trained under a distant light assumption. Once the prediction of surface normal $\mathbf{n}_i$ and reflectances ρ_{ij} are obtained via the photometric stereo networks, we re-render the scene observations based on the image formation model Eq. (2) and estimate the scene shape $\mathbf{p}_i$ by minimizing the re-rendering loss. Figure 1 illustrates an overview of the proposed method.

4.1 Formulation

We first define a pseudo observation m'_{ij} as

$$m'_{ij} = m_{ij} \frac{\|\mathbf{s}_j - \mathbf{p}_i\|_2^2}{\Phi_{ij}}, \tag{4}$$

in which the light fall-off $\frac{1}{\|\mathbf{s}_j - \mathbf{p}_i\|_2^2}$ and anisotropic radiant intensity Φ_{ij} are discounted from the actual observation m_{ij} in Eq. (2). With this expression, Eq. (2) can be rewritten as

$$m'_{ij} = \rho_{ij} \max(0, \mathbf{l}_{ij}^\top \mathbf{n}_i), \tag{5}$$

which is equivalent to the distant light image formation model except that we do not know the surface point $\mathbf{p}_i$ included in both m'_{ij} and $\mathbf{l}_{ij}$. Under f point light sources, measurements at the i^{th} surface point form a pseudo-observation vector $\mathbf{m}'_i = [m'_{i1}, \cdots, m'_{if}]^\top$, and the corresponding light matrix $\mathbf{L}_i$ can be defined as $\mathbf{L} = [\mathbf{l}_{i1}, \cdots, \mathbf{l}_{if}]^\top$.

Now suppose that we have a guess about the surface position $\mathbf{p}_i$. Then we can compute both the light matrix $\mathbf{L}_i$ and the measurement vector $\mathbf{m}'_i$ from Eqs. (1) and (4), respectively. With the light matrix $\mathbf{L}_i$ and measurement vector $\mathbf{m}'_i$, our method solves for surface normal $\mathbf{n}_i$ and reflectance ρ_{ij} at the i^{th} surface point using two differentiable networks; namely the surface normal estimation network `PS` and the reflectance estimation network `R`:

$$\begin{cases} \mathbf{n}_i^* = \texttt{PS}(\mathbf{m}'_i, \mathbf{L}_i), \\ \rho_{ij}^* = \texttt{R}_j(\mathbf{m}'_i, \mathbf{L}_i), \end{cases} \tag{6}$$

where $\mathbf{n}_i^* \in \mathcal{S}^2$ and $\rho_{ij}^* \in \mathbb{R}$ are the prediction of the surface normal and the reflectance under the j^{th} light, respectively. Unlike previous works which depend on a parametric reflectance model such as the Lambertian model, the capability of handling a variety of BRDFs in the proposed method stems from Eq. (6), whose detail is explained in the next section. Here, we assumed a given guess of the surface position $\mathbf{p}_i$ as input for the networks. However, since the surface normal estimation network `PS` and reflectance estimation network `R` are pretrained and treated as deterministic functions, by substituting Eqs. (1) and (4) into Eq. (6), the prediction of the surface normal $\mathbf{n}_i^*$ and reflectance ρ_{ij}^* become (differentiable) functions of the surface position $\mathbf{p}_i$.

The partial derivative of the surface position $\mathbf{p}_i$, written as $\boldsymbol{\nu}(\mathbf{p}_i)$, also represents surface normal prediction . While the partial derivative $\boldsymbol{\nu}(\mathbf{p}_i)$ is directly calculated from the prediction of the surface position $\mathbf{p}_i$, the estimated normal $\mathbf{n}_i^*$ is constrained by learned prior knowledge in the estimation network `PS`. For robust estimation, we define the estimated surface normal as the weighted mean of the partial derivative of the surface point $\mathbf{p}_i$ and the surface normal $\mathbf{n}_i^*$ obtained by the estimation network:

$$\hat{\mathbf{n}}_i(\mathbf{p}_i) = \boldsymbol{u}_1((1-\kappa)\boldsymbol{u}_1(\boldsymbol{\nu}(\mathbf{p}_i)) + \kappa \mathbf{n}_i^*(\mathbf{p}_i)), \tag{7}$$

where κ (set to 0.5 in our implementation) balances the two surface normals.

In addition, to ensure that the estimates of position $\mathbf{p}_i$ and surface normal $\mathbf{n}_i^*$ are consistent with each other, we use an objective function $\mathcal{L}_n$ defined over the partial derivative of the surface point $\mathbf{p}_i$ and the predicted surface normal $\mathbf{n}_i^*$,

$$\mathcal{L}_n(\{\mathbf{p}_i|\forall i\}) = \sum_i \left\{1 - \boldsymbol{u}_1(\boldsymbol{\nu}(\mathbf{p}_i))^\top \mathbf{n}_i^*(\mathbf{p}_i)\right\}.$$

Once we obtain the predicted surface normal $\hat{\mathbf{n}}_i$ and reflectances ρ_{ij}^*, we reconstruct the re-rendered observations $\hat{m}_{ij} \in \mathbb{R}$ using Eq. (2) as

$$\hat{m}_{ij} = \frac{\Phi_{ij}}{\|\mathbf{s}_j - \mathbf{p}_i\|_2^2} \max\left(0, \mathbf{l}_{ij}^\top \hat{\mathbf{n}}_i\right) \rho_{ij}^*. \tag{8}$$

The re-rendering $\hat{m}_{ij}$ is differentiable[1] with respect to the surface position $\mathbf{p}_i$ because the networks in Eq. (6) as well as Eqs. (3), (1), (4), (7) are all differentiable. Therefore, we minimize the following objective function for estimating surface point $\mathbf{p}_i$ for all i and j starting with an initial guess for $\mathbf{p}_i$:

$$\mathcal{L}_m(\{\mathbf{p}_i|\forall i\}) = \frac{1}{f} \sum_i \sum_j \left\{u_2(m_{ij}) - u_2(\hat{m}_{ij}(\mathbf{p}_i))\right\}^2,$$

in which $u_2(\cdot)$ represents the normalization operation for observations and is defined as $u_2(x_{ij}) = x_{ij}/\|\mathbf{X}\|_F$, $\mathbf{X} = [x_{ij}]$, taking care of the global scaling in observations, as used in [31].

As a result, the final form of the objective function becomes

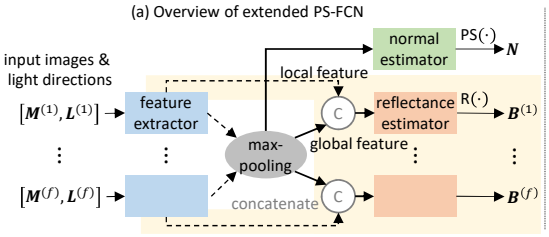
$$\mathcal{L} = (1 - \lambda)\mathcal{L}_m + \lambda \mathcal{L}_n, \tag{9}$$

where the scalar weight $\lambda \in (0, 1)$ balances the two objective functions. Finally, we obtain estimates of the position $\hat{\mathbf{p}}_i = \text{argmin}_{\mathbf{p}} \mathcal{L}$, the surface normal $\hat{\mathbf{n}}_i(\hat{\mathbf{p}}_i)$, and the reflectance $\rho_{ij}^*(\hat{\mathbf{p}}_i)$. Since the objective function $\mathcal{L}$ is non-convex, the proposed method requires an initial guess as with most existing near-light photometric stereo methods. For initialization, we assume that the distance from the camera to the scene is given by a rough measurement and we use a plane as initial scene shape. Equation (8) is defined for grayscale observations. For multi-channel observations, we calculate the re-rendered observations for each color channel and take the sum of the re-rendering loss $\mathcal{L}_m$ from each channel.

4.2 Normal and reflectance estimation networks

The proposed method uses a surface normal network `PS` and a reflectance estimation network `R`. For the surface normal estimation network `PS` we adopt PS-FCN [6]. PS-FCN is an end-to-end differentiable network that takes input

[1] $\max(0, x)$ is differentiable everywhere except at $x = 0$, which is in practice not a problem with numerical differentiation as in many other works.



(b) Network architecture of reflectance estimator

	Layer	Output shape
	Input features (local & global)	$H/2 \times W/2 \times (128 + 128)$
1	Conv + LeakyReLU	$H/2 \times W/2 \times 128$
2	Conv + LeakyReLU	$H/2 \times W/2 \times 128$
3	Deconv + LeakyReLU	$H \times W \times 64$
4	Conv	$H \times W \times c$

Fig. 2. *(a)* Architecture of the surface normal and reflectance estimation networks. The feature extractor and normal estimator are the same as in PS-FCN [6] and we add the reflectance estimator shown in the yellow box. *(b)* Details of the reflectance estimator. "Conv", "Deconv", and "LeakyReLU" mean a convolution layer with a 3×3 kernel, a deconvolution layer with a 3×3 kernel and stride 2, and a Leaky ReLU with a scale factor of $\alpha = 0.1$, respectively. $H \times W$ is the input image size. The input is the concatenated features of the local and global features, where both features have a size of $H/2 \times W/2 \times 128$ and the output is c channel reflectance maps $\mathbf{B}_m^{(j)}$ in the form of the image shape $H \times W \times c$. The weights are shared for all lightings.

images concatenated with vectors of light directions and outputs the corresponding surface normal map. Its authors showed that PS-FCN works well for scenes with spatially varying, diverse real-world BRDFs through a benchmark comparison on a real-world dataset. We extend the original PS-FCN for simultaneous estimation of surface normals and reflectances.

Figure 2 shows an overview of the extended PS-FCN. We add the reflectance estimator $\mathtt{R}$ to the original PS-FCN, which estimates the reflectances $\mathbf{B} \in \mathbb{R}^{p \times f}$, in which an element $B_i^{(j)} \in \mathbb{R}$ represents the reflectance at the i^{th} point (corresponding to the pixel) under the j^{th} lighting. p and f are the numbers of pixels and light sources, respectively. We denote the reflectance map for all pixels under the j^{th} lighting as $\mathbf{B}^{(j)} \in \mathbb{R}^p$ and the reflectances at the i^{th} pixel under all lightings as $\mathbf{B}_i \in \mathbb{R}^f$. The reflectance estimator takes as input the local features that are concatenated with the global feature, and outputs the prediction of reflectance map $\mathbf{B}^{(\cdot)}$ for all lightings. The global feature provides global information such as the object's shape, while the local feature accounts for the reflectances under individual lightings. For the network architecture of the reflectance estimator $\mathtt{R}$, we use an architecture identical to the one for surface normal estimation except for the normalization and output shape.

For training, in addition to the original cosine similarity loss for the surface normals, we use the following re-rendering loss $\mathcal{L}_{\mathbf{B}}$ for the reflectance estimator that is defined with the estimated reflectance $B_i^{(j)}$:

$$\mathcal{L}_{\mathbf{B}} = \sum_i \sum_j \left\{ \check{m}_{ij} - B_i^{(j)} \max\left(0, \mathbf{l}_{ij}^\top \check{\mathbf{n}}_i\right) \right\}^2 . \tag{10}$$

In the above equation, $\check{m}_{ij} \in \mathbb{R}$ and $\check{\mathbf{n}}_i \in \mathcal{S}^2$ represent the ground truth measurements and the surface normal at the i^{th} surface point under the j^{th} lighting. Since this network assumes distant lighting, $\mathbf{l}_{ij}$ represents the lighting direction.

We use the same training dataset as the original PS-FCN, but normalize the input images to remove a global scaling ambiguity. Specifically, a ground truth measurement $\check{\mathbf{m}}_i = [\check{m}_{i1}, \cdots, \check{m}_{if}]$ is normalized for each pixel by two factors: its original norm $\|\check{\mathbf{m}}_i\|_2$ and the number of lights/images f. We normalize the measurements with a scaling factor $s = f^{-\frac{1}{2}} \|\check{\mathbf{m}}_i\|_2^{-1}$. Since scaling the observations also scales the reflectances, we need to undo that by inversely scaling them with s^{-1} to keep them consistent with the original reflectances in the images. We show the evaluation of our extended PS-FCN on a distant-light photometric stereo dataset in our supplementary material. Although Eq. (10) is for grayscale observations, for multi-channel observations, we change the output shape of the reflectance estimator to $\mathbf{B}_m^{(j)} \in \mathbb{R}^{p \times c}$ where c is the number of color channels.

4.3 Implementation

The proposed method obtains predictions of the scene's shape by minimizing the objective function of Eq. (9). In our implementation, to minimize the objective $\mathcal{L}$ we use the Adam optimizer [14] with default settings ($\beta_1 = 0.9$ and $\beta_2 = 0.99$) and set the balancing weight λ to 0.05. We stop iterating when one of the following is met: (1) $\mathcal{L}^{(t+1)} - \mathcal{L}^{(t)} < \tau$ or (2) $t > T_{\text{iter}}$, where $\mathcal{L}^{(t)}$ is the value of the loss $\mathcal{L}$ after the t^{th} iteration. We use $\tau = 10^{-6}$ and $T_{\text{iter}} = 10^4$.

During the iterations, we randomly sample light sources to construct mini-batch data. In our implementation, each iteration uses 32 randomly selected images to reduce the usage of computational resources. The problem would otherwise not fit into the GPU memory if there is a large number of light sources. Too small mini-batches, on the other hand, result in unstable predictions.

Following Quéau [22], for better and faster convergence we use hierarchical scaling optimization, *i.e.*, we reduce the input image resolution and use the resulting solution to initialize the optimization at a higher resolution. In our experiments, we use a coarse-to-fine approach starting from 1/8×, 1/4×, 1/2×, to 1× of the input image resolution. The surface position $\mathbf{p}_i$ at the lowest resolution is initialized with a planar depth map.

The learning rate depends on the scaling of the shape $\mathbf{p}_i$. Using the initial depth d and focal length f, we set the initial learning rate to $0.5 \times \frac{d}{f}$ where the second term corresponds to the physical size of one pixel. For each finer resolution in the coarse-to-fine approach we then set the learning rate to half of the previous coarser resolution's.

5 Experiments

To evaluate the proposed method, we conducted experiments using both synthetic and real-world scenes. For comparison, we used MQ16 [17] and DQ18 [22] as existing methods for near-light photometric stereo, and PS-FCN [6] for distant-light photometric stereo. For the input of PS-FCN, we calculated the directions of each light source using the distance between the camera and the target object, which is used for the initial guess in our near-light photometric stereo

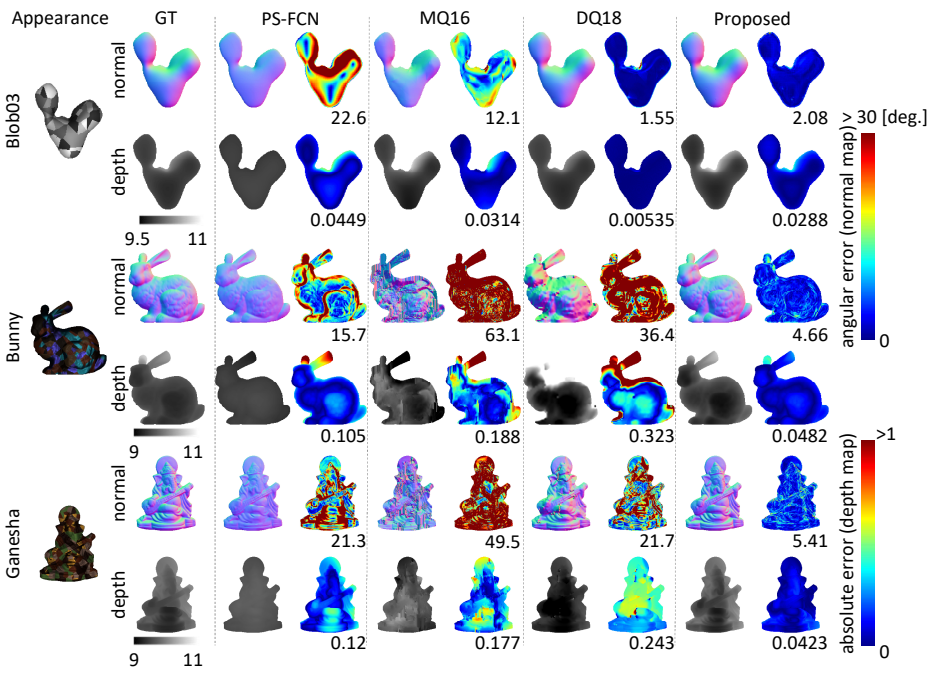


Fig. 3. Estimated results for our synthetic dataset. For each scene and each method, we show the estimated surface normal map, the depth map, the normal error map, and the depth error map. "GT" shows the ground truth of normals and depth. The numbers underneath the error maps show mean angular error in degrees for the surface normal maps, and mean absolute error for the depth maps.

method. Since PS-FCN only estimates the surface normal, we used a quadratic integration-based method [23] to obtain depth maps from estimated normal maps. As for MQ16, the surface normal maps are calculated from the estimated depth maps. Our method is implemented in PyTorch[2]. On an NVIDIA Quadro RTX 8000 GPU it took about 0.6 s per iteration and 30–60 min until full convergence for a scene with 256×256 px.

5.1 Evaluation with synthetic data

We first evaluated our method using a synthetic dataset. For the evaluation, we used three scenes: *Blob03* [13], the Stanford *Bunny* [33], and *Ganesha* [25]. *Blob03* is rendered using spatially varying Lambertian reflectances, and *Bunny* and *Ganesha* consist of five different BRDFs, respectively, that are sampled from measured BRDFs (MERL BRDF database [16]).

We rendered the scenes using the Mitsuba renderer[3]. The camera's focal length was set to 120 mm (35 mm-equivalent), and the image resolution was set

[2] PyTorch v1.1.0: http://pytorch.org
[3] Mitsuba v0.5.0: http://mitsuba-renderer.org

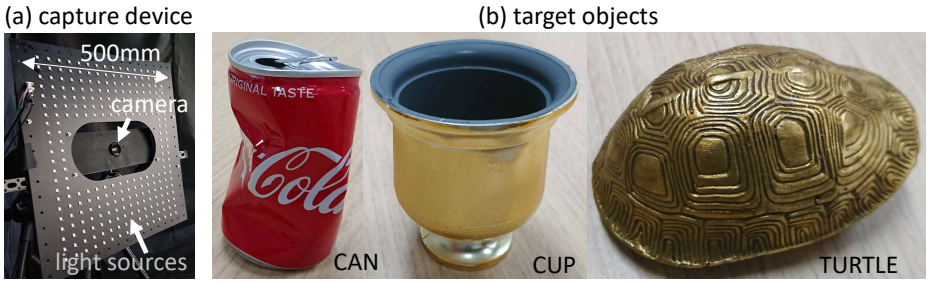


Fig. 4. Real-world experiment setup. *(a):* Our capture device has 256 LEDs on a printed circuit board (500×500 mm) at known positions. The CCD camera (FLIR Blackfly S; 3072×2048 resolution) is fixed to the board at a known position by a 3D-printed frame. We put the target objects about 300 mm away from the device. *(b):* The three target objects used in our experiments: *CAN*, *CUP*, and *TURTLE*.

to 256×256. We defined the scene size so that the distance from camera to target is 10, which means $(X, Y, Z) = (0, 0, 10)$ in camera coordinates, and put 100 point light sources randomly in the range $(X, Y, Z) = (\pm 5, \pm 5, 6 \pm 1)$. The light sources had identical intensity and ideal uniform radiant patterns ($\psi_j = 1$ and $\mu_j = 0$ for all j in Eq. (3)). For the initialization of the shape $\mathbf{p}_i$, we used a planar depth map whose distance from the camera to the target was obtained by the mean distance of the ground truth.

Figure 3 shows the scene and estimation results of our method and the comparison methods PS-FCN [6], MQ16 [17], and DQ18 [22]. While MQ16 and DQ18 work better for the Lambertian *Blob03*, for the other two scenes with more general BRDFs our method shows superior accuracy in both surface normals and depth estimates. MQ16's mean absolute errors for the depth maps are slightly better than DQ18's; however, MQ16 exhibits unstable results in scenes with diverse BRDFs, which can also be seen in the accuracy of the surface normal maps. One of the reasons is that, as discussed by Quéau [22], PDE-based methods are sensitive to the initialization and MQ16 depends on the initialization of both the depth map and the reflectance parameter, *i.e.*, the shininess of the Blinn–Phong model. Although PS-FCN can handle non-Lambertian BRDFs such as the MERL BRDFs, the estimated results are flatter than the ground truth because, in contrast to the proposed method, it ignores the near-light effects.

5.2 Evaluation with real-world data

To evaluate the performance of our method in real-world scenes, we performed a real-world experiment in the setup shown in Fig. 4. We carefully designed our capture device so that the LEDs and the camera are fixed at known positions. We assumed that all LEDs have identical intensity $\psi_j = 1$ and used the radiant intensity distribution obtained from the LED datasheet to calibrate the light emission ($\mu_j = 1$ in Eq. (3)). For the target objects, we used (1) a crushed aluminum can (*CAN*), (2) a plastic cup (*CUP*), and (3) a brass turtle shell (*TURTLE*).

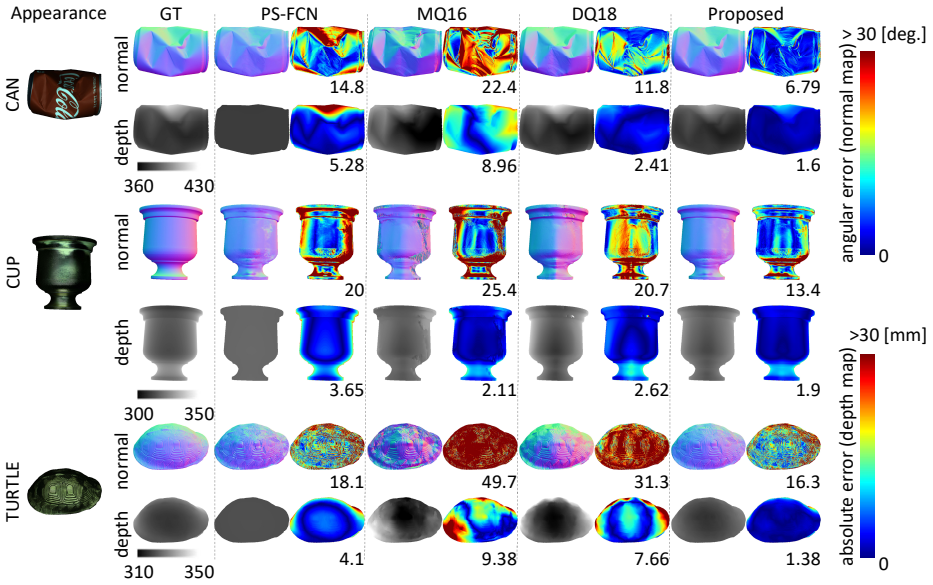


Fig. 5. Estimation results for our real-world dataset. For each scene and each method, we show the estimated surface normal map, the depth map, the normal error map, and the depth error map. "GT" shows the ground truth of normals and depth. The numbers underneath the error maps show mean angular error in degrees for the surface normal maps, and mean absolute error for the depth maps.

All of them are made of different materials and include specular reflectances. To obtain ground truth, we used a structured-light scanner (EinScan Pro) and followed the alignment procedures of the DiLiGenT dataset [28]. Note that, in the evaluation of depth maps we align the estimated depth map to the ground truth because the estimate may have a shift even when we use the initialization calculated from the ground truth. Unlike in the synthetic experiments, we obtained the ground truth by shape-to-image alignment and the absolute depth value of the ground-truth depth map is sensitive to this alignment. The input images are first cropped based on the object mask to avoid redundant computation and are then resized so that the image resolution does not exceed 600×600 pixels due to GPU memory limitations. A more detailed discussion about this limitation can be found in Sec. 6.

Figure 5 shows the evaluation results on our dataset. As can be seen, MQ16 and DQ18 are heavily affected by the specular reflections whereas our method handles them significantly better. For example on the *CUP*, MQ16 is slightly better than DQ18, especially around the center part of the object, because it can handle the non-Lambertian reflectances with the Blinn–Phong model, but it still exhibits large errors due to the instability of the optimization. In contrast, the proposed method works consistenly better on all scenes. Although the pro-

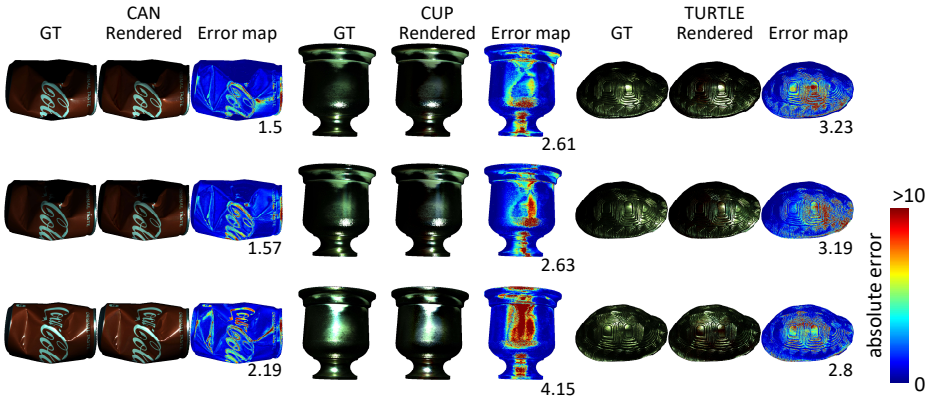


Fig. 6. Rerendering results of our method for our real-world dataset. For each scene, we selected 3 lighting conditions out of the 256 available ones. Each row corresponds to one lighting condition. "GT" and "Rendered" are the ground truth observations (*i.e.*, input images) and rendered images using the estimated reflectances, respectively. The error maps visualize the absolute error and the numbers underneath the error maps are the mean absolute errors in a scaled intensity of 0 to 255. For better visualization, we applied the same brightness correction to both the ground truth and rendered images.

posed method utilizes PS-FCN for normal estimation, it achieves more accurate estimations than PS-FCN by taking near-light effects into account.

To demonstrate the performance of reflectance estimation, Fig. 6 shows rerenderings of the scenes using the estimated reflectances. The proposed method estimates per-pixel and per-light reflectances, and can therefore handle the spatially varying real-world BRDFs. We can see that the obtained reflectance estimations are quite good in all scenes. The estimated reflectances can potentially be used for applications such as a material recognition and parametric BRDF estimation, as well as rendering.

6 Discussion

In this paper, we presented a near-light photometric stereo method for spatially varying reflectances using deep neural networks. Based on the assumption that lighting can be regarded as distant (*i.e.*, parallel) in a small surface area, our formulation allows us to use distant-light photometric stereo in near-light settings. The proposed method uses a state-of-the-art deep learning-based photometric stereo method, PS-FCN, as surface normal and reflectance estimation network, which can handle diverse, spatially varying reflectances. Compared to existing near-light methods which assume over-simplified parametric reflectance model, we showed that our method is superior for scenes with diverse materials. In what follows, we discuss the current limitations and future directions.

Depth discontinuity Since most photometric stereo methods assume continuous surfaces, estimation fails at depth discontinuities. This is also the case for our method (in Fig. 3 we can see that the *Bunny* ears have poor accuracy) since partial derivatives of the surface position $\mathbf{p}_i$ require differentiability.

Limitation in image resolution Since the proposed method is based on deep neural networks, it has a limitation due to the available GPU memory. In our implementation, the optimization for a scene with 460×630 px consumes about 40 GB GPU memory which fits on the *NVIDIA Quadro RTX 8000*'s 48 GB. Since we use mini-batch training with respect to the light sources as described in Sec. 4.3, the number of light sources does not matter. The most memory intensive block in our network is the reflectance estimation in a per-light manner. To reduce GPU memory consumption, one possible approach would be to lower the resolution of the reflectance maps, assuming that scenes do not have high spatial frequency in the reflectances.

Effect of perspective projection While the light source conditions, distant or nearby, and the camera projection model, orthographic or perspective, are independent configurations, near-light photometric stereo typically assumes perspective projection. However, our network, the extended PS-FCN from Fig. 2, assumes orthographic projection in a small patch area. In Sec. 5.1, we only showed results with a fixed focal length. In our supplemental material we show the effects of perspective projection and demonstrate that the result deterioration is not very significant. To handle perspective projection better, one possible extension would be to use a surface normal and reflectance estimation network that works in a per-pixel manner as discussed below.

Alternative networks for surface normal and reflectance estimation In this paper, we presented our framework based on PS-FCN. However, our method can use any differentiable photometric stereo method for the surface normal and reflectance estimation network (Eq. (6)).

One possible alternative would be CNN-PS [11], which estimates surface normals from an observation map which represents per-pixel observations in a fixed shape and achieves the best accuracy in the DiLiGenT benchmark for distant-light photometric stereo [28]. Since CNN-PS works in a per-pixel manner, it is more suitable for the assumptions in Eq. (6). However, to use CNN-PS in our method, in future work we would have to develop (1) a differentiable representation of the observation map with respect to both observations and lighting directions and (2) a simultaneous estimation of reflectances.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP19H01123. Hiroaki Santo and Michael Waechter are grateful for support through a JSPS Research Fellowship for Young Scientists (JP19J10326) and JSPS Postdoctoral Fellowship (JP17F17350), respectively.

References

1. Ackermann, J., Fuhrmann, S., Goesele, M.: Geometric point light source calibration. In: Vision, Modeling, and Visualization. pp. 161–168 (2013)
2. Ahmad, J., Sun, J., Smith, L., Smith, M.: An improved photometric stereo through distance estimation and light vector optimization from diffused maxima region. Pattern Recognition Letters **50**, 15–22 (2014)
3. Blinn, J.F.: Models of light reflection for computer synthesized pictures. In: SIGGRAPH (1977)
4. Bony, A., Bringier, B., Khoudeir, M.: Tridimensional reconstruction by photometric stereo with near spot light sources. In: European Signal Processing Conference (2013)
5. Burley, B.: Physically-based shading at Disney. In: SIGGRAPH 2012 Course Notes (2012)
6. Chen, G., Han, K., Wong, K.Y.K.: PS-FCN: A flexible learning framework for photometric stereo. In: European Conference on Computer Vision (ECCV) (2018)
7. Chen, L., Zheng, Y., Shi, B., Subpa-Asa, A., Sato, I.: A microfacet-based reflectance model for photometric stereo with highly specular surfaces. In: International Conference on Computer Vision (ICCV) (2017)
8. Collins, T., Bartoli, A.: 3D reconstruction in laparoscopy with close-range photometric stereo. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2012)
9. Georghiades, A.S.: Incorporating the Torrance and Sparrow model of reflectance in uncalibrated photometric stereo. In: International Conference on Computer Vision (ICCV) (2003)
10. Huang, X., Walton, M., Bearman, G., Cossairt, O.: Near light correction for image relighting and 3D shape recovery. In: 2015 Digital Heritage (2015)
11. Ikehata, S.: CNN-PS: CNN-based photometric stereo for general non-convex surfaces. In: European Conference on Computer Vision (ECCV) (2018)
12. Iwahori, Y., Sugie, H., Ishii, N.: Reconstructing shape from shading images under point light source illumination. In: International Conference on Pattern Recognition (ICPR) (1990)
13. Johnson, M.K., Adelson, E.H.: Shape estimation in natural illumination. In: Computer Vision and Pattern Recognition (CVPR) (2011)
14. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR) (2014)
15. Ma, L., Liu, J., Pei, X., Hu, Y., Sun, F.: Calibration of position and orientation for point light source synchronously with single image in photometric stereo. Optics Express **27**(4), 4024–4033 (2019)
16. Matusik, W., Pfister, H., Brand, M., McMillan, L.: A data-driven reflectance model. Transactions on Graphics (TOG) **22**(3), 759–769 (Jul 2003)
17. Mecca, R., Quéau, Y.: Unifying diffuse and specular reflections for the photometric stereo problem. In: Winter Conference on Applications of Computer Vision (WACV) (2016)
18. Mecca, R., Rodolà, E., Cremers, D.: Realistic photometric stereo using partial differential irradiance equation ratios. Computers & Graphics **51**, 8–16 (2015)
19. Mecca, R., Wetzler, A., Bruckstein, A.M., Kimmel, R.: Near field photometric stereo with point light sources. SIAM Journal on Imaging Sciences **7**(4), 2732–2770 (2014)

20. Nie, Y., Song, Z.: A novel photometric stereo method with nonisotropic point light sources. In: International Conference on Pattern Recognition (ICPR). pp. 1737–1742. IEEE (2016)
21. Park, J., Sinha, S.N., Matsushita, Y., Tai, Y., Kweon, I.: Calibrating a non-isotropic near point light source using a plane. In: Computer Vision and Pattern Recognition (CVPR). pp. 2267–2274 (2014)
22. Quéau, Y., Durix, B., Wu, T., Cremers, D., Lauze, F., Durou, J.D.: LED-based photometric stereo: Modeling, calibration and numerical solution. Journal of Mathematical Imaging and Vision **60**(3), 313–340 (2018)
23. Quéau, Y., Durou, J.D., Aujol, J.F.: Variational methods for normal integration. Journal of Mathematical Imaging and Vision **60**(4), 609–632 (2018)
24. Quéau, Y., Wu, T., Lauze, F., Durou, J.D., Cremers, D.: A non-convex variational approach to photometric stereo under inaccurate lighting. In: Computer Vision and Pattern Recognition (CVPR) (2017)
25. Rodolà, E., Albarelli, A., Bergamasco, F., Torsello, A.: A scale independent selection process for 3D object recognition in cluttered scenes. International Journal of Computer Vision (IJCV) **102**(1-3), 129–145 (2013)
26. Santo, H., Samejima, M., Sugano, Y., Shi, B., Matsushita, Y.: Deep photometric stereo network. In: ICCV Workshop on Physics Based Vision meets Deep Learning (PBDL) (2017)
27. Santo, H., Waechter, M., Lin, W.Y., Sugano, Y., Matsushita, Y.: Light structure from pin motion: Geometric point light source calibration. International Journal of Computer Vision (IJCV) **128**(7), 1889–1912 (2020)
28. Shi, B., Mo, Z., Wu, Z., Duan, D., Yeung, S.K., Tan, P.: A benchmark dataset and evaluation for non-Lambertian and uncalibrated photometric stereo. Transactions on Pattern Analysis and Machine Intelligence (PAMI) **41**(2), 271–284 (2019)
29. Shi, B., Tan, P., Matsushita, Y., Ikeuchi, K.: Bi-polynomial modeling of low-frequency reflectances. Transactions on Pattern Analysis and Machine Intelligence (PAMI) **36**(6), 1078–1091 (2014)
30. Silver, W.M.: Determining shape and reflectance using multiple images. Master's thesis, Massachusetts Institute of Technology (1980)
31. Taniai, T., Maehara, T.: Neural inverse rendering for general reflectance photometric stereo. In: International Conference on Machine Learning (ICML) (2018)
32. Torrance, K.E., Sparrow, E.M.: Theory for off-specular reflection from roughened surfaces. JOSA **57**(9), 1105–1114 (1967)
33. Turk, G., Levoy, M.: Zippered polygon meshes from range images. In: SIGGRAPH. ACM (1994)
34. Woodham, R.J.: Photometric method for determining surface orientation from multiple images. Optical engineering **19**(1), 139–144 (1980)