DeepPS2: Revisiting Photometric Stereo using Two Differently Illuminated Images

Ashish Tiwari[1] and Shanmuganathan Raman[2]

[1] Prime Minister Research Fellow
[2] Jibaben Patel Chair in Artificial Intelligence
CVIG Lab, IIT Gandhinagar, Gujarat, India
{ashish.tiwari,shanmuga}@iitgn.ac.in

Abstract. Estimating 3D surface normals through photometric stereo has been of great interest in computer vision research. Despite the success of existing traditional and deep learning-based methods, it is still challenging due to: (i) the requirement of three or more differently illuminated images, (ii) the inability to model unknown general reflectance, and (iii) the requirement of accurate 3D ground truth surface normals and known lighting information for training. In this work, we attempt to address an under-explored problem of photometric stereo using just two differently illuminated images, referred to as the PS2 problem. It is an intermediate case between a single image-based reconstruction method like Shape from Shading (SfS) and the traditional Photometric Stereo (PS), which requires three or more images. We propose an inverse rendering-based deep learning framework, called DeepPS2, that jointly performs surface normal, albedo, lighting estimation, and image relighting in a completely self-supervised manner with no requirement of ground truth data. We demonstrate how image relighting in conjunction with image reconstruction enhances the lighting estimation in a self-supervised setting.[3]

Keywords: Photometric Stereo, Deep Learning, Inverse Rendering, Image Relighting

1 Introduction

Inferring the 3D shape of the objects using digital images is a fundamental and challenging task in computer vision research. It directly extends to quality control, virtual/augmented reality, medical diagnosis, e-commerce, etc. The widely used geometric approaches to shape recovery such as binocular [21,42] or multi-view stereo [38,11,25,24,26] methods require images from different views to triangulate the 3D points. However, they rely heavily on the success of image feature matching techniques and fall short of recovering finer details such as indentations, imprints, and scratches. The photometric methods for 3D shape reconstruction use shading cues from either a single image - *Shape from Shading*

[3] Supported by SERB IMPRINT 2 Grant

(SfS) [15] or at least three images - *Photometric Stereo (PS)* [46] to recover surface normals and are known to better preserve the finer surface details.

What are the bottlenecks? The SfS problem is ill-posed due to the underlying convex/concave ambiguity and the fact that infinite surface normals exist to explain the intensity at each pixel [33]. The PS methods are known to handle such ambiguities and provide a unique surface normal defining the intensity at each pixel by using three or more differently illuminated images [14]. However, the well-posed traditional photometric stereo problem (as introduced by Woodhman [46]) assumes the surfaces to be purely Lambertian, which seldom is the case in the real world. Several recent methods [17,50,9,8,7,10] have also addressed shape estimation for non-Lambertian surfaces with unknown reflectance properties. However, they require more images ($\sim 50 - 100$) as input. While there are methods that require as few as six (or even fewer) images [28], our goal is to resort to just two images under a photometric stereo setting.

Scope of the PS2 problem. The scope of this work is to address the photometric stereo problem in an intermediate setting with two images ($m = 2$) between SfS ($m = 1$) and the traditional PS ($m \geq 3$). It can essentially be viewed as a degenerative case of lack of meaningful information due to shadows in a typical three-source photometric stereo setting [13]. Another use case of a PS2 problem arises in the 3D reconstruction of the non-rigid objects [12]. When an object is imaged under three light sources, one could be occluded by the object, and only the other two would provide meaningful cues. Further, the PS2 problem arises when $m \geq 3$ and light sources are coplanar. Such a situation typically occurs when the scene is illuminated by the sun and hence, applies to outdoor PS as well [37,19].

Constraints in addressing the PS2 problem. Several normal fields can offer solutions to the PS2 problem. One can perform an exhaustive search among these normal fields and find the one that best fits the underlying shape satisfying the smoothness constraint [30]. The differential PS formulation implicitly enforces such smoothness. However, it requires explicit knowledge of the surface boundary conditions [29], which is rarely available or requires regularization [13], which is generally tedious owing to heavy parameter tuning. A few methods [29,34] have put forward ways to address the PS2 problem based on the non-differential formulation by recasting it as a binary labeling problem. While such optimization problems can be solved using graph-cut-based algorithms [5], they require the albedo to be known.

Can deep neural networks offer a solution? We use deep neural network to model unknown general surfaces with complex Bidirectional Reflectance Distribution Functions (BRDFs) while addressing the PS2 problem. The photometric stereo problem using deep neural networks has been addressed either under a *calibrated* (known lightings) or an *uncalibrated* (unknown lightings) setting. While most of these methods require 3D ground truth supervision [17,50,9,8,7,10,27,41], a little progress has been made to address PS in a self-supervised manner [20]. However, such self-supervised and uncalibrated methods still require ground truth supervision for lighting estimation.

In this work, we introduce an inverse-rendering-based deep learning framework, called DeepPS2, to address the PS2 problem and work towards developing a completely uncalibrated and self-supervised method. The core idea is to utilize the shading cues from two differently illuminated images to obtain the 3D surface normals. DeepPS2 is designed to perform albedo estimation, lighting estimation, image relighting, and image reconstruction without any ground truth supervision. While image reconstruction is commonly adopted in the existing unsupervised/self-supervised approaches, the appropriate design considerations to perform image relighting using the estimated lightings bring out several interesting insights about the proposed framework.

Contributions The following are the key contributions of this work.

- We introduce DeepPS2, an uncalibrated inverse-rendering based photometric stereo method that jointly performs surface normal, albedo, and lighting estimation in a self-supervised setting[4].
- We propose a self-supervised lighting estimation through light space discretization and perform image relighting (using the estimated lightings) along with image reconstruction.
- We model the specularities explicitly using estimated illumination and albedo refinement .
- To the best of our knowledge, ours is the first work to introduce the PS2 problem using deep learning in a self-supervised manner.

2 Related Work

This section reviews the literature on the PS2 problem and some recent deep learning-based photometric stereo methods.

The PS2 Problem. Onn and Bruckstein [30] discussed the ambiguities in determining surface normals using two images and proposed to use integrability constraint to handle such ambiguities. Sato and Ikeuchi [37] used their method to solve the problem with $m \geq 3$ images under solar illumination, which in a sense addresses the PS2 problem [46]. Later, Yang *et al.* [48] studied the problem, particularly for the convex objects. Kozera provided an analytical resolution to the differential formulation of PS2 [23]. Since 1995 (for over ten years later), only Ikeda [16] addressed the PS2 problem by essentially considering the second image as an auxiliary to better solve the SFS problem. Queau *et al.* [34] addressed the PS2 problem using a graph cut based optimization method. Further, the problem of outdoor PS is being re-explored in several works [1,2]. While these methods attempt to provide a numerical resolution to the PS problem [29,34], we intend to address it using the capacity of deep neural networks.

Deep Learning-based methods. Deep learning has seen great progress addressing photometric stereo [50,9,7,10,17,36]. Santo *et al.* [36] were the first to propose a deep learning-based method to obtain per-pixel surface normals. However, they were limited by the pre-defined order of pixels at the input. Later,

[4] `https://github.com/ashisht96/DeepPS2`

Chen *et al.* in their subsequent works [9,7,10] proposed to model the spatial information using feature-extractor and features-pooling based strategies for photometric stereo. Further, the works by Yao *et al.* [49] and Wang *et al.* [44] proposed to extract and combine the local and global features for better photometric understanding. However, all these methods require ground truth surface normals for supervision which is generally difficult to obtain. Recently, Taniai & Maehara [41] proposed a self-supervised network to directly output the surface normal using a set of images and reconstruct them but with known lightings as input. Kaya *et al.* [20] expanded their method to deal with inter-reflections under an uncalibrated setting, however, the lighting estimation was still supervised. Other methods such as Lichy *et al.* [28], and Boss *et al.* [4] predicted shape and material using three or less and two images (one with and one without flash), respectively. While LERPS [43] infers lighting and surface normal from a single image, it requires multiple images (one at a time) for training. We work towards an uncalibrated photometric stereo method that uses only two differently illuminated images as the input while estimating lightings, surface normals, and albedos, all in a self-supervised manner.

3 Understanding PS2: Photometric Stereo using Two Images

Before describing the PS2 problem, we review some key features of the SfS [15] and the traditional PS problem [46]. We assume that an orthographic camera images the surface under uniform directional lighting with viewing direction $\boldsymbol{v} \in \mathbb{R}^3$ pointing along the z-direction and the image plane parallel to the XY plane of the 3D Cartesian coordinate system XYZ.

3.1 Shape from Shading (SfS)

Consider an anisotropic non-Lambertian surface f with the Bidirectional Reflectance Distribution Function (BRDF) ρ. Let the surface point (x, y) be characterized by the surface normal $\boldsymbol{n} \in \mathbb{R}^3$, illuminated by the light source in the direction $\boldsymbol{\ell} \in \mathbb{R}^3$, and viewed from the direction $\boldsymbol{v} \in \mathbb{R}^3$. The image formation of such a surface is given as per Equation 1.

$$\boldsymbol{I}(x, y) = \rho(\boldsymbol{n}, \boldsymbol{\ell}, \boldsymbol{v})\psi_{f,s}(x, y)\left[\boldsymbol{n}(x, y)^T\boldsymbol{\ell}\right] + \epsilon \tag{1}$$

Here, $\psi_{f,s}(x, y)$ specifies the attached and the cast shadows. It is equal to 0, if (x, y) is shadowed and equal to 1, otherwise. ϵ incorporates the global illumination and noise effect. $\boldsymbol{I}(x, y)$ is the normalized gray level with respect to the light source intensity. Clearly, with albedo and lightings being known apriori, the surface normals $\boldsymbol{n}(x, y)$ in the revolution cone around the lighting direction $\boldsymbol{\ell}$ constitute the set of infinite solutions to Equation 1. Therefore, it becomes an ill-posed problem and is difficult to solve locally.

3.2 Photometric Stereo (PS)

The simplest solution to overcome the ill-posedness of SfS is to have $m \geq 2$ differently illuminated images of the object taken from the same viewpoint. In general, for multiple light sources, Equation 1 extends to the following.

$$\boldsymbol{I}_j(x,y) = \rho(\boldsymbol{n}, \boldsymbol{\ell}_j, \boldsymbol{v})\psi_{f,s}(x,y)\left[\boldsymbol{n}(x,y)^T\boldsymbol{\ell}_j\right] + \epsilon_j \tag{2}$$

Here, the equation is specific to the j^{th} light source. For $m \geq 3$ and a Lambertian surface, Equation 2 formulates a photometric stereo problem (the traditional one for $m = 3$). Solving such a system is advantageous as it is well-posed and can be solved locally, unlike SfS.

3.3 The PS2 problem

With such a non-differential formulation (as in Equation 2), the three unknowns (n_x, n_y, n_z) can be obtained by solving three or more linear equations. However, such a formulation is tricky to solve under two scenarios: (i) when the light sources are coplanar (rank-deficit formulation) and (ii) when $m = 2$. These scenarios lead us to the formulation of the PS2 problem - photometric stereo with two images, as described in Equation 3.

$$\begin{aligned}
\rho(\boldsymbol{n}, \boldsymbol{\ell}_1, \boldsymbol{v})\psi_{f,s}(x,y)\left[\boldsymbol{n}(x,y)^T\boldsymbol{\ell}_1\right] + \epsilon_1 &= \boldsymbol{I}_1(x,y) \\
\rho(\boldsymbol{n}, \boldsymbol{\ell}_2, \boldsymbol{v})\psi_{f,s}(x,y)\left[\boldsymbol{n}(x,y)^T\boldsymbol{\ell}_2\right] + \epsilon_2 &= \boldsymbol{I}_2(x,y) \\
n_x(x,y)^2 + n_y(x,y)^2 + n_z(x,y)^2 &= 1
\end{aligned} \tag{3}$$

The non-linearity in the third part of Equation 3 could give non-unique solution [18]. Adding one more image (under non-coplanar light source configuration) can straightaway solve the problem. However, it will fail when the surface is arbitrarily complex in its reflectance properties. Further, the problem becomes even more difficult to solve when albedo is unknown.

4 Method

In this section, we describe DeepPS2, a deep learning-based solution to the PS2 problem. Further, we describe several design considerations, light space sampling and discretization, and share the training strategy.

4.1 Network Architecture

Let $I_1, I_2 \in \mathbb{R}^{\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ be the two images corresponding to the lighting directions $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$, respectively. The two images along with the object mask

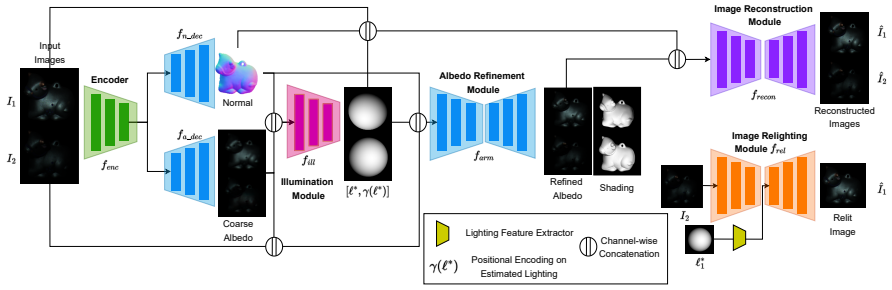


Fig. 1. The proposed inverse rendering framework, called DeepPS2, for shape, material, and illumination estimation. The encoder-decoder design is inspired by Hourglass networks [47]. Layer-wise skip connections are avoided for visual clarity

$M \in \mathbb{R}^{1\times H\times W}$ are fed to the encoder f_{enc} to obtain an abstract feature map $\boldsymbol{\phi}_{img}$, as described in Equation 4.

$$\boldsymbol{\phi}_{img} = f_{enc}([I_1, I_2, M]; \boldsymbol{\theta}_{enc}) \tag{4}$$

Here, $[\cdot]$ represents channel-wise concatenation and $\boldsymbol{\theta}_{enc}$ represents the parameters of the encoder.

Surface Normal and Albedo Estimation. We use $\boldsymbol{\phi}_{img}$ to obtain an estimate of surface normal map $\hat{N}$ and the albedo $\hat{A}$ through the decoders f_{n_dec} and f_{a_dec}, respectively, as described in Equation 5.

$$\hat{N} = f_{n\_dec}(\boldsymbol{\phi}_{img}; \boldsymbol{\theta}_{n\_dec})$$

$$\hat{A} = f_{a\_dec}(\boldsymbol{\phi}_{img}; \boldsymbol{\theta}_{a\_dec}) \tag{5}$$

Here, $\hat{A} = [\hat{A}_1, \hat{A}_2]$ represents the albedos of two images I_1 and I_2 together. The design of each encoder-decoder combination [5] is inspired by that of the Hourglass network [47].

Lighting Estimation. A straightforward way to estimate lighting directions could be to use another fully connected branch and train the network to regress to the desired lightings directly from $\boldsymbol{\phi}_{img}$. However, fully connected layers require a large number of parameters. Further, obtaining precise lighting information directly just from the image features would be difficult since it would not have the explicit knowledge of the structure and reflectance properties of the underlying surface. With an intent to keep the entire architecture fully convolutional, we propose an *illumination module* (f_{ill}) to predict the desired lighting directions by using the estimated normal map and albedos, as described in Equation 6.

$$\hat{l}_i = f_{ill}([\hat{N}, \hat{A}_i]; \boldsymbol{\theta}_{lem}) \tag{6}$$

[5] The detailed layer-wise architecture can be found in our supplementary material.

Here, $i = 1, 2$ corresponding to two images I_1 and I_2, respectively.

At this stage, one straightforward approach could be to use the estimated normal, albedos, and lightings in order to reconstruct the original images through the image rendering equation (see Equation 11). However, the estimated albedo $\hat{A}$ without lighting estimates fails to capture the complex specularities on the surface (see Figure 4). Also, the estimated lightings were a little far from the desired ones.

Thus, the question now is - *how do we validate the accuracy of the estimated albedos and lightings*, especially when there is no ground truth supervision? The albedos and lightings go hand-in-hand and are dependent on each other as far as image rendering is considered, of course, in addition to the surface normal (see Generalized Bas Relief (GBR) ambiguity [3]). To address the aforementioned concerns, we propose two crucial resolves: (i) *albedo refinement* before image reconstruction and (ii) *image relighting* using the estimated lightings.

Albedo Refinement by Specularity Modeling. As discussed earlier, the estimated albedo $\hat{A}$ failed to represent the specularities directly from the image features. Most of the existing deep photometric stereo methods have implicitly handled specularities using multiple differently illuminated images through max-pooling and global-local feature-fusion. However, it is crucial to understand that the specularities are essentially the reflections on the surface, and information about surface geometry can help model such specularities better. Understanding surface geometry becomes even more crucial when we have just one or two images to model the surface reflection. Therefore, we choose to explicitly model these specularities and refine the albedo estimate using a few reasonable and realistic assumptions.

We assume that the specular BRDF is isotropic and is only the function of the half-vector $\boldsymbol{h}$ and the surface normal $\boldsymbol{n}$ at any point on the surface as the BRDF can be re-parameterized to a half-vector based function [35]. In doing so, we could omit the Fresnel Reflection coefficients and geometric attenuation associated with modelling BRDFs. The authors in [31,6] found that the isotropic BRDF can also be modeled simply by two parameters $\theta_h = cos^{-1}(\boldsymbol{n}^T\boldsymbol{h})$ and $\theta_d = cos^{-1}(\boldsymbol{v}^T\boldsymbol{h})$. Therefore, we use the estimated lighting ℓ_i to compute $cos(\theta_h)$ and $cos(\theta_d)$ to further refine the albedo. Additionally, we use positional encoding to model the high-frequency specularities in the refined albedo. In short, we construct the L_i as per Equation 7.

$$L_i = [\boldsymbol{p}_i, \gamma(\boldsymbol{p}_i)]$$

$$\boldsymbol{p}_i = [\boldsymbol{n}^T\boldsymbol{h}_i, \boldsymbol{v}^T\boldsymbol{h}_i] \tag{7}$$

Here, $\gamma(\eta) = [sin(2^0\pi\eta), cos(2^0\pi\eta), ..., sin(2^{m-1}\pi\eta), cos(2^{m-1}\pi\eta)]$. We choose $m = 3$ in our method. Futher, $\boldsymbol{h}_i = \frac{\hat{\boldsymbol{l}}_i + \boldsymbol{v}}{||\hat{\boldsymbol{l}}_i + \boldsymbol{v}||}$.

Following these observations, we use an encoder-decoder based *albedo refinement module* (f_{arm}) to obtain the refined albedo by considering the estimated lightings L_i, albedos $\hat{A}$, surface normal $\hat{N}$, and the underlying images as its

input. Equation 8 describes the information flow.

$$\hat{A}_{i(ref)} = f_{arm}([I_i, \hat{N}, \hat{A}_i, L_i,]; \boldsymbol{\theta}_{arm}) \tag{8}$$

Image Relighting. Generally, at this stage, the existing approaches proceed further to use the rendering equation and reconstruct the input image(s). However, the lightings are either known or have been estimated with ground truth supervision. This allows stable training and offers convincing results. However, in our case, the lightings are estimated without any explicit supervision and are expected to produce learning instabilities. So the question is, *how can we ensure that the estimated lightings are close to the desired ones without any ground truth supervision?*

As an additional check on the authenticity of the estimated lightings, we propose to use them for the image relighting task. We use an *image relighting module* (f_{rel}) to relight one image into the other using the estimated lighting as the target lighting and measure the quality of the relit image, as described in Equation 9.

$$\hat{I}_{1(rel)} = f_{rel}(I_2, \boldsymbol{\phi}(\hat{\boldsymbol{\ell}}_1); \boldsymbol{\theta}_{rel}) \tag{9}$$

Here, $\boldsymbol{\phi}(\hat{\boldsymbol{\ell}}_1)$ is the lighting feature extracted from the desired target lighting $\hat{\boldsymbol{\ell}}_1$. The quality of the relit image fosters the lighting estimates to be close to the desired ones.

Image Reconstruction. Having obtained the estimates of surface normal, albedo, and lightings, we finally use them to obtain the reflectance map $\boldsymbol{R}_i$ using the encoder-decoder based *image reconstruction module* (f_{recon}), as described in Equation 10.

$$\boldsymbol{R}_i = f_{recon}([I_i, \hat{N}, \hat{A}_{i(ref)}, \hat{\boldsymbol{\ell}}_i]; \boldsymbol{\theta}_{recon}) \tag{10}$$

The reflectance image $\boldsymbol{R}_i$ is then used to reconstruct the associated image $\hat{I}_i$, as described in Equation 11.

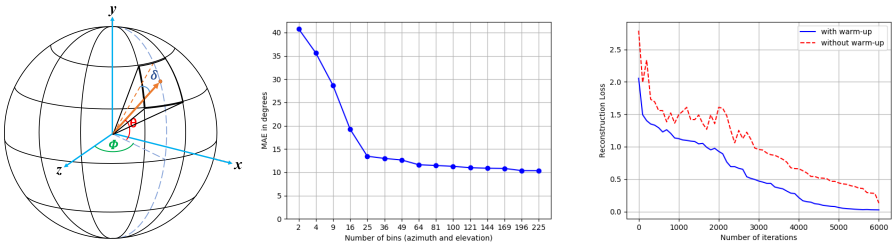
$$\hat{I}_i = \boldsymbol{R}_i \odot max(\hat{\boldsymbol{\ell}}_i^{\,T} \hat{N}, 0) \tag{11}$$

Here, $\odot$ refers to the element-wise multiplication.

In this way, the proposed DeepPS2 produces estimates of surface normal, albedos, and lightings as well as relights the image under target lightings by using only two images as input and no additional ground truth supervision. Based on the network performance, we show that the PS2 problem can be well addressed using the benefits of deep learning framework.

4.2 More on Lighting Estimation: The Light Space Sampling

As discussed earlier, an intuitive approach to estimate light source directions would be to directly regress them from image(s). However, regressing these values to the exact ones is difficult and can cause learning difficulties [7]. Further, under the distant light source assumption, it is easier and better to specify a region in the light space rather than the exact direction while locating the light



(a) (b) (c)

Fig. 2. (a) Light space discretization into $K = 25$ bins. $\delta = 180/2K$ is the maximum angular deviation. (b) Variation of MAE with K. (c) Effect of early stage warm-up

source. Additionally, this eases the light source calibration during data acquisition. Therefore, we choose to formulate the lighting estimation as a classification problem. A few methods in the recent past have adopted the classification formulation [7,10] and weak calibration setting [28] for lighting and shape estimation and have produced excellent results.

In this work, we discretize the light space (upper hemisphere) into $K = 25$ bins (as shown in Fig. 2(a)) i.e. 5 bins along the azimuth direction $\phi \in [0°, 180°]$ centered at $[18°, 54°, 90°, 126°, 162°]$ and 5 bins along the elevation direction $\theta \in [-90°, 90°]$ centered at $[-72°, -36°, 0°, 36°, 72°]$. While each bin suffers a maximum angular deviation of 18° along each direction (Fig. 2(a)), they offer a relatively simpler light source configuration during data acquisition. They can be realized using hand-held lighting devices. Further, learning under such discretized light space configuration allows the network to better tolerate errors in the estimated lightings and the subsequent downstream tasks. During training, the network must select the appropriate bin in the light space to understand the light source configuration from the input image, the estimated normal map, and the albedos.

4.3 Network Training

We use the standard DiLiGenT benchmark dataset [39] having the 10 objects imaged under 96 different light directions with complex non-Lambertian surfaces. We implement DeepPS2 in Pytorch [32] with Adam optimizer [22] and initial learning rate of 1×10^{-4} for 25 epochs and batch size 32 on NVIDIA RTX 5000 GPU. The learning rate is reduced to half after every 5 epochs. It is observed that if the object under consideration has relatively simple reflectance properties, even a randomly initialized network trained with the image reconstruction loss can lead to good solutions. However, for complex scenes, it is better to warm up the network by initializing the weights through weak supervision only at the early stages of training [41,20]. In our case, we perform this warming up for normal, albedo, and lighting estimation through weak supervision using L_1-loss ($\mathcal{L}_{L_1}$), L_2-loss ($\mathcal{L}_{L_2}$), and the perceptual loss ($\mathcal{L}_{perp}$) for first 2000 iterations, as

described in Section 4.4. For weak supervision, we randomly sample 10 images (preferably, each one from a different lighting bin) and estimate the normal map using the least-squares formulation [46], as per Equation 12.

$$\hat{N}' = L^{-1}I \tag{12}$$

It is important to note that the lighting directions in L are from the discretized light space setting, where we compute the lighting direction as the one pointing towards the center of the selected bin. Since we have the images, the normal map $\hat{N}'$, and the discretized lightings L, we compute the *diffuse shading* ($\boldsymbol{n}^T\boldsymbol{\ell}$) and *specular highlights* (regions where $\boldsymbol{n}$ is close to the half-angle $\boldsymbol{h}$ of $\boldsymbol{\ell}$ and viewing direction $\boldsymbol{v} = [0, 0, 1]^T$). Once we have the shadings (diffuse and specular), we compute the albedos ($\hat{A}'$) to use them for weak supervision since an image is the product of the albedo and the shading.

4.4 Loss Functions

In this section, we describe the loss function used for training the entire framework. Equation 13 describes the combination of L_1-loss and the perceptual loss $\mathcal{L}_{perp}$ used for both image reconstruction and relighting.

$$\mathcal{L}_T(X, \hat{X}) = \lambda_1\mathcal{L}_1(X, \hat{X}) + \lambda_2\mathcal{L}_2(X, \hat{X}) + \lambda_{perp}\mathcal{L}_{perp}(X, \hat{X}) \tag{13}$$

Here,

$$\mathcal{L}_1(X, \hat{X}) = \parallel X - \hat{X} \parallel_1$$
$$\mathcal{L}_2(X, \hat{X}) = \parallel X - \hat{X} \parallel_2^2$$

$$\mathcal{L}_{perp}(X, \hat{X}) = \frac{1}{WHC}\sum_{x=1}^{W}\sum_{y=1}^{H}\sum_{z=1}^{C} \parallel \phi(X)_{x,y,z} - \phi(\hat{X})_{x,y,z} \parallel_1 \tag{14}$$

Here, ϕ is the output of VGG-19 [40] network and W, H, C are the width, height, and depth of the extracted feature ϕ, respectively. $\lambda_1 = \lambda_2 = 0.5$ and $\lambda_{perp} = 1.0$.

Weak Supervision. We use the $\mathcal{L}_T$ and the standard cross-entropy loss to provide weak supervision (for first 2000 iterations) for albedos and lightings, respectively. However, for surface normals, we use Equation 15.

$$\mathcal{L}_{norm}(\hat{N}, \hat{N}') = \frac{1}{M}\sum_{p} \parallel \hat{N}_p - \hat{N}'_p \parallel_2^2 \tag{15}$$

5 Experimental Results

In this section, we show the qualitative and quantitative comparison of the DeepPS2 with several baseline approaches. The classical methods [34,29] have

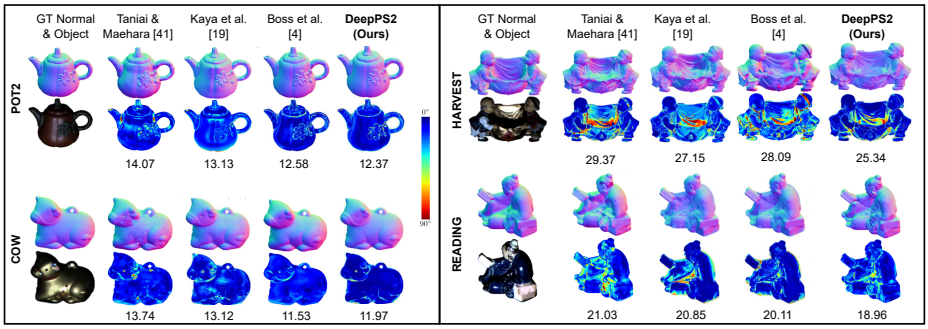


Fig. 3. Surface normal maps obtained using a randomly chosen input image pair. More results are available in the supplementary material

provided the numerical resolution to the underlying ambiguities in PS2. However, the code and results on the DiLiGenT benchmark are not available for comparison. Moreover, since deep learning-based methods have significantly outperformed the traditional photometric stereo methods (even in handling ambiguities), we resort to comparing our work only with the state-of-the-art deep learning-based methods such as UPS-FCN [9], SDPS-Net [7], IRPS [41], Kaya *et al.* [20], Lichy *et al.* [28], and Boss *et al.* [4]. They have been chosen carefully as they can be modified to align with our problem setting by re-training them with two images as input for a fair comparison.

Results on Normal Estimation. Table 1 shows a quantitative comparison of the proposed framework with the other deep learning-based methods. All the methods have been trained with two images as input, and the Mean Angular Error (MAE) is reported to quantify the surface normal estimation. Since IRPS [41] is designed to take two images (one with frontal flash), we evaluate it using pairs of images where one image is lit frontally i.e., from the bin corresponding to $\theta = 0°$ and $\phi = 90°$. From Table 1, we observe that the proposed DeepPS2 obtains the best average MAE value and best (or at least second best) individual scores for eight different objects (except POT1 and BEAR). Even though our framework performs best in the calibrated setting, it outperforms the other baselines under the uncalibrated setting as well. Furthermore, even with no ground truth supervision, our method outperforms other supervised (row 1-6) and self-supervised (row 7-8) methods. To appreciate the results qualitatively, we show a visual comparison of READING, HARVEST, COW, and POT2 with the self-supervised baselines [41,20], and a two-image based supervised method [4] in Fig. 3. Interestingly, DeepPS2 performs the best on objects like HARVEST and READING, having complex shadows and inter-reflections with spatially-varying material.

Results on Albedo Estimation. In Fig. 4, we present a qualitative assessment of the albedos obtained using our method. We observe that the learned albedos are able to handle the complex shadows and specular highlights, especially after refinement using the estimated lightings.

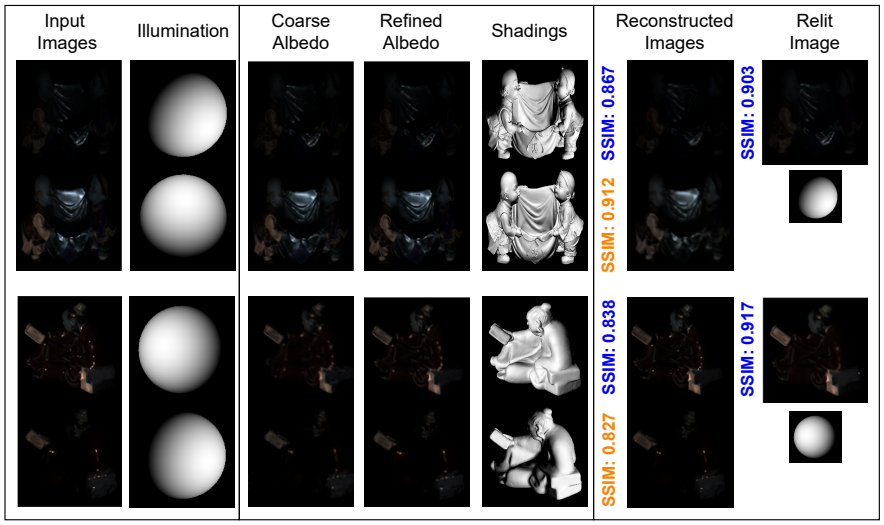


Fig. 4. Inverse rendering results on HARVEST and READING objects. The reconstruction and relighting module yield the SSIM of 0.837 and 0.779, respectively, when averaged over all the objects on the DiLiGenT Benchmark. More results are available in the supplementary material

Results on Lighting Estimation. The goal of discretized lighting is to remove the network's dependence on precise lighting calibration. Therefore, we attempt to model the illumination using the weakly calibrated lighting directions such as front, front-right/left, top, top-right/left, bottom, bottom-right/left, etc. Given that the light space discretization yields an MAE of 18° numerically, we intend to establish that the network may not need precise calibration at all times. A rough and/or abstract understanding of lighting directions should help guide the network towards realistic shape estimation. To better evaluate the performance of the *illumination module*, we visualize the learned illumination over a sphere in Fig. 4. It is observed that the illumination module captures the distribution of light sources essential for modeling the complex specularities in the refined albedos at the later stage.

Results on Image Relighting and Reconstruction. We report the widely used Structural Similarity Index (SSIM) [45] to quantify the quality of the reconstructed and relight images. However, these results are best appreciated visually. Therefore, we use Fig. 4 to show the quality of the generated images. The quality of the results establishes that our inverse rendering results are sufficiently stable for realistic relighting and reconstruction.

5.1 Ablation Studies

In this section, we discuss several design choices in DeepPS2 under different experimental settings.

Table 1. Mean Angular Error (MAE) over 10 randomly chosen image pairs per object from the DiLiGenT Benchmark [39]. **GREEN** and **YELLOW** coloured cells indicate the best and the second best performing methods, respectively. Rows 1-6 and 7-8 correspond to supervised and self-supervised approaches, respectively

Type of Method	Objects → Method ↓	Ball	Cat	Pot1	Bear	Pot2	Buddha	Goblet	Reading	Cow	Harvest	Average
Calibrated	PS-FCN [9]	6.41	20.04	19.67	16.95	21.12	23.04	24.81	29.93	17.23	34.68	21.38 ± 2.05
Uncalibrated	UPS-FCN [9]	9.71	18.97	17.85	15.12	18.62	19.77	22.14	27.36	14.83	31.25	19.56 ± 1.58
Calibrated	SDPS-Net [7]	7.97	19.88	18.12	12.51	18.25	25.12	26.36	27.47	15.21	30.59	20.14 ± 1.17
Uncalibrated	SDPS-Net [7]	7.81	21.74	19.73	13.25	20.47	27.81	29.66	31.12	18.94	34.14	22.6 ± 1.02
Uncalibrated	Boss *et al.* [4]	7.71	14.81	10.17	8.01	12.89	15.98	18.18	21.54	11.96	27.36	14.85 ± 0.98
Uncalibrated	Lichy*et al.* [28]	7.42	20.34	11.87	9.94	11.12	18.75	19.38	21.51	12.93	29.52	16.27 ± 1.01
Calibrated	Taniai & Maehara [41]	7.03	10.02	11.62	8.74	12.58	18.25	16.85	21.31	14.97	28.89	15.03 ± 0.96
Uncalibrated	Kaya *et al.* [20]	6.97	9.57	10.14	8.69	13.81	17.57	15.93	21.87	14.81	28.72	14.81 ± 0.89
Calibrated	DeepPS2 (Ours)	6.17	9.62	10.35	8.87	12.78	14.78	13.29	18.34	10.13	25.18	12.95 ± 0.64
Uncalibrated	DeepPS2 (Ours)	6.28	9.87	10.73	9.67	12.09	14.51	14.22	19.94	11.08	26.06	13.44 ± 0.67

Ablation 1: ***What if we do not include lighting estimation in the framework?*** We attempt to understand the effect of including the lighting information explicitly in the surface normal estimation through such an inverse rendering-based framework. In Table 2, comparing the experiment IDs 1 and 2, we observe that lighting estimation is crucial for the task at hand. This observation is in line with the classical rendering equation that requires lighting directions to understand the reflectance properties and shadows on the surface. Further, we intended to know the deviation in MAE for surface normal estimation when actual lightings (calibrated setting) are used. Although the network performs better under the calibrated setting (see Table 1), the error difference is not very large (0.49 units). This supports our idea of using weaker calibrations for surface normal estimation under distant lightings.

Table 2. Quantitative comparison of various design choices. LE: Lighting Estimation, AR: Albedo Refinement, PE: Positional Encoding, and IR: Image Relighting. Experiments IDs 1-6 include warm-up

ID	LE	AR	PE	IR	Ball	Cat	Pot1	Bear	Pot2	Buddha	Goblet	Reading	Cow	Harvest	Average
1	✗	✗	✗	✗	9.87	36.55	19.39	12.42	14.52	13.19	20.57	58.96	19.75	55.51	26.07
2	✓	✗	✗	✗	9.32	15.62	16.41	10.96	15.77	19.93	18.37	32.34	16.17	30.26	18.51
3	✓	✓	✗	✗	7.37	15.64	10.58	9.37	14.72	15.06	18.1	23.78	16.31	27.17	15.85
4	✓	✓	✓	✗	6.88	12.16	11.13	9.79	15.11	14.89	16.07	20.46	11.85	27.22	14.55
5	✓	✓	✓	✓	6.28	9.87	10.73	9.67	12.09	14.51	14.22	19.94	11.08	26.06	13.44
6	frontally-lit image				6.74	9.38	10.13	9.08	13.18	14.58	14.63	17.84	11.98	24.87	13.24
7	w/o warm-up				12.43	25.01	22.82	15.44	20.57	25.76	29.16	52.16	25.53	44.45	27.33
8	fully supervised				5.14	8.97	10.28	8.92	9.89	12.76	12.38	18.52	9.81	23.22	11.98

Ablation 2: ***Effect of discretizing the light space on normal estimation.*** Fig. 2 (b) shows the effect of a different number of bins on the MAE evaluated over the DiLiGenT benchmark. We resort to choosing $K = 25$ bins as the reduction in the MAE plateaus (roughly) after that point. Further, the light space discretization not only reduces the computational overhead but also helps the network understand the lighting dynamics more holistically. This is evident from the MAE reported in Table 1 and quality of the refined albedos in Fig. 4.

Ablation 3: ***Do albedo refinement and image relighting help in modeling the illumination?*** Qualitative results in Fig. 4 show how well the refined albedos capture the specularities on the surface. Table 2 (IDs 2 and 3) shows the performance improvement by including the *albedo refinement module*. The explicit specularity modeling is observed to produce realistic albedos. The performance is further enhanced through the use of positional encoding (Table 2 ID 4) as it helps the module to better capture the high-frequency characteristics in the refined albedo. Finally, the inclusion of the *image relighting module* further reduces the MAE (Table 2 ID 5). Since the relighting module is solely driven by the estimated lightings, relighting helps in obtaining better surface normal estimates through better lighting estimation as an additional task.

Ablation 4: ***What is the effect of warming up the network with weak supervision at the early stages of training?*** We also consider understanding the effect of weak supervision during the early stage warm-up. Table 2 (IDs 5 and 7) clearly establishes the benefit of warming-up. Fig. 2 (c) shows the the convergence with and without the warm-up. Clearly, an early-stage warm-up provides stable and faster convergence as the outliers in the images are excluded at the early stages during weak supervision.

Ablation 5: ***What if the lighting directions of one image at the input is known?*** We evaluate an interesting and practical case where one of the two input images is captured with collocated light source and camera i.e., $\boldsymbol{\ell} = \boldsymbol{v} = [0, 0, 1]^T$. Since the lighting direction is known, we provide (auxiliary) supervision to the illumination module to obtain a better lighting estimate for the other image. Table 2 (ID 6) shows the results obtained over image pairs having one image sampled from the frontal lighting bin i.e. $\theta = 0°, \phi = 90°$. Under this setting, the method performs better than the completely self-supervised version because frontally-lit (flashed) images offer a better understanding of specularities on complex surfaces. Finally, we also show the performance of DeepPS2 under a fully supervised setting (Table 2 (ID 8)) to establish the upper bound of DeepPS2.

6 Conclusion

In this work, we address the PS2 problem (photometric stereo with two images) using a self-supervised deep learning framework called DeepPS2. In addition to surface normals, the proposed method also estimates albedos and lightings and performs image relighting, all without any ground truth supervision. Interestingly, we demonstrate that weakly calibrated lightings can be enough for the network to learn the underlying shape of an object. In conjunction with image reconstruction, image relighting helps in better lighting estimation. While other uncalibrated methods have used ground truth supervision for learning to estimate lightings, we do so entirely in a self-supervised manner. To the best of our knowledge, we are the first to address photometric stereo using two images in a deep learning setting.

References

1. Abrams, A., Hawley, C., Pless, R.: Heliometric stereo: Shape from sun position. In: European conference on computer vision. pp. 357–370. Springer (2012)
2. Ackermann, J., Langguth, F., Fuhrmann, S., Goesele, M.: Photometric stereo for outdoor webcams. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 262–269. IEEE (2012)
3. Belhumeur, P.N., Kriegman, D.J., Yuille, A.L.: The bas-relief ambiguity. International journal of computer vision **35**(1), 33–44 (1999)
4. Boss, M., Jampani, V., Kim, K., Lensch, H., Kautz, J.: Two-shot spatially-varying brdf and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3982–3991 (2020)
5. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on pattern analysis and machine intelligence **23**(11), 1222–1239 (2001)
6. Burley, B., Studios, W.D.A.: Physically-based shading at disney. In: ACM SIGGRAPH. vol. 2012, pp. 1–7. vol. 2012 (2012)
7. Chen, G., Han, K., Shi, B., Matsushita, Y., Wong, K.Y.K.: Self-calibrating deep photometric stereo networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8739–8747 (2019)
8. Chen, G., Han, K., Shi, B., Matsushita, Y., Wong, K.Y.K.: Deep photometric stereo for non-lambertian surfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
9. Chen, G., Han, K., Wong, K.Y.K.: Ps-fcn: A flexible learning framework for photometric stereo. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–18 (2018)
10. Chen, G., Waechter, M., Shi, B., Wong, K.Y.K., Matsushita, Y.: What is learned in deep uncalibrated photometric stereo? In: European Conference on Computer Vision. pp. 745–762. Springer (2020)
11. Furukawa, Y., Ponce, J.: Accurate, dense, and robust multiview stereopsis. IEEE transactions on pattern analysis and machine intelligence **32**(8), 1362–1376 (2009)
12. Hernández, C., Vogiatzis, G., Brostow, G.J., Stenger, B., Cipolla, R.: Non-rigid photometric stereo with colored lights. In: 2007 IEEE 11th International Conference on Computer Vision. pp. 1–8. IEEE (2007)
13. Hernández, C., Vogiatzis, G., Cipolla, R.: Overcoming shadows in 3-source photometric stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(2), 419–426 (2010)
14. Horn, B., Klaus, B., Horn, P.: Robot vision. MIT press (1986)
15. Horn, B.K.: Shape from shading: A method for obtaining the shape of a smooth opaque object from one view (1970)
16. Ikeda, O.: A robust shape-from-shading algorithm using two images and control of boundary conditions. In: Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429). vol. 1, pp. I–405. IEEE (2003)
17. Ikehata, S.: Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–18 (2018)
18. Ikeuchi, K., Horn, B.K.: Numerical shape from shading and occluding boundaries. Artificial intelligence **17**(1-3), 141–184 (1981)
19. Jung, J., Lee, J.Y., So Kweon, I.: One-day outdoor photometric stereo via skylight estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4521–4529 (2015)

20. Kaya, B., Kumar, S., Oliveira, C., Ferrari, V., Van Gool, L.: Uncalibrated neural inverse rendering for photometric stereo of general surfaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3804–3814 (2021)
21. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: Proceedings of the IEEE international conference on computer vision. pp. 66–75 (2017)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Kozera, R.: On shape recovery from two shading patterns. International Journal of Pattern Recognition and Artificial Intelligence **6**(04), 673–698 (1992)
24. Kumar, S.: Jumping manifolds: Geometry aware dense non-rigid structure from motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5346–5355 (2019)
25. Kumar, S., Dai, Y., Li, H.: Monocular dense 3d reconstruction of a complex dynamic scene from two perspective frames. In: Proceedings of the IEEE international conference on computer vision. pp. 4649–4657 (2017)
26. Kumar, S., Dai, Y., Li, H.: Superpixel soup: Monocular dense 3d reconstruction of a complex dynamic scene. IEEE transactions on pattern analysis and machine intelligence **43**(5), 1705–1717 (2019)
27. Li, J., Robles-Kelly, A., You, S., Matsushita, Y.: Learning to minify photometric stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7568–7576 (2019)
28. Lichy, D., Wu, J., Sengupta, S., Jacobs, D.W.: Shape and material capture at home. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6123–6133 (2021)
29. Mecca, R., Durou, J.D.: Unambiguous photometric stereo using two images. In: International Conference on Image Analysis and Processing. pp. 286–295. Springer (2011)
30. Onn, R., Bruckstein, A.: Integrability disambiguates surface recovery in two-image photometric stereo. International Journal of Computer Vision **5**(1), 105–113 (1990)
31. Pacanowski, R., Celis, O.S., Schlick, C., Granier, X., Poulin, P., Cuyt, A.: Rational brdf. IEEE transactions on visualization and computer graphics **18**(11), 1824–1835 (2012)
32. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
33. Prados, E., Faugeras, O.: Shape from shading: a well-posed problem? In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). vol. 2, pp. 870–877. IEEE (2005)
34. Quéau, Y., Mecca, R., Durou, J.D., Descombes, X.: Photometric stereo with only two images: A theoretical study and numerical resolution. Image and Vision Computing **57**, 175–191 (2017)
35. Rusinkiewicz, S.M.: A new change of variables for efficient brdf representation. In: Eurographics Workshop on Rendering Techniques. pp. 11–22. Springer (1998)
36. Santo, H., Samejima, M., Sugano, Y., Shi, B., Matsushita, Y.: Deep photometric stereo network. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 501–509 (2017)
37. Sato, Y., Ikeuchi, K.: Reflectance analysis under solar illumination. In: Proceedings of the Workshop on Physics-Based Modeling in Computer Vision. pp. 180–187. IEEE (1995)

38. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
39. Shi, B., Wu, Z., Mo, Z., Duan, D., Yeung, S.K., Tan, P.: A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3707–3716 (2016)
40. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
41. Taniai, T., Maehara, T.: Neural inverse rendering for general reflectance photometric stereo. In: International Conference on Machine Learning. pp. 4857–4866. PMLR (2018)
42. Taniai, T., Matsushita, Y., Sato, Y., Naemura, T.: Continuous 3d label stereo matching using local expansion moves. IEEE transactions on pattern analysis and machine intelligence **40**(11), 2725–2739 (2017)
43. Tiwari, A., Raman, S.: Lerps: Lighting estimation and relighting for photometric stereo. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2060–2064. IEEE (2022)
44. Wang, X., Jian, Z., Ren, M.: Non-lambertian photometric stereo network based on inverse reflectance model with collocated light. IEEE Transactions on Image Processing **29**, 6032–6042 (2020)
45. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402. Ieee (2003)
46. Woodham, R.J.: Photometric method for determining surface orientation from multiple images. Optical engineering **19**(1), 139–144 (1980)
47. Yang, J., Liu, Q., Zhang, K.: Stacked hourglass network for robust facial landmark localisation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 79–87 (2017)
48. Yang, J., Ohnishi, N., Sugie, N.: Two-image photometric stereo method. In: Intelligent Robots and Computer Vision XI: Biological, Neural Net, and 3D Methods. vol. 1826, pp. 452–463. SPIE (1992)
49. Yao, Z., Li, K., Fu, Y., Hu, H., Shi, B.: Gps-net: Graph-based photometric stereo network. Advances in Neural Information Processing Systems **33**, 10306–10316 (2020)
50. Zheng, Q., Jia, Y., Shi, B., Jiang, X., Duan, L.Y., Kot, A.C.: Spline-net: Sparse photometric stereo through lighting interpolation and normal estimation networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8549–8558 (2019)