Self-calibrating Photometric Stereo by Neural Inverse Rendering

Junxuan Li and Hongdong Li

Australian National University
{junxuan.li,hongdong.li}@anu.edu.au

Abstract. This paper tackles the task of uncalibrated photometric stereo for 3D object reconstruction, where both the object shape, object reflectance, and lighting directions are unknown. This is an extremely difficult task, and the challenge is further compounded with the existence of the well-known generalized bas-relief (GBR) ambiguity in photometric stereo. Previous methods to resolve this ambiguity either rely on an overly simplified reflectance model, or assume special light distribution. We propose a new method that jointly optimizes object shape, light directions, and light intensities, all under general surfaces and lights assumptions. The specularities are used explicitly to solve uncalibrated photometric stereo via a neural inverse rendering process. We gradually fit specularities from shiny to rough using novel progressive specular bases. Our method leverages a physically based rendering equation by minimizing the reconstruction error on a per-object-basis. Our method demonstrates state-of-the-art accuracy in light estimation and shape recovery on real-world datasets.

Keywords: Uncalibrated photometric stereo; generalized bas-relief ambiguity; neural network; inverse rendering.

1 Introduction

Photometric Stereo (PS) aims to reconstruct the 3D shape of an object given a set of images taken under different lights. Calibrated photometric stereo methods assume the light directions are known in all images [48,49,19,34,50,39,17,9]. However, it is quite a tedious and laborious effort to calibrate the light sources in all input images in practice, often requiring instrumented imaging environment and expert knowledge. How to solve uncalibrated photometric stereo is therefore a crucial milestone to bring PS to practical use.

Recovering the surface shape with unknown light sources and general reflectance is difficult. Previous methods tackle this problem by assuming the Lambertian surfaces. However, Lambertian surfaces in uncalibrated photometric stereo have an inherent 3×3 parameters ambiguity in normals and light directions [4]. When the surface integrability constraint is introduced, this ambiguity can be further reduced to a 3-parameter generalized bas-relief (GBR) ambiguity. Additional information is required to further resolve this ambiguity.

Existing methods to resolve the GBR ambiguity resort to introducing additional knowledge, such as priors on the albedo distribution [2], color intensity profiles [41,29], and symmetric BRDFs [51,30,28]. Drbohlav *et al.* [13] leveraged the mirror-like specularities on a surface to resolve the ambiguity. But they need to manually label the mirror-like specularities for computation. Georghiades [14] addressed the ambiguity by using the TS reflectance model [45]. However, to avoid the local minima, they further assumed the uniformly distributed albedos. These methods either rely on unrealistic assumptions or are unstable to solve. Hence, there is still a gap in applying this technique to more generalized real-world datasets. Recent deep learning-based methods push the boundary of light estimation and surface normal estimation [8,21]. These methods treated light estimation as a classification task. Hence, they lose the ability to continuously represent the lights.

In this paper, we present an inverse rendering approach for uncalibrated photometric stereo. We propose a model which explicitly uses specular effects on the object's surface to estimate both the lights and surface normals. We show that by incorporating our model, the GBR ambiguity can be resolved up to a binary convex/concave ambiguity. Our neural network is optimized via the inverse rendering error. Hence, there is no need to manually label the specular effects during the process. To avoid local minima during the optimization, we propose *progressive specular bases* to fit the specularities from shiny to rough. The key idea of the above technique is to leverage the mirror-like specularities to reduce GBR ambiguity in the early stage of optimization. We propose a neural representation to continuously represent the lighting, normal and spatially-varying albedos. By fitting both the specular and diffuse photometric components via the inverse rendering process, our neural network can jointly optimize and refine the estimation of light directions, light intensities, surface normals, and spatially-varying albedos. In summary, our contributions in this paper are:

- We propose a neural representation that jointly estimates surface normals, light sources, and albedos via inverse rendering.
- We propose progressive specular bases to guide the network during optimization, effectively escaping local minima.

Extensive evaluations on challenging real-world datasets show that our method achieves state-of-the-art performance on lighting estimation and shape recovery.

2 Related Work

Calibrated Photometric Stereo. By assuming the surface of objects to be ideal Lambertian, shapes can be revealed in closed-form with three or more known lights [48]. This restricted assumption is gradually relaxed by following studies [49,34,50,19], where error terms were introduced to account for the deviations from the Lambertian assumption. A regression-based inverse rendering framework [18] was also used for dealing with more general surfaces. Also, in recent years, deep learning-based methods have been widely used in the context of photometric stereo [39,17,24,9,52,47,53]. Santo *et al.* [39] proposed the first photometric stereo neural network, which feeds image pixels into the network

in a predetermined order. Some later works rearranged the pixels into an observation map and then solved the problem per-pixelly [17,24,53,26]. Other deep learning-based approaches used both local and global images cues for normal estimation [9,52,47,16,20]. However, their works assumed both the light directions and intensities to be known. Calibrating light sources may be a tedious process that requires professional knowledge. It will be more convenient to the public if no ground truth light directions are needed for photometric stereo.

Uncalibrated Photometric Stereo. Under the Lambertian surface assumption, there is an inherent generalized bas-relief ambiguity in solving uncalibrated photometric stereo [4]. Traditional works explored many directions to resolve this ambiguity by providing additional knowledge to the system, such as specularities [13], TS model [14], priors on the albedo distribution [2], shadows [43], color intensity profiles [41,29], perspective views[35], inter-reflections [7], local diffuse reflectance maxima [36], symmetric BRDFs [51,30,28], and total variation [37]. In the presence of inaccurate lighting, Quéau *et al*. [38] refined the initial lighting estimation by explicitly modeling the outliers among Lambertian assumption. Other works aim at solving the uncalibrated photometric stereo under natural illumination [33,15]; and semi-calibrated lighting where light directions are known but light intensities are unknown [27,12]. With the advance of the neural network, deep learning-based methods produced state-of-the-art performance in this area. Chen *et al*. [10] proposed a neural network that directly takes images as input, and outputs the surface normal. Later works [8,11] further improved this pipeline by predicting both the light directions and surface normal at the same time. A recent work [40] proposes a way to search for the most efficient neural architecture for uncalibrated photometric stereo. These neural network methods learn prior information for solving the GBR ambiguity from a large amount of training data with ground truth.

Neural Inverse Rendering. Taniai *et al*. [44] proposed the first neural inverse rendering framework for photometric stereo. They proposed a convolutional neural network that takes images at the input and directly outputs the surface normal. Li *et al*. [23] proposed an MLP framework for solving the geometry and reflectance via the reconstruction errors. But their works require the light direction at inputs. Kaya *et al*. [21] use a pre-trained light estimation network to deal with unknown lights. However, their work cannot propagate the reconstruction error back to the light directions and intensities.

In this paper, we propose a neural representation that explicitly models the specularities and uses it for resolving the GBR ambiguity via an inverse rendering process. Our model allows the re-rendered errors to be back-propagated to the light sources and refines them jointly with the normals. Hence, our method is also robust when accounting for inaccurate lighting.

3 Specularities Reduce GBR Ambiguity

In this section, we introduce the notations and formulations of image rendering in the context of uncalibrated photometric stereo under general surfaces. We

discuss the GBR ambiguity under Lambertian surfaces. We further demonstrate that the GBR ambiguity can be resolved under non-Lambertian surfaces with the presence of specularities.

3.1 GBR ambiguity

Given any point in an object's surface, we assume its surface normal to be $\boldsymbol{n} \in \mathbb{R}^3$. It is illuminated by a distant light with direction to be $\boldsymbol{l} \in \mathbb{R}^3$ and light intensity to be $e \in \mathbb{R}^+$. If we observe the surface point from view direction $\boldsymbol{v} \in \mathbb{R}^3$, its pixel intensity $m \in \mathbb{R}^+$ can be modeled as: $m = e\rho(\boldsymbol{n}, \boldsymbol{v}, \boldsymbol{l}) \max(\boldsymbol{n}^T\boldsymbol{l}, 0)$. Here, the $\rho(\boldsymbol{n}, \boldsymbol{v}, \boldsymbol{l})$ denotes a surface point's BRDF function, which is influenced by the surface normal, view direction, and lighting direction. The noise, interreflections, and cast-shadows that deviate from the rendering equation are ignored.

In the above equation, traditional methods assume the surface material to be ideal Lambertian, which makes the BRDF function to be a constant: $\rho(\boldsymbol{n}, \boldsymbol{v}, \boldsymbol{l}) = \rho_d \in \mathbb{R}$. For simplicity, we omit the attached-shadows operator $\max(\cdot)$, and incorporate the diffuse albedo and light intensities into the surface normal and light direction. The equation can be rewrite as

$$\boldsymbol{M} = \boldsymbol{B}^T\boldsymbol{S}, \tag{1}$$

where $\boldsymbol{B} = [\rho_{d_1}\boldsymbol{n}_1, \cdots, \rho_{d_p}\boldsymbol{n}_p] \in \mathbb{R}^{3\times p}$ denotes the normal matrix with p different pixels in a image; $\boldsymbol{S} = [e_1\boldsymbol{l}_1, \cdots, e_n\boldsymbol{l}_n] \in \mathbb{R}^{3\times n}$ denotes the light matrix with n different light sources; $\boldsymbol{M} \in \mathbb{R}^{p\times n}$ denotes the p pixels' intensities under n different light sources. Under this simplified assumption, once the surface point is illuminated by three or more known light sources, the equation has a closed-form solution on surface normals [48].

Under the uncalibrated photometric stereo setting, both the light directions and light intensities are unknown. The above equation will have a set of solutions in a 3×3 linear space. By applying the surface integration constraints, it can be further reduced to a 3 parameters space, which is also known as the generalize bas-relief (GBR) ambiguity in the form as below

$$\boldsymbol{G} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \mu & \nu & \lambda \end{bmatrix} \tag{2}$$

where $\lambda \neq 0; \mu, \nu \in \mathbb{R}$. The transformed normal $\widehat{\boldsymbol{B}} = \boldsymbol{G}^{-T}\boldsymbol{B}$, transformed light $\widehat{\boldsymbol{S}} = \boldsymbol{G}\boldsymbol{S}$. So that both sides of $\boldsymbol{M} = \boldsymbol{B}^T\boldsymbol{G}^{-1}\boldsymbol{G}\boldsymbol{S}$ remain equivalent after the transformation. Additional knowledge need to be introduced for solving the GBR ambiguity above.

3.2 Resolving the ambiguity with specularities

We now explain how specularities on object surfaces provide additional information for reducing the GBR ambiguity. For simplicity, we incorporate the diffuse

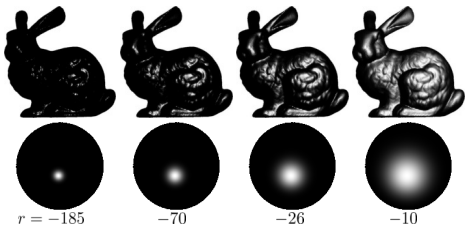


Fig. 1: Rendered "Bunny" and "Sphere" using different specular bases with different roughness. The roughness term controls the sharpness of a specular lobe. The basis presents narrow specular spikes when roughness is small, which is close to the mirror-like reflection.

albedo into surface normal $\boldsymbol{b} = \rho_d \boldsymbol{n}$, and incorporate the light intensity into light direction $\boldsymbol{s} = e\boldsymbol{l}$, and only consider the illuminated points. As stated above, the GBR ambiguity exists when we assume the surface to be Lambertian. There exists a transformed surface normal and light direction $\widehat{\boldsymbol{b}} = \boldsymbol{G}^{-T}\boldsymbol{b}$, $\widehat{\boldsymbol{s}} = \boldsymbol{G}\boldsymbol{s}$, so that the transformed surface and lights will compose the identical pixel observation $\widehat{m} = \widehat{\boldsymbol{b}}^T\widehat{\boldsymbol{s}} = \boldsymbol{b}^T\boldsymbol{G}^{-1}\boldsymbol{G}\boldsymbol{s} = m$.

However, in the presence of specularities, the surface BRDF is not constant anymore. Georghiades [14] models the reflectance as the combination of the diffuse and specular parts as below[1]

$$\rho(\boldsymbol{n}, \boldsymbol{v}, \boldsymbol{l}) = \rho_d + \rho_s \exp(r(1 - \boldsymbol{n}^T\boldsymbol{h})), \tag{3}$$

where the ρ_d is the diffuse albedo, ρ_s denotes the specular albedo, and the $r \in \mathbb{R}^-$ denotes the roughness. $\boldsymbol{h} = \frac{\boldsymbol{v}+\boldsymbol{l}}{||\boldsymbol{v}+\boldsymbol{l}||}$ is the half-unit-vector between view direction $\boldsymbol{v}$ and light direction $\boldsymbol{l}$. Therefore, with the specular terms, the image intensity after the GBR transformation is

$$\begin{aligned}\widehat{m} &= \widehat{\boldsymbol{b}}^T\widehat{\boldsymbol{s}} + \rho_s||\widehat{\boldsymbol{s}}|| \exp(r(1 - \frac{\widehat{\boldsymbol{b}}^T}{||\widehat{\boldsymbol{b}}||}\frac{\boldsymbol{v} + \frac{\widehat{\boldsymbol{s}}}{||\widehat{\boldsymbol{s}}||}}{||\boldsymbol{v} + \frac{\widehat{\boldsymbol{s}}}{||\widehat{\boldsymbol{s}}||}||})) \\ &= \boldsymbol{b}^T\boldsymbol{G}^{-1}\boldsymbol{G}\boldsymbol{s} + \rho_s||\boldsymbol{G}\boldsymbol{s}|| \exp(r(1 - \frac{\boldsymbol{b}^T\boldsymbol{G}^{-1}}{||\boldsymbol{G}^{-T}\boldsymbol{b}||}\frac{\boldsymbol{v} + \frac{\boldsymbol{G}\boldsymbol{s}}{||\boldsymbol{G}\boldsymbol{s}||}}{||\boldsymbol{v} + \frac{\boldsymbol{G}\boldsymbol{s}}{||\boldsymbol{G}\boldsymbol{s}||}||})),\end{aligned} \tag{4}$$

where $||\cdot||$ denotes the length of a vector. In general, $m = \widehat{m}$ only holds for all pixels in the images when the GBR transformation matrix $\boldsymbol{G}$ is identity matrix. Theoretical proof was made by Georghiades [14] that when providing with four different $(\boldsymbol{b}, \boldsymbol{s})$ pairs, it is sufficient to solve the GBR ambiguity up to the binary convex/concave ambiguity, *i.e.* $\lambda = \pm 1; \mu, \nu = 0$. However, even a global minimum exists on $\boldsymbol{G}$, there is no guarantee that no local minima in the 3 parameters space of λ, μ and ν. Solving the above equation is still challenging given the existence of noise, shadows and inter-reflections in real world images. To avoid the local minima, Georghiades [14] assumed that the specular albedo ρ_s is uniform across the surface. This uniform specular albedo assumption prevents their method from being applied to more general objects. In this paper, we are

[1] For simplicity, we omit some terms from [14] without affecting the correctness of their proof.

aiming to solve this problem in more general surfaces, *i.e.* under spatially-varying diffuse and spatially-varying specular albedo.

Resolving the GBR ambiguity by Specular Spikes In fact, the GBR ambiguity can also be resolved by merely four or more pairs of mirror-like reflection effects (*i.e.* specular spikes) on a surface [13]. The roughness term r in equation 3 controls the sharpness of a specular lobe. As shown in Fig. 1, when roughness is small, the resulted material is very close to a mirror-like material (see $r_i = -185$). The specular basis reaches its highest value when $1 - \boldsymbol{n}^T \frac{\boldsymbol{v}+\boldsymbol{l}}{||\boldsymbol{v}+\boldsymbol{l}||} = 0$. Since all the three vectors here are unit vectors, the above equation holds when surface normal $\boldsymbol{n}$ is a bisector between the viewing direction $\boldsymbol{v}$ and the light direction $\boldsymbol{l}$. Hence, we have the following equation when the basis function reach its highest value, *i.e.* where the mirror-like specularities happens

$$\boldsymbol{v} = 2(\boldsymbol{l}^T\boldsymbol{n})\boldsymbol{n} - \boldsymbol{l}. \tag{5}$$

From the *consistent viewpoint constraint* [13], the GBR ambiguity can be reduced to two-parameteric group of transformations (rotation around the viewing vector and isotropic scaling). However, the mirror-like specular spikes needed to be manually labeled in previous method [13]. While in our paper, these mirror-like specular effects can be automatically fitted via our neural network.

4 Proposed Method

We propose a neural network based method that aims at inverse rendering the object by factoring the lighting, surface normal, diffuse albedo, and specular components. This section describes our model for solving uncalibrated photometric stereo in the presence of specularities.

4.1 Proposed image rendering equation

Following previous works on uncalibrated photometric stereo, we make the following assumptions on the problem. We assume that the images are taken in orthographic views. Hence the view direction are consistent across the object surface, $\boldsymbol{v} = [0, 0, 1]^T$. The object is only illuminated once by distance lights with unknown direction $\boldsymbol{l}$ and intensities e. Given the above assumptions, we now rewrite the rendering equation as

$$m = e\rho(\boldsymbol{n}, \boldsymbol{l}) \max(\boldsymbol{n}^T\boldsymbol{l}, 0). \tag{6}$$

Here, the only information we have is the observation of the surface point's pixel intensity m. Our target is to inverse this rendering equation to get all the other unknown terms, such as surface normal $\boldsymbol{n}$, light direction $\boldsymbol{l}$, light intensity e, and surface BRDF function $\rho(\cdot)$. In the following sections, we present our model to parameterize and optimize these terms.

4.2 BRDF modeling

As discussed by equation 2 above, the Lambertian surface assumption alone will lead to GBR ambiguity in solving uncalibrated photometric stereo problem. Hence, we model the reflectance as the combination of the diffuse and specular parts as $\rho(\boldsymbol{n}, \boldsymbol{l}) = \rho_d + \rho_s(\boldsymbol{n}, \boldsymbol{l})$, where the ρ_d is the diffuse albedo, and $\rho_s(\boldsymbol{n}, \boldsymbol{l})$ is the specular terms. We further model the specular term as the summation of a set of specular bases as below

$$\rho_s(\boldsymbol{n}, \boldsymbol{l}) = \sum_{i=1}^{k} \rho_{s_i} \exp(r_i(1 - \boldsymbol{n}^T \boldsymbol{h})), \tag{7}$$

where $\boldsymbol{h} = \frac{\boldsymbol{v}+\boldsymbol{l}}{||\boldsymbol{v}+\boldsymbol{l}||}$ is the half-unit-vector between view direction $\boldsymbol{v}$ and light direction $\boldsymbol{l}$; k is the number of bases. Here, we adopted the Spherical Gaussian [46] as our basis function. The ρ_{s_i} denotes the specular albedo, and the $r_i \in \mathbb{R}^-$ denotes the roughness. The lower the roughness, the more shiny the material will be. We rendered two objects with the proposed specular basis, as shown in Fig. 1.

In summary, our BRDF modeling takes both the diffuse and specular component into consideration and estimate them jointly. We also model the specularities as a summation of a set of bases, which enable the material to range from shiny to rough. We can now rewrite the rendering equation as below

$$m = e(\rho_d + \sum_{i=1}^{k} \rho_{s_i} \exp(r_i(1 - \boldsymbol{n}^T \boldsymbol{h}))) \max(\boldsymbol{n}^T \boldsymbol{l}, 0). \tag{8}$$

4.3 Progressive Specular Bases

Inspired by the two ways of resolving GBR in Sec. 3.2, we proposed the novel *progressive specular bases* to solve the uncalibrated photometric stereo robustly. The key idea of progressive specular bases is to first fit the surface with only mirror-like specular bases (bases with small roughness term r_i); then, we gradually enable the other specular bases for more diffuse effects (bases with large roughness).

At the early stage of optimization, we only enable mirror-like specular bases, the network will attempt to solve uncalibrated photometric stereo using only the mirror-like specular spikes. Then, as the optimization progresses, other specular bases for the network are gradually enabled to fit those diffuse effects. Our progressive specular bases will guide the network away from local minima at the early stage of optimization, resulting better optimized results in the end.

The progressive specular bases is achieved by applying a smooth mask on the different specular basis (from small roughness with mirror-like effects to large roughness with less sharp effects) over the course of optimization. The weights applied to the different specular bases are defined as below

$$\rho_s(\boldsymbol{n}, \boldsymbol{l}) = \sum_{i=1}^{k} \omega_i(\alpha) \rho_{s_i} \exp(r_i(1 - \boldsymbol{n}^T \boldsymbol{h})), \tag{9}$$

where the weight $\omega_i(\alpha)$ is defined as

$$\omega_i(\alpha) = \begin{cases} 0 & \text{if } \alpha < i \\ \frac{1-\cos((\alpha-i)\pi)}{2} & \text{if } 0 \leq \alpha - i < 1 \\ 1 & \text{if } \alpha - i \geq 1 \end{cases} \tag{10}$$

$\alpha \in [0, k]$ will gradually increase during the optimization progress. The defined weights above are inspired by a recent coarse-to-fine positional encoding strategy on camera pose estimation [25]. In the early stage of optimization, the α is small, hence the weight $\omega(\alpha) = 0$ will be zero for those specular bases with roughness r_i, where $i > \alpha$. As the optimization progress, we gradually activate the specular bases one by one. When $\alpha = k$, all the specular bases is used, hence, equation 9 is identical to equation 7 in the final stage. In practice, we set the specular roughness terms $\boldsymbol{r} = \{r_i | i \in \{1, \cdots, k\}\}$ in ascending order. So that the above weight will gradually activate the specular bases from small roughness to large roughness.

To sum up, when applying progressive specular bases, the network will focus on fitting the bright specular spikes at an early stage; then, as the optimization progress, more specular bases are available for the network to fit on the diffuse effects.

4.4 Neural representation for surfaces

Here, we describe our neural representation for object surface modeling. Inspired by the recently proposed coordinately-based multilayer-perceptron (MLP) works [32], we proposed two coordinately-based networks which take only the pixel coordinates (x, y) at input, and output the corresponding surface normal and diffuse albedos and specular albedos.

$$\boldsymbol{n} = N_{\Theta}(x, y), \tag{11}$$

$$\rho_d, \boldsymbol{a} = M_{\Phi}(x, y). \tag{12}$$

Where $N_{\Theta}(\cdot), M_{\Phi}(\cdot)$ are MLPs with Θ, Φ to be their parameters respectively. Given a image pixel coordinates (x, y), the two MLPs directly output the surface normal $\boldsymbol{n}$, diffuse albedo ρ_d, and specular albedos $\boldsymbol{a} = \{\rho_{s_i} | i \in \{1, \cdots, k\}\}$ of that position.

4.5 Neural representation for lighting

Next, we describe the parameterization of the light direction and intensity. Let $\boldsymbol{I} \in \mathbb{R}^{h \times w}$ denotes the image taken under a light source, where h, w denote the height and width of the input image. The direction and intensity of that light source are directly predicted by feeding this image into a convolutional neural network:

$$e, \boldsymbol{l} = L_{\Psi}(\boldsymbol{I}). \tag{13}$$

where $L_{\Psi}(\cdot)$ is a convolutional neural network with its parameters Ψ. The network $L_{\Psi}(\cdot)$ takes only the image $\boldsymbol{I}$ as input, directly output the corresponding light direction $\boldsymbol{l}$ and light intensity e. Unlike previous deep learning based lighting estimation network [8,11,21], we do not fix the lighting estimation at testing. Instead, the lighting estimation is further refined (*i.e.* fine-tuned) on the testing images by jointly optimizing the lighting, surface normals, and albedos via the reconstruction loss.

5 Implementation

This section describes the detail of network architectures, hyperparameters selection, and loss functions.

Network architectures The surface normal net $N_{\Theta}(\cdot)$ uses 8 fully-connected layers with 256 channels, followed by a ReLU activation function except for the last layer. The material net $M_{\Phi}(\cdot)$ uses the same structure but with 12 fully-connected ReLU layers. We apply a positional encoding strategy with 10 levels of Fourier functions to the input pixel coordinates (x, y) before feeding them to the normal and material MLPs. The lighting network $L_{\Psi}(\cdot)$ consists of 7 convolutional ReLU layers and 3 fully connected layers. Please see the supplementary material for detailed network architectures.

For the choices of specular bases, we initialize the roughness value for each basis range from $-r_t$ to $-r_b$ with logarithm intervals

$$r_i = -\exp(\ln r_t - (\ln r_t - \ln r_b)\frac{i-1}{k-1}), \tag{14}$$

where $i \in [1, \cdots, k]$ denotes the index of basis. In testing, we empirically set the number of bases $k = 12$, $r_t = 300$, and $r_b = 10$.

Pre-training light model The light model $L_{\Psi}(\cdot)$ is pre-trained on a public available synthetic dataset, Blobby and Sculpture datasets [10]. We trained the $L_{\Psi}(\cdot)$ for 100 epoches, with batch size to be 64. We adopt the Adam optimizer [22] for updating the network parameter Ψ with learning rate 5.0×10^{-4}. The light network is pre-trained for once, based on the pre-train loss $\mathcal{L}_{\text{pre}} = (1-\boldsymbol{l}^T\bar{\boldsymbol{l}})+(e-\bar{e})^2$, where the first term is cosine loss for light directions, the second term is mean-square-error for intensities. The same network is then used for all other testing datasets at test time.

Testing At testing stage, we continue refining the lighting from pre-trained light net $L_{\Psi}(\cdot)$, while the proposed normal-MLP $N_{\Theta}(\cdot)$ and material-MLP $M_{\Phi}(\cdot)$ are optimized from scratch via the reconstruction loss. As the reconstruction loss, we use the mean absolute difference, which is the absolute difference between observed intensity $\boldsymbol{M} \in \mathbb{R}^{p \times n}$ and re-rendered intensity $\overline{\boldsymbol{M}}$.

$$\mathcal{L} = \frac{1}{pn}\sum_{i=1}^{p}\sum_{j=1}^{n}|\boldsymbol{M}_{i,j} - \overline{\boldsymbol{M}}_{i,j}|, \tag{15}$$

Table 1: Ablation study on effectiveness of progressive specular bases (PSB). We compare the models with and without progressive specular bases. Applying progressive specular bases will consistently improve estimation accuracy.

Model	direction	intensity	normal
$\boldsymbol{r}$	5.30	0.0400	9.39
$\boldsymbol{r}$ + PSB	4.42	0.0382	7.71
trainable $\boldsymbol{r}$	4.75	0.0372	8.57
trainable $\boldsymbol{r}$ + PSB	**4.02**	**0.0365**	**7.05**

Table 2: Quantitative results on DiLiGenT where our model takes different lighting as initialization. Our model performs consistently well when varying levels of noise are applied to the lighting estimation. It shows that our model is robust against the errors in lighting estimation.

	L_Ψ	$+20^\circ$	$+30^\circ$	$+50^\circ$	$+70^\circ$
direction	4.02	4.07	4.14	4.34	4.40
intensity	0.0365	0.0337	0.0356	0.0358	0.0355
normal	7.05	7.40	7.40	7.40	7.44

where the above summation is over all p pixels under n different light sources. At each iteration, we sampled pixels from 8 images and feed them to the networks. The iterations per-epoch depends on the number of images of the scene. We run 2000 epoches in total. The Adam optimizer is used with the learning rate being 10^{-3} for all parameters.

Training and testing time Our framework is implemented in PyTorch and runs on a single NVIDIA RTX3090 GPU. The pre-training time of our light model L_Ψ only takes around 2 hours. In comparison, previous deep methods [11,21] take more than 22 hours in training light models. The reason is that we shift part of the burden of solving lightings from the neural light model to the inverse rendering procedural. Hence, our light model can be relatively lightweight and easy to train compared to previous deep learning based light estimation networks [11,21].

In testing, our method takes an average of 16 minutes to process each of the ten objects in DiLiGenT [42] benchmark, ranging from 13 minutes to 21 minutes. In comparison, previous CNN-based inverse rendering methods [44,21] take on average 53 minutes per object in testing. The reason is that both of our object modeling net N_Θ, M_Φ are simple MLPs. Hence, we can achieve a much faster forward-backward time when optimizing the MLP-based network than previous CNN-based methods.

6 Experiments

6.1 Testing Dataset

We conduct experiments in following public real-world datasets: DiLiGenT [42], Gourd&Apple dataset [1], and Light Stage Data Gallery [6]. They all provide calibrated light directions and light intensities as ground truth for evaluation. DiLiGenT contains 10 objects; each object has 96 images captured under different lighting conditions; a high-end laser scanner captured ground truth surface

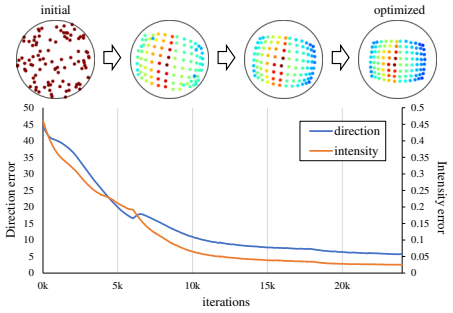


Fig. 2: Visualization of the lighting optimization under noised input. The predicted light distribution over a sphere is represented as the spheres above. The left-most lighting sphere shows the noised lighting estimation at the start of optimization. Our model gradually refines the incorrect lighting during optimization (from left to right) and provides the optimized result.

normal is available for evaluation. Gourd&Apple contains three objects; each object has around 100 images captured under different lighting conditions; Light Stage Data Gallery contains 252 images per object; following previous works [8], we select in total 133 images illuminated by forward-facing lights for photometric stereo. Unfortunately, Gourd&Apple and Light Stage Data Gallery did not provide ground truth surface normal for quantitative evaluation.

6.2 Evaluation metrics

In this paper, we use mean angular errors (MAE) as an evaluation metric for surface normal estimate and light direction estimation. Lower MAE is preferred.

As light intensities among different images can only be estimated up to a scale factor, we follow previous work [8] to use scale-invariant relative error

$$E_{si} = \frac{1}{n}\sum_{i}^{n}\frac{|se_i - \overline{e_i}|}{\overline{e_i}}, \tag{16}$$

where $e_i, \overline{e_i}$ denote the estimated and ground truth light intensity of i-th light respectively; s is the scale factor computed by solving $\text{argmin}_s \sum_i^n (se_i - \overline{e_i})^2$ with least squares. Lower scale-invariant relative error is preferred.

6.3 Ablation study

Effectiveness of the Progressive Specular Bases To show the effectiveness of the proposed progressive specular bases, we conduct the ablation study as shown in Tab. 1. The models are evaluated in the DiLiGenT dataset. In the first and second row of the table, we evaluate the model with and without the progressive specular bases. It shows that the model with progressive specular bases achieves lower error in both lighting and normal estimation. In the third and fourth row of the table, instead of using fixed (as defined by equation 14) roughness terms, we set these roughness terms as trainable parameters. Results also demonstrate that when the roughness terms are trainable, progressive specular bases consistently improve the estimation accuracy. We also observe that the trainable roughness

Table 3: Evaluation results on DiLiGenT benchmark. Here, **bold** indicates the best results and underline denotes the second best results.

(a) Normal estimation results on DiLiGenT benchmark.

Method	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	average
SM10[41]	8.90	11.98	15.54	19.84	22.73	48.79	73.86	16.68	50.68	26.93	29.59
WT13[51]	4.39	6.42	13.19	36.55	19.75	20.57	55.51	9.39	14.52	58.96	23.93
PF14[36]	4.77	9.07	14.92	9.54	19.53	29.93	29.21	9.51	15.90	24.18	16.66
LC18[28]	9.30	10.90	19.00	12.60	15.00	18.30	28.00	12.40	15.70	22.30	16.30
UPS-FCN[10]	6.62	11.23	15.87	14.68	11.91	20.72	27.79	13.98	14.19	23.26	16.02
BK21[21]	3.78	5.96	13.14	7.91	10.85	11.94	25.49	8.75	10.17	18.22	11.62
SDPS-Net[8]	2.77	6.89	8.97	8.06	8.48	11.91	17.43	8.14	7.50	14.90	9.51
SK21[40]	3.46	5.48	10.00	8.94	6.04	9.78	17.97	7.76	7.10	15.02	9.15
GCNet[11]+PS-FCN[10]	2.50	5.60	**8.60**	7.90	7.80	9.60	16.20	7.20	7.10	14.90	8.70
Ours	**1.24**	**3.82**	9.28	**4.72**	**5.53**	**7.12**	**14.96**	**6.73**	**6.50**	**10.54**	**7.05**

(b) Light intensity estimation results on DiLiGenT benchmark.

Method	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	average
PF14[36]	0.0360	0.0980	0.0530	0.0590	0.0740	0.2230	0.1560	**0.0170**	**0.0440**	0.1220	0.0882
LCNet[8]	0.0390	0.0610	0.0480	0.0950	0.0730	0.0670	0.0820	0.0580	0.0480	0.1050	0.0676
GCNet[11]	0.0270	0.1010	0.0320	0.0750	**0.0310**	0.0420	0.0650	0.0390	0.0590	0.0480	0.0519
Ours	**0.0194**	**0.0186**	**0.0206**	**0.0321**	0.0621	**0.0418**	**0.0230**	0.0303	0.0816	**0.0352**	**0.0365**

(c) Light direction estimation results on DiLiGenT benchmark.

Method	Ball	Bear	Buddha	Cat	Cow	Goblet	Harvest	Pot1	Pot2	Reading	average
PF14[36]	4.90	5.24	9.76	5.31	16.34	33.22	24.99	2.43	13.52	21.77	13.75
LCNet[8]	3.27	3.47	4.34	**4.08**	4.52	10.36	6.32	5.44	2.87	**4.50**	4.92
GCNet[11]	1.75	2.44	**2.86**	4.58	**3.15**	2.98	**5.74**	**1.41**	**2.81**	5.47	**3.32**
Ours	**1.43**	**1.56**	4.22	4.41	4.94	**2.26**	6.41	3.46	4.19	7.34	4.02

with progressive specular bases achieves the best performance. Our analysis is that, by relaxing the training of specular roughness, the network can adjust these terms for more accurate material estimation. Hence, it also leads to more accurate lighting and normals.

Robustness on light modeling As stated above, our model shifts part of the burden of solving lightings from the neural light modeling to the later inverse rendering procedural. Hence, even if our pre-trained light model L_Ψ does not provide perfect lighting estimations, our later procedural can continue refining its estimation via the reconstruction error. To demonstrate the robustness of our model against the errors of the light model, we conduct the experiments where different levels of noise are added to the lightings, as shown in Tab. 2 and Fig. 2. In Tab. 2, the first column shows the results of our model on DiLiGenT with the pre-trained L_Ψ. From the second column, different levels of noise (noise that is up to certain degrees) are applied to the lightings. The light directions are randomly shifted, and the light intensities are all re-set to ones. Then, we further refined this noised lighting estimation via the inverse rendering procedural at the testing

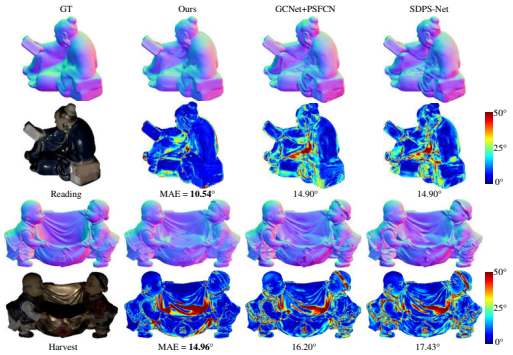


Fig. 3: Visualized comparisons of normal estimation for "Reading" and "Harvest" in DiLiGenT. Our method produces better normal estimation than others, particularly in regions with specularities (*e.g.* see the head of "reading", the golden sack of "Harvest").

stage. As illustrated in Tab. 2, even when the light directions are randomly shifted up to 70 degrees, our model still achieves comparable performance after the optimization. In Fig. 2, we visualize how our model gradually refines the lighting estimation during the course of optimization.

6.4 Evaluation on DiLiGenT benchmark

Results on normal estimation We evaluate our method on the challenging DiLiGenT benchmark and compare our method with previous works. The quantitative result on normal estimation is shown in Tab. 3a. We achieve the best average performance, and we outperform the second-best method by 1.65 degrees on average. Thanks to the proposed progressive specular bases, our method performs particularly well on those objects with specularities. There are a large number of specularities in "Reading" and "Goblet", where our method outperforms the others by 4.36 and 2.48 degrees, respectively. Figure 3 shows qualitative comparison on "Reading" and "Harvest'. Our method produces much better results on those specular regions.

Results on light estimation The quantitative results on light intensity estimation are shown in Tab. 3b. Our model achieves the best performance on average, which is 0.0365 in relative error. The results on light direction estimation are presented in Tab. 3c, where we also demonstrate a comparable result to previous methods. Figure 4 showcases the visualization of the lighting results. As LCNet [8] discretely represents the light direction into bins, their estimation looks very noisy (see lighting in "Reading"). In contrast, our model can continuously refine the lights. Hence, our lighting estimation preserves smoother pattern overall.

6.5 Evaluation on other real world dataset

We then evaluate our method on other challenging real-world datasets. Tab. 4 shows that our method achieves the best performance in lighting estimation in the Gourd&Apple dataset. In Fig. 4, We visualized the estimated lighting

Table 4: Evaluation results on Gourd&Apple dataset.

(a) Results on light intensity.

Method	Apple	Gourd1	Gourd2	Avg.
PF14[36]	0.1090	0.0960	0.3290	0.1780
LCNet[8]	0.1060	0.0480	**0.1860**	0.1130
GCNet[11]	0.0940	0.0420	0.1990	0.1120
Ours	**0.0162**	**0.0272**	0.2330	**0.0921**

(b) Results on light direction.

Method	Apple	Gourd1	Gourd2	Avg.
PF14[36]	6.68	21.23	25.87	17.92
LCNet[8]	9.31	4.07	7.11	6.83
GCNet[11]	10.91	4.29	7.13	7.44
Ours	**1.87**	**2.34**	**2.01**	**2.07**

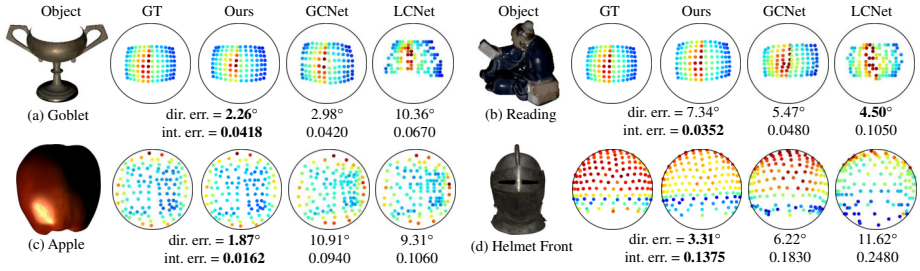


Fig. 4: Visualized comparisons of the ground-truth and estimated lighting distribution for the DiLiGenT dataset, Gourd&Apple dataset, and Light Stage dataset. It demonstrates that our lighting estimation is also robust in different datasets under different lighting distributions.

in "Apple" from Gourd&Apple dataset, and "Helmet Front" from Light Stage dataset. These results manifest that our method can also reliably recover the lighting in different light distributions. Our specular modeling is also applicable to different datasets under different materials. Please refer to supplementary material for more results.

7 Discussions and Conclusions

In this paper, we propose a neural representation for lighting and surface normal estimation via inverse rendering. The surface is explicitly modeled with diffuse and specular components, and the GBR ambiguity is resolved by fitting on these photometric cues. To avoid local minima during optimization, we propose *progressive specular bases* for fitting the specularities. Our method provides state-of-the-art performance on lighting estimation and shape recovery on challenging real-world datasets.

Limitations and future work: The inter-reflections, subsurface scattering, and image noises are not considered in our image rendering equation 6. Our model may fail if these terms are prominent on an object's surface. Explicitly modeling these terms and jointly refining them within the same framework will be an intriguing direction to pursue.

Acknowledgments This research is funded in part by ARC-Discovery grants (DP190102261 and DP220100800), a gift from Baidu RAL, as well as a Ford Alliance grant to Hongdong Li.

References

1. Alldrin, N., Zickler, T., Kriegman, D.: Photometric stereo with non-parametric and spatially-varying reflectance. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
2. Alldrin, N.G., Mallick, S.P., Kriegman, D.J.: Resolving the generalized bas-relief ambiguity by entropy minimization. In: 2007 IEEE conference on computer vision and pattern recognition. pp. 1–7. IEEE (2007)
3. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. IEEE transactions on pattern analysis and machine intelligence **37**(8), 1670–1687 (2014)
4. Belhumeur, P.N., Kriegman, D.J., Yuille, A.L.: The bas-relief ambiguity. International journal of computer vision **35**(1), 33–44 (1999)
5. Burley, B., Studios, W.D.A.: Physically-based shading at disney. In: ACM SIGGRAPH. vol. 2012, pp. 1–7. vol. 2012 (2012)
6. Chabert, C.F., Einarsson, P., Jones, A., Lamond, B., Ma, W.C., Sylwan, S., Hawkins, T., Debevec, P.: Relighting human locomotion with flowed reflectance fields. In: ACM SIGGRAPH 2006 Sketches, pp. 76–es (2006)
7. Chandraker, M.K., Kahl, F., Kriegman, D.J.: Reflections on the generalized bas-relief ambiguity. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 788–795. IEEE (2005)
8. Chen, G., Han, K., Shi, B., Matsushita, Y., Wong, K.Y.K.: Self-calibrating deep photometric stereo networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8739–8747 (2019)
9. 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)
10. Chen, G., Han, K., Wong, K.Y.K.: Ps-fcn: A flexible learning framework for photometric stereo. In: European Conference on Computer Vision. pp. 3–19. Springer (2018)
11. 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)
12. Cho, D., Matsushita, Y., Tai, Y.W., Kweon, I.S.: Semi-calibrated photometric stereo. IEEE transactions on pattern analysis and machine intelligence **42**(1), 232–245 (2018)
13. Drbohlav, O., Šára, R.: Specularities reduce ambiguity of uncalibrated photometric stereo. In: European Conference on Computer Vision. pp. 46–60. Springer (2002)
14. Georghiades, A.S.: Incorporating the torrance and sparrow model of reflectance in uncalibrated photometric stereo. In: Computer Vision, IEEE International Conference on. vol. 3, pp. 816–816. IEEE Computer Society (2003)
15. Haefner, B., Ye, Z., Gao, M., Wu, T., Quéau, Y., Cremers, D.: Variational uncalibrated photometric stereo under general lighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8539–8548 (2019)
16. Honzátko, D., Türetken, E., Fua, P., Dunbar, L.A.: Leveraging spatial and photometric context for calibrated non-lambertian photometric stereo. arXiv preprint arXiv:2103.12106 (2021)

17. Ikehata, S.: Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces. In: European Conference on Computer Vision. pp. 3–19. Springer (2018)
18. Ikehata, S., Aizawa, K.: Photometric stereo using constrained bivariate regression for general isotropic surfaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2179–2186 (2014)
19. Ikehata, S., Wipf, D., Matsushita, Y., Aizawa, K.: Robust photometric stereo using sparse regression. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 318–325. IEEE (2012)
20. Ju, Y., Dong, J., Chen, S.: Recovering surface normal and arbitrary images: A dual regression network for photometric stereo. IEEE Transactions on Image Processing **30**, 3676–3690 (2021)
21. 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)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Li, J., Li, H.: Neural reflectance for shape recovery with shadow handling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16221–16230 (2022)
24. 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)
25. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. In: IEEE International Conference on Computer Vision (ICCV) (2021)
26. Logothetis, F., Budvytis, I., Mecca, R., Cipolla, R.: Px-net: Simple and efficient pixel-wise training of photometric stereo networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12757–12766 (2021)
27. Logothetis, F., Mecca, R., Cipolla, R.: Semi-calibrated near field photometric stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 941–950 (2017)
28. Lu, F., Chen, X., Sato, I., Sato, Y.: Symps: Brdf symmetry guided photometric stereo for shape and light source estimation. IEEE transactions on pattern analysis and machine intelligence **40**(1), 221–234 (2017)
29. Lu, F., Matsushita, Y., Sato, I., Okabe, T., Sato, Y.: Uncalibrated photometric stereo for unknown isotropic reflectances. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1490–1497 (2013)
30. Lu, F., Sato, I., Sato, Y.: Uncalibrated photometric stereo based on elevation angle recovery from brdf symmetry of isotropic materials. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 168–176 (2015)
31. Matusik, W., Pfister, H., Brand, M., McMillan, L.: A data-driven reflectance model. ACM Transactions on Graphics **22**(3), 759–769 (Jul 2003)
32. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision. pp. 405–421. Springer (2020)
33. Mo, Z., Shi, B., Lu, F., Yeung, S.K., Matsushita, Y.: Uncalibrated photometric stereo under natural illumination. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2936–2945. IEEE Computer Society (2018)
34. Mukaigawa, Y., Ishii, Y., Shakunaga, T.: Analysis of photometric factors based on photometric linearization. JOSA A **24**(10), 3326–3334 (2007)

35. Papadhimitri, T., Favaro, P.: A new perspective on uncalibrated photometric stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1474–1481 (2013)
36. Papadhimitri, T., Favaro, P.: A closed-form, consistent and robust solution to uncalibrated photometric stereo via local diffuse reflectance maxima. International journal of computer vision **107**(2), 139–154 (2014)
37. Quéau, Y., Lauze, F., Durou, J.D.: Solving uncalibrated photometric stereo using total variation. Journal of Mathematical Imaging and Vision **52**(1), 87–107 (2015)
38. Quéau, Y., Wu, T., Lauze, F., Durou, J.D., Cremers, D.: A non-convex variational approach to photometric stereo under inaccurate lighting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 99–108 (2017)
39. Santo, H., Samejima, M., Sugano, Y., Shi, B., Matsushita, Y.: Deep photometric stereo network. In: Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on. pp. 501–509. IEEE (2017)
40. Sarno, F., Kumar, S., Kaya, B., Huang, Z., Ferrari, V., Van Gool, L.: Neural architecture search for efficient uncalibrated deep photometric stereo. arXiv preprint arXiv:2110.05621 (2021)
41. Shi, B., Matsushita, Y., Wei, Y., Xu, C., Tan, P.: Self-calibrating photometric stereo. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 1118–1125. IEEE (2010)
42. 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. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
43. Sunkavalli, K., Zickler, T., Pfister, H.: Visibility subspaces: Uncalibrated photometric stereo with shadows. In: European Conference on Computer Vision. pp. 251–264. Springer (2010)
44. Taniai, T., Maehara, T.: Neural inverse rendering for general reflectance photometric stereo. In: International Conference on Machine Learning. pp. 4864–4873 (2018)
45. Torrance, K.E., Sparrow, E.M.: Theory for off-specular reflection from roughened surfaces. Josa **57**(9), 1105–1114 (1967)
46. Wang, J., Ren, P., Gong, M., Snyder, J., Guo, B.: All-frequency rendering of dynamic, spatially-varying reflectance. ACM Transactions on Graphics (TOG) **28**(5), 1–10 (2009)
47. 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)
48. Woodham, R.J.: Photometric method for determining surface orientation from multiple images. Optical engineering **19**(1), 191139 (1980)
49. Wu, L., Ganesh, A., Shi, B., Matsushita, Y., Wang, Y., Ma, Y.: Robust photometric stereo via low-rank matrix completion and recovery. In: Asian Conference on Computer Vision. pp. 703–717. Springer (2010)
50. Wu, T.P., Tang, C.K.: Photometric stereo via expectation maximization. IEEE transactions on pattern analysis and machine intelligence **32**(3), 546–560 (2010)
51. Wu, Z., Tan, P.: Calibrating photometric stereo by holistic reflectance symmetry analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1498–1505 (2013)
52. Yao, Z., Li, K., Fu, Y., Hu, H., Shi, B.: Gps-net: Graph-based photometric stereo network. Advances in Neural Information Processing Systems **33** (2020)

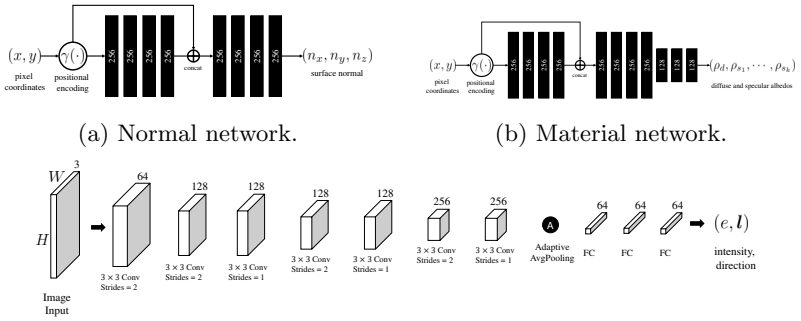
53. Zheng, Q., Jia, Y., Shi, B., Jiang, X., Duan, L.Y., Kot, A.C.: Spline-net: Sparse photometric stereo through lighting interpolation and normal estimation networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8549–8558 (2019)

Supplementary Material

A Implementation Details

Network architectures Here, we describe our network architectures in detail. In Fig. 5a, the surface normal net $N_{\Theta}(\cdot)$ uses 8 fully-connected layers with 256 channels. It takes positional encoded [32] pixel coordinates $\gamma(x), \gamma(y)$ as input, directly output the surface normal at that position $\boldsymbol{n} = [n_x, n_y, n_z]^T$. As shown in Fig. 5b, the material net $M_{\Phi}(\cdot)$ uses the same structure but with 3 more fully-connected layers than normal net. It takes the same positional encoded pixel coordinates input and outputs the diffuse and specular albedos of the surface point. As shown in Fig. 5c, the lighting network $L_{\Psi}(\cdot)$ consists of 7 convolutional ReLU layers and 3 fully connected layers. It takes the image with size $H \times W \times 3$ as input, directly outputs the light intensity e and light direction $\boldsymbol{l}$ of that image.

Early supervision Following previous works [3,23], we additionally use the surface smoothness constraints and shape-from-contour priors as the early supervision in our network. After early-stage training in the first half iterations, we discard these priors and train the network via photometric loss.



(a) Normal network. (b) Material network.

(c) Lighting network.

Fig. 5: The network architectures of our networks.

B Visualization of the effectiveness of PSB

Recall that, although the GBR ambiguity can be reduced up to a binary concave/convex ambiguity under our model, there is no guarantee that no local minima exist during the optimization. To effectively avoid the local minimas during the optimization, we propose the progressive specular bases (PSB) for the network. In Fig. 6, we provide the visual comparison between the model with PSB and the model without PSB. The first row displays the observed ground truth image under a light source, the ground truth light distribution, and the ground truth surface normal. The second row and third row display the reconstructed image, the estimated light distribution, the estimated normal, the error map of estimated normal, and the estimated shape from our "with PSB model" and "without PSB model" respectively.

As we can see, the "without PSB" produces a worse light and normal estimation. Both the light and normal are "shifted" along the z axis. However, its reconstructed image still presents a similar quality to the observed ground truth (PSNR: 40.06dB). This observation coincides with the observation from Belhumeur [4], where they also observed that the differences in shape are hard to be discerned from the frontal images given a small scale along the z axis.

The PSB can provide prior information to the network and limit the space of possible solutions by forcing the network to fit on the shiny specularities first in the early stage of optimization. Hence, by applying with the PSB, even with a poor network initialization, our network can still effectively avoid the local minimas to achieve better results.

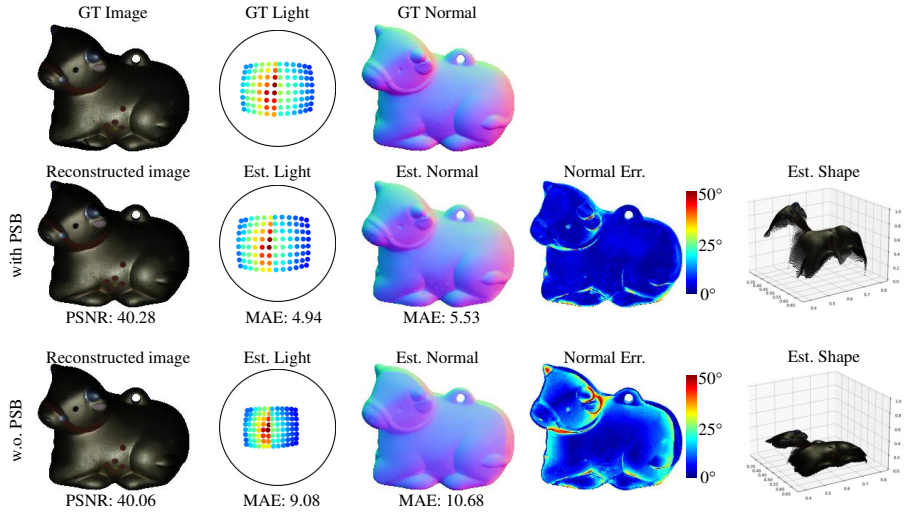


Fig. 6: Visualized comparisons of with/without using the *progressive specular basis* (PSB).

C Ablation study on lighting model

In this section, we conduct two experiments to showcase the effectiveness of the lighting network. As shown in Fig. 7: on the first row, we showcase the observed(GT) image, ground truth lights and normals; on the second row, we display reconstructed image and estimated result using the model with lighting network $L_\Psi(\cdot)$; on the third row, we present results using the model without lighting network and takes randomized lights as initialization.

Without using the lighting network, we take randomized lights as initialization. Our network may sometimes produce a flipped surface as a result, as shown in the third row in Fig. 7. As we can see in the second row and third row in Fig. 7, the estimated lights and normals are flipped in the x, y axis. In the third row, the mean angular error (MAE) for light direction is 55.47 degrees, and normal error is 91.07 degrees. However, its reconstructed image is almost identical to the observed ground truth image.

During the experiments, we observed that this convex/concave ambiguity can be easily resolved by providing the model with a coarse lighting estimation, as shown in second row in Fig. 7. Our lighting model $L_\Psi(\cdot)$ can provide a coarse lighting estimation as the starting point, which is sufficient to for the followed self-supervised network to further refine the coarse results and produce the correct lights.

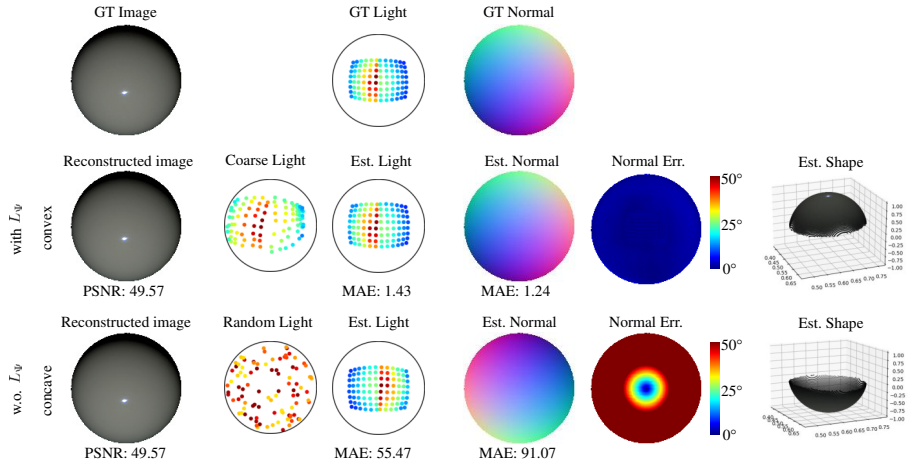


Fig. 7: Visualization of the effectiveness of lighting model.

D Robustness on Sparse Inputs

In this section, we present the results on DiLiGenT [42] dataset with only 16 images at the inputs. Following previous works on sparse inputs for photometric stereo [24], we selected 16 images as input for our method and others for comparison. The errors are shows in Tab. 5. As we can see from the table, our method still outperform the state-of-the-art with only 16 images. Besides, with 16 images as input, our method only drop 0.72 degrees in MAE in normal estimation, while GCNet[11]+PSFCN[10] drops 2.04 degrees in MAE. It also demonstrate that our method is robust against sparse input.

Table 5: Quantitative comparison on DiLiGenT with only 16 images at input.

(a) MAE of surface normal.

model	All images	16 images
Ours	7.05	7.77
GCNet[11]+PSFCN[10]	8.70	10.74

(b) Scale-invariant relative error of light intensities.

model	All images	16 images
Ours	0.0365	0.0548
GCNet[11]	0.0519	0.0550

(c) MAE of light directions.

model	All images	16 images
Ours	4.02	5.02
GCNet[11]	3.32	4.04

E Results on DiLiGenT benchmark

In this section, we present the results on DiLiGenT [42] dataset, as shown in the following Fig. 10, 11 and 12. For each object, the first row displays the ground truth lighting, ours estimated lighting, and lighting results from GCNet [11] and SDPS-Net [8]. The second row displays the ground truth surface normal and estimated surface normal by ours and competing methods. The last row display the observed image and the error map of the estimated surface normal. We also present the quantitative evaluation for lighting and normal below the lighting and error map. Note that UPS-FCN [10] can not estimate the lighting.

Results on almost Lambertian surface As we can see from the results, our method works well for specular objects, as well as objects that appear to be very diffuse, such as "Cat". In order to better understand why our method also works well on objects like "Cat", we visualized the reconstructed terms ρ_d and ρ_s in Fig. 8. Figure 8 shows that the "Cat" is not purely diffuse and contains very soft specularities. Our method is able to capture and use these soft specularities as clues for estimating the surface normal.

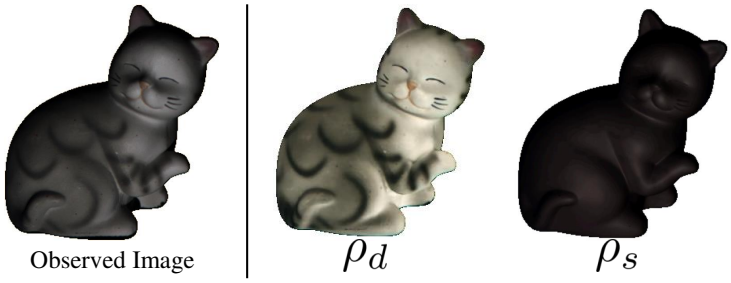


Fig. 8: Visualization of reconstructed ρ_d and ρ_s in object "Cat".

F Results on Apple&Gourd dataset

In this section, we present the results on Apple&Gourd [1] dataset. In Fig. 13, for each object, the first row displays the ground truth lighting, ours estimated lighting, and lighting results from GCNet [11]. The second row displays the observed image and estimated surface normal by ours and competing methods. Note that there is no ground truth surface normal available in this dataset, so we only visualized compare the normal results. As shown in "Gourd2", it is clear that our estimated normal present higher quality than previous state-of-the-art method GCNet [11]+PSFCN [10].

G Results on synthetic dataset with 100 MERL BRDFs

To evaluation our method across different surface materials and BRDFs, we test our method on a publicly available synthetic dataset[2]: GCNet-Synthetic [11]. The dataset consists of two rendered synthetic objects: Dragon and Armadillo for testing. This dataset was rendered with 100 MERL [31] BRDFs under 82 random light directions using physically based renderer Mitsuba[3].

We showcase the results in Fig. 14 and Fig. 15. As we can see from the figures, our method produce comparable results to GCNet [11].

[2] `https://github.com/guanyingc/UPS-GCNet`
[3] `http://mitsuba-renderer.org/`

We dive into the MERL dataset and found that our method fails to fit the materials such as "steel", "chrome", and "chrome-steel", where they generally present asymmetric highlights as shown in Fig. 9. Prior work [5] believed that these anomaly asymmetric highlights could be caused by the lens flare. We believe that using a different BRDF model to account for these effects can improve the performance on these materials. We are happy to consider this as a future direction.

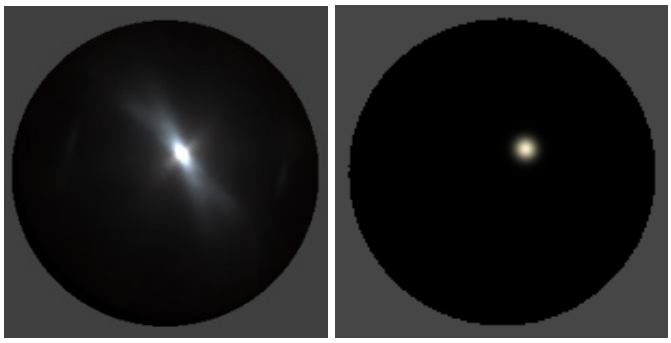


Fig. 9: Rendered sphere of "steel". Left is the data from MERL. Right is our estimated result.

H Self-captured images outside of the laboratory

We captured 55 images with a Nikon camera and a handheld flashlight. The target object is captured in a regular livingroom environment with lights off. The captured image and our estimated results (normals, shadings, and lights) are shown in Fig. 16. As we can see from the results, our method still performs very well in a non-laboratory environment.

I Future works

We believe that our method, with some adaptations, can be extended to solve the problem under many other assumptions, such as specularity detection, multi-view photometric stereo, photometric stereo under multi-light-sources and natural illumination. Our method inverse renders the object to shapes and materials. Hence, the specularity detection is also available at output, as shown in Fig. 8. A possible adaptation for multi-view photometric stereo is to apply our algorithm to each view of the object, and then fuse the normal map from different views to obtain the full geometry. We can also model the environment map as Spherical-Gaussians to enable fast integration of BRDF and lighting in natural-illumination and multi-light-sources.

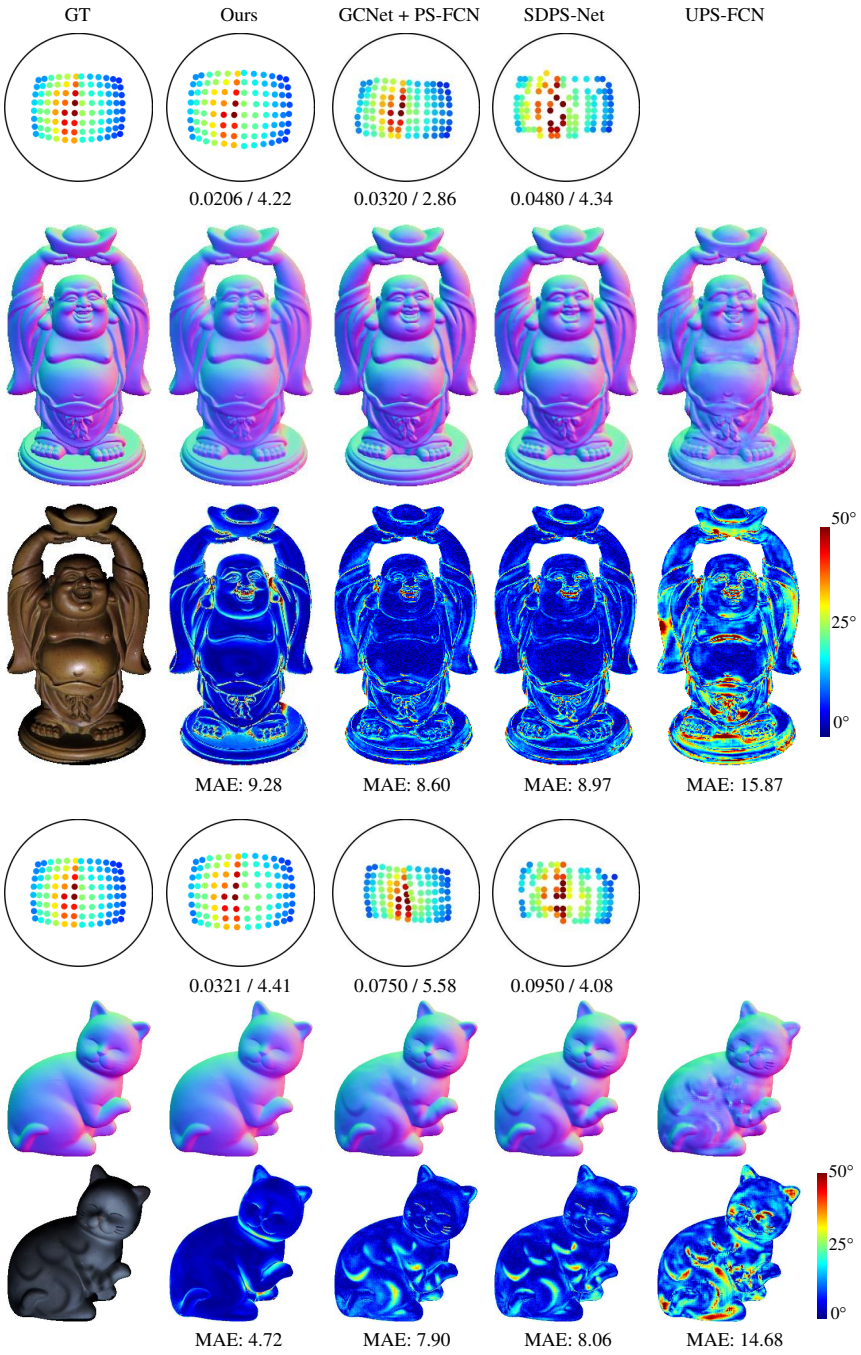


Fig. 10: Results for "Buddha" and "Cat" from DiLiGenT dataset.

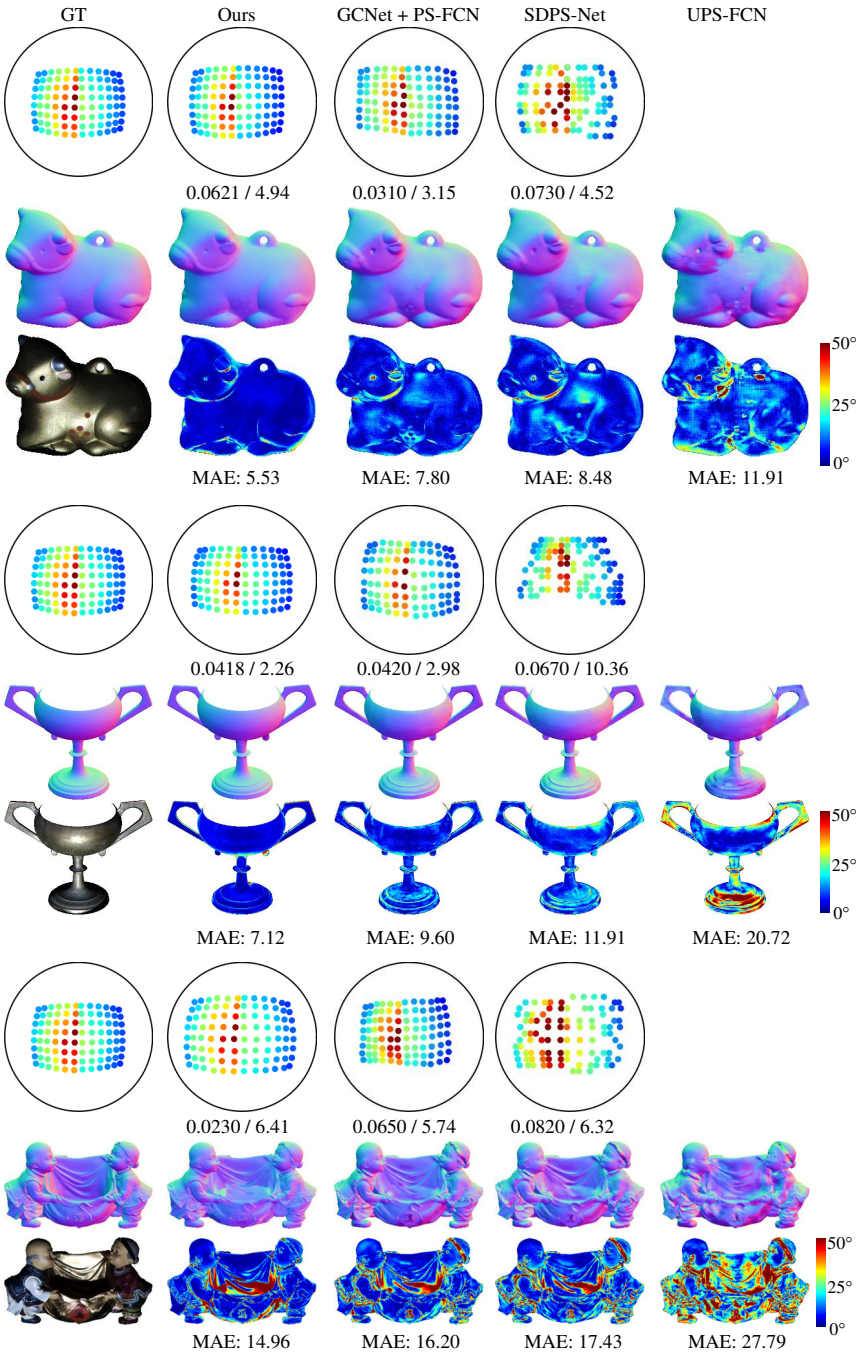


Fig. 11: Results for "Cow" , "Goblet", and "Harvest" from DiLiGenT dataset.

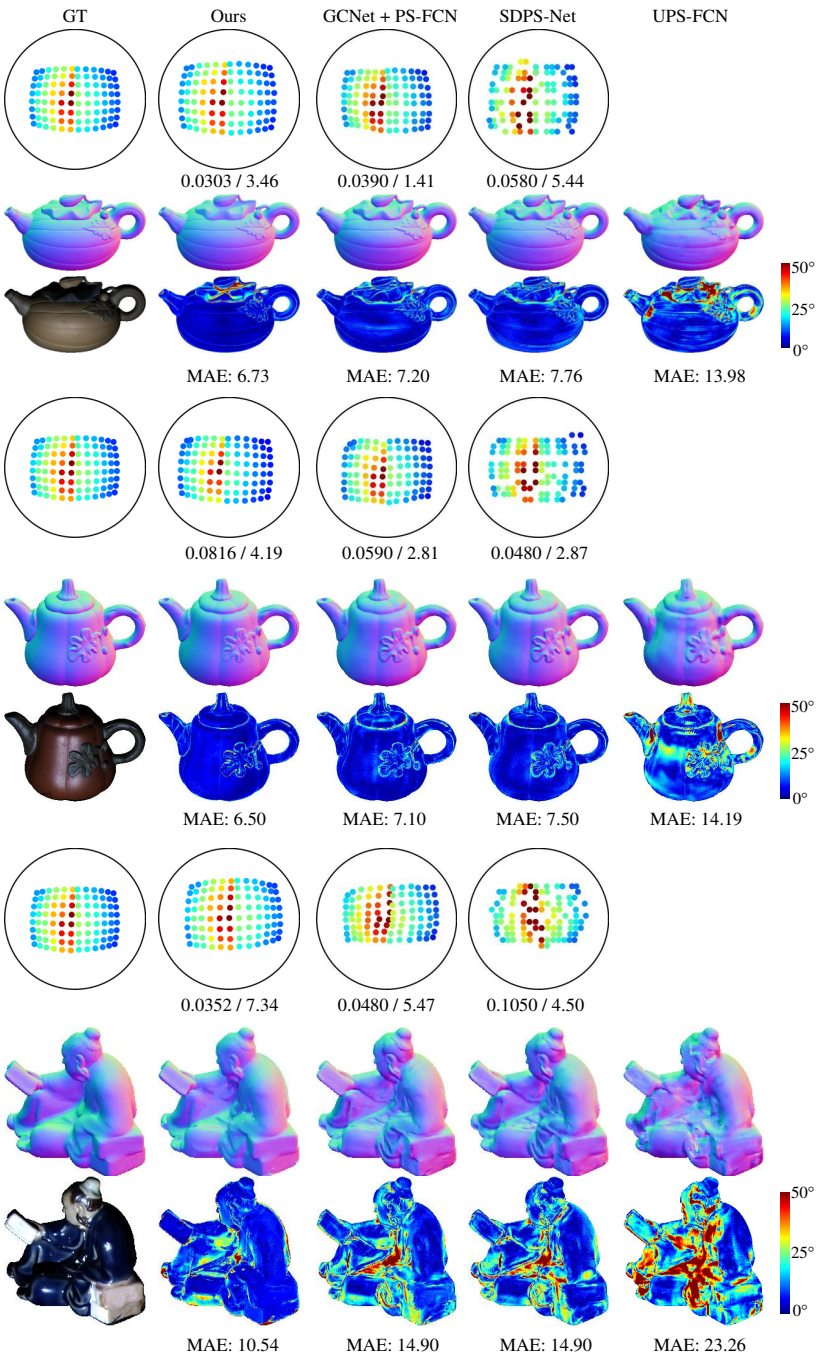


Fig. 12: Results for "Pot1" , "Pot2", and "Reading" from DiLiGenT dataset.

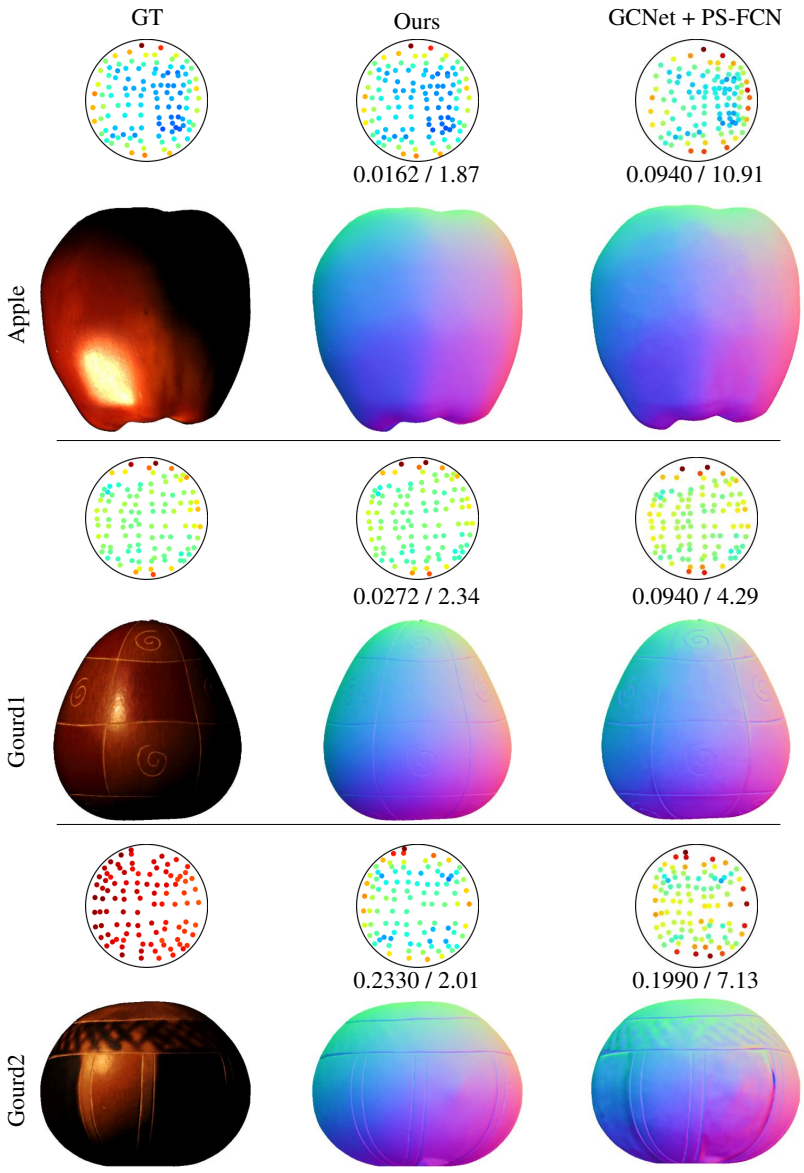
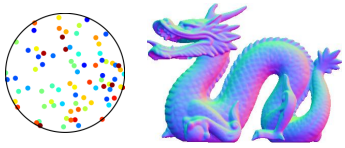
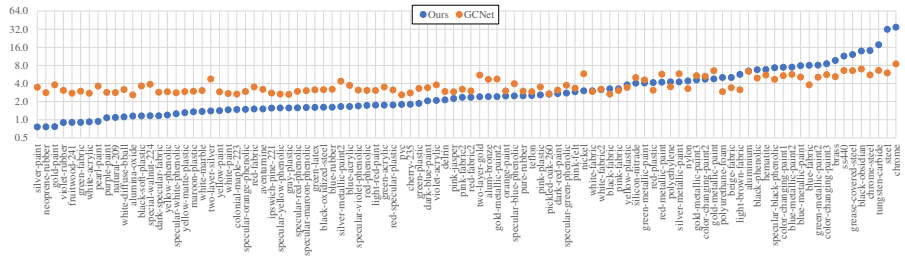


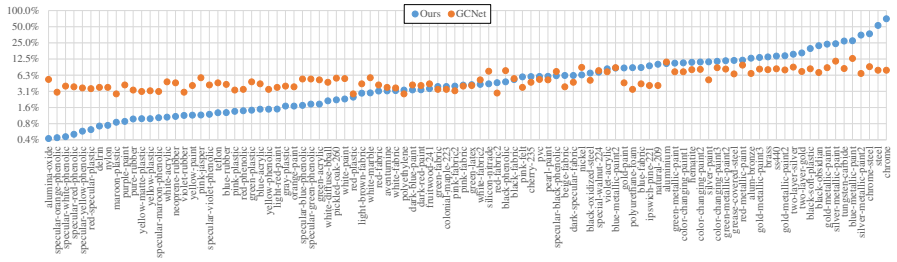
Fig. 13: Results for "Apple" , "Gourd1", and "Gourd2" from Apple&Gourd dataset.



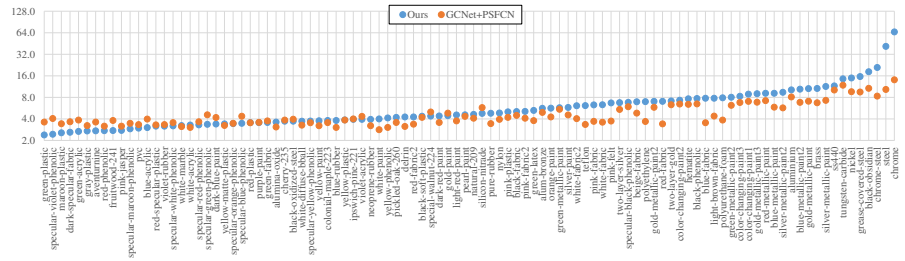
(a) Ground truth of lights and surface normals.



(b) MAE of light directions.

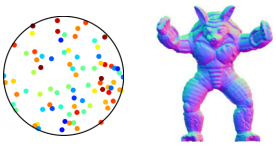


(c) Scale-invariant relative error in percentage of light intensities.

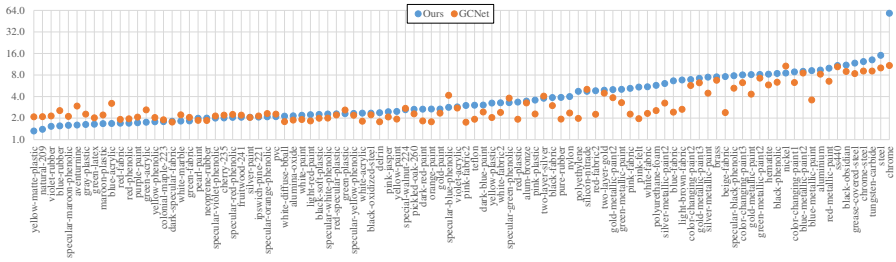


(d) MAE of surface normals.

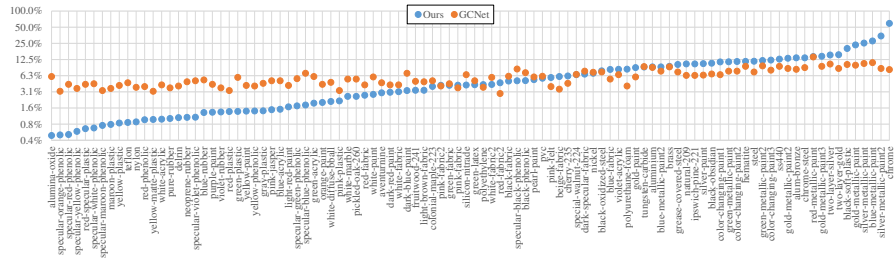
Fig. 14: Comparison on object "Dragon" rendered with 100 MERL BRDFs.



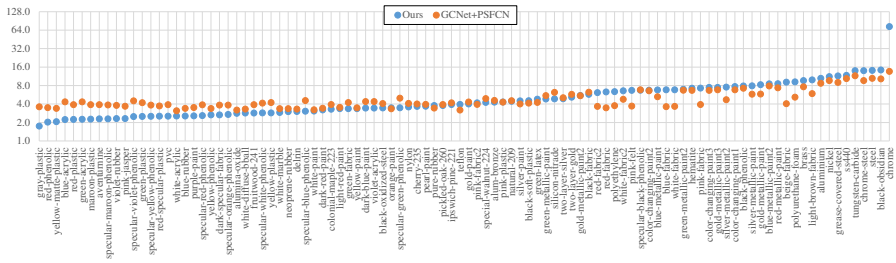
(a) Ground truth of lights and surface normals.



(b) MAE of light directions.



(c) Scale-invariant relative error in percentage of light intensities.



(d) MAE of surface normals.

Fig. 15: Comparison on object "Armadillo" rendered with 100 MERL BRDFs.

Fig. 16: The captured image of “CokeCan” and our estimations.