

What is Learned in Deep Uncalibrated Photometric Stereo?

Guanying Chen¹ Michael Waechter² Boxin Shi^{3,4}
Kwan-Yee K. Wong¹ Yasuyuki Matsushita²

¹The University of Hong Kong ²Osaka University ³Peking University
⁴Peng Cheng Laboratory

Abstract. This paper targets at discovering what a deep uncalibrated photometric stereo network learns to resolve the problem’s inherent ambiguity, and designing an effective network architecture based on the new insight to improve the performance. The recently proposed deep uncalibrated photometric stereo method achieved promising results in estimating directional lightings. However, what specifically inside the network contributes to its success remains a mystery. In this paper, we analyze the features learned by this method and find that they strikingly resemble attached shadows, shadings, and specular highlights, which are known to provide useful clues in resolving the generalized bas-relief (GBR) ambiguity. Based on this insight, we propose a guided calibration network, named *GCNet*, that explicitly leverages object shape and shading information for improved lighting estimation. Experiments on synthetic and real datasets show that GCNet achieves improved results in lighting estimation for photometric stereo, which echoes the findings of our analysis. We further demonstrate that GCNet can be directly integrated with existing calibrated methods to achieve improved results on surface normal estimation. Our code and model can be found at <https://guanyingc.github.io/UPS-GCNet>.

Keywords: Uncalibrated photometric stereo, generalized bas-relief ambiguity, deep neural network

1 Introduction

Photometric stereo aims at recovering the surface normals of a scene from single-viewpoint imagery captured under varying light directions [50, 47]. In contrast to multi-view stereo [41], photometric stereo works well for textureless surfaces and can recover highly detailed surface geometry.

Following the conventional assumption, this paper assumes the scene is illuminated by a single light direction in each image. Most existing photometric stereo methods [46, 23, 22] require *calibrated* light directions as input. *Uncalibrated* photometric stereo, on the other hand, simultaneously estimates light directions and surface normals. In multi-view stereo, this problem of auto-calibration (*i.e.*, calibration from images of the scene without the use

of any explicit calibration targets) has been solved satisfactorily on “wild” imagery such as those from Internet photo sharing sites [3]. Auto-calibration for photometric stereo is without a doubt an important goal since it makes photometric stereo applicable to wild data [43] and useful for amateurs who know nothing about tedious calibration procedures [42, 1, 39]. Existing methods for uncalibrated photometric stereo [4, 35] often assume a Lambertian reflectance model and their focus has been on resolving the generalized bas-relief (GBR) ambiguity [6]. Manifold embedding based methods [40, 29] can deal with surfaces with general isotropic reflectances, but they rely on a roughly uniform lighting distribution which is usually not satisfied in real-world datasets.

Despite the impressive results on complex reflectances reported by recent deep learning methods for calibrated photometric stereo [38, 49, 10, 22, 27, 53], not much work has been done on learning-based uncalibrated photometric stereo. Recently, Chen *et al.* [8] introduced a deep uncalibrated photometric stereo network, called Lighting Calibration Network (LCNet), to estimate light directions and intensities from input images, and a normal estimation network to predict surface normals. Compared with UPS-FCN [10] which directly estimates surface normals from images, Chen *et al.*’s two-stage approach achieves considerably better results. However, the features learned by LCNet to resolve the ambiguity in lighting estimation remain unknown.

This paper focuses on demystifying the problem of how deep uncalibrated photometric stereo learns to resolve the GBR ambiguity, and how to improve it for higher accuracy in lighting estimation. Our contributions are:

- We discuss the differences between the learning-based LCNet [8] and traditional uncalibrated methods, and analyze the features learned by LCNet to resolve the GBR ambiguity.
- We find that attached shadows, shadings, and specular highlights are key elements for lighting estimation, and that LCNet extracts features independently from each input image without exploiting any inter-image information (“inter-image” means information shared by all images).
- Based on our findings, we propose a guided calibration network (GCNet) that explicitly utilizes object shape and shading information as guidances for better lighting estimation.

2 Related Work

In this section, we briefly review recent deep learning methods for calibrated photometric stereo and existing methods for uncalibrated photometric stereo. Readers are referred to [20, 2, 45] for more comprehensive surveys of photometric stereo algorithms. In the rest of this paper, we will use “lighting” to refer to light direction and light intensity.

Deep calibrated photometric stereo Recently, deep learning methods have been proposed in the context of photometric stereo. Compared with traditional methods that often adopt a simplified reflectance model, learning-based methods can directly learn the mapping from observations to surface normals and

achieve state-of-the-art results on a real-world benchmark [45] with complex reflectances. Santo *et al.* [38] first introduced a fully-connected deep photometric stereo network to estimate pixel-wise surface normals from a fixed number of observations. To handle a variable number of input images in an order-agnostic manner, Ikehata [22] proposed a fixed shape observation map representation, while Chen *et al.* [9] adopted an element-wise max-pooling operation to fuse features stemming from multiple inputs. Li *et al.* [27] and Zheng *et al.* [53] focused on reducing the number of required images while maintaining similar accuracy under the framework proposed by Ikehata [22]. Different from the above supervised methods that require a synthetic dataset for training, Tani and Maehara [49] proposed an unsupervised framework to estimate surface normals via an on-line optimization process.

Uncalibrated photometric stereo Most existing uncalibrated photometric stereo methods are based on matrix factorization (*e.g.*, singular value decomposition) and assume a Lambertian reflectance model. A Lambertian surface’s normals can be recovered up to a 3×3 linear ambiguity when light directions are unknown [19]. By considering the surface integrability constraint, this linear ambiguity can be reduced to a 3-parameter generalized bas-relief (GBR) ambiguity [15, 6, 52, 26]. To further resolve the GBR ambiguity, many methods make use of additional clues like inter-reflections [7], specularities [13, 16, 12], albedo priors [4, 44], isotropic reflectance symmetry [48, 51], special light source distributions [54], or Lambertian diffuse reflectance maxima [35].

Manifold embedding methods [40, 34, 29, 30] can handle surfaces with general isotropic reflectance based on the observation that the distance between two surface points’ intensity profiles is closely related to their surface normals’ angular difference. However, these methods often assume a uniform lighting distribution. Other methods related to uncalibrated photometric stereo include exemplar-based methods [21], regression-based methods [32], semi-calibrated photometric stereo [11], inaccurate lighting refinement [37], and photometric stereo under general lighting [5, 33, 18].

Recently, Chen *et al.* [8] introduced a Lighting Calibration Network (LCNet) to estimate lightings from images and then estimate surface normals based on the lightings. This two-stage method achieves considerably better results than the single-stage method [10]. It also has slightly better interpretability because the lightings estimated in the first stage can be visualized. However, the features learned by LCNet to estimate lightings remain unknown.

3 Learning for Lighting Calibration

In this section, we discuss the inherent ambiguity in uncalibrated photometric stereo of Lambertian surfaces, the fact that it can be resolved for non-Lambertian surfaces, and the features learned by LCNet [8] to resolve such ambiguity.

Lambertian surfaces and the GBR ambiguity When ignoring shadows (*i.e.*, attached and cast shadows) and inter-reflections, the image formation of a

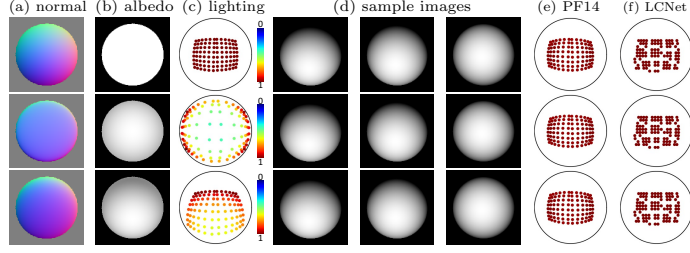


Fig. 1. Row 1 is the true shape of a *Sphere*, while rows 2 and 3 are shapes under two different GBR transformations. In column (c), the points’ positions and colors indicate light direction and relative intensity, respectively. Columns (e) and (f) show the lightings estimated by PF14 [35] and LCNet [8].

Lambertian surface with P pixels captured under F lightings can be written as

$$\mathbf{M} = \mathbf{N}^\top \mathbf{L}, \quad (1)$$

where $\mathbf{M} \in \mathbb{R}^{P \times F}$ is the measurement matrix. $\mathbf{N} \in \mathbb{R}^{3 \times P}$ is the surface normal matrix whose columns are albedo scaled normals $\mathbf{N}_{:,p} = \rho_p \mathbf{n}_p$, where ρ_p and \mathbf{n}_p are the albedo and unit-length surface normal of pixel p . $\mathbf{L} \in \mathbb{R}^{3 \times F}$ is the lighting matrix whose columns are intensity scaled light directions $\mathbf{L}_{:,f} = e_f \mathbf{l}_f$, where e_f and \mathbf{l}_f are the light intensity and unit-length light direction of image f .

By matrix factorization and applying the surface integrability constraint, \mathbf{N} and \mathbf{L} can be recovered up to an unknown 3-parameter GBR transformation \mathbf{G} [6] such that $\mathbf{M} = (\mathbf{G}^{-\top} \mathbf{N})^\top (\mathbf{G} \mathbf{L})$. This GBR ambiguity indicates that there are infinitely many combinations of albedo ρ , normal \mathbf{n} , light direction \mathbf{l} , and light intensity e that produce the same appearance \mathbf{M} (see Fig. 1 (a)-(d)):

$$\hat{\rho} = \rho |\mathbf{G}^{-\top} \mathbf{n}|, \quad \hat{\mathbf{n}} = \frac{\mathbf{G}^{-\top} \mathbf{n}}{|\mathbf{G}^{-\top} \mathbf{n}|}, \quad \hat{\mathbf{l}} = \frac{\mathbf{G} \mathbf{l}}{|\mathbf{G} \mathbf{l}|}, \quad \hat{e} = e |\mathbf{G} \mathbf{l}|. \quad (2)$$

Although the surface’s appearance remains the same after GBR transformation (*i.e.*, $\hat{\rho} \hat{\mathbf{n}}^\top \hat{\mathbf{l}} \hat{e} = \rho \mathbf{n}^\top \mathbf{l} e$, see Fig. 1 (d)), a surface point’s albedo will be scaled by $|\mathbf{G}^{-\top} \mathbf{n}|$. As a result, the albedo of an object will change gradually and become spatially-varying. Because this kind of spatially-varying albedo distribution resulting from GBR transformations rarely occurs on real world objects, some previous methods make explicit assumptions on the albedo distribution (*e.g.*, constant albedo [6, 35] or low entropy [4]) to resolve the ambiguity.

PF14 [35], a state-of-the-art non-learning uncalibrated method [45], detects Lambertian diffuse reflectance maxima (*i.e.*, image points satisfying $\mathbf{n}^\top \mathbf{l} = 1$) to estimate \mathbf{G} ’s 3 parameters. We will later use it as a non-learning benchmark in our comparative experiments.

LCNet and the GBR ambiguity LCNet [8] is a state-of-the-art lighting calibration network for uncalibrated photometric stereo (see Fig. 2). Figure 1 (e)-(f) compare the results of LCNet and PF14 on surfaces that differ by GBR transformations. Since the input images are the same in all cases, LCNet estimates the

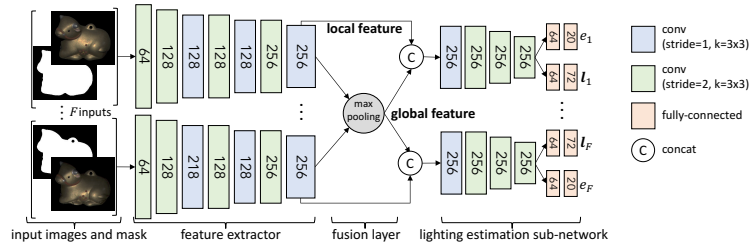


Fig. 2. Network architecture of LCNet [8]. Each layer’s value indicates its output channel number. LCNet first extracts a local feature for each input with a shared-weight feature extractor. All local features are aggregated by element-wise max-pooling to produce the global feature, each local feature is concatenated with the global feature, and is fed into a shared-weight lighting estimation sub-network to estimate a 3D light direction \mathbf{l} and a scalar light intensity e for each image.

Table 1. Light direction estimation results of PF14 [35] and LCNet [8] on a *Sphere* rendered with different BRDF types. Non-Lambertian BRDFs are taken from the MERL dataset [31].






						
model	Lambertian	fabric	plastic	phenolic	metallic	avg.
PF14	7.19	14.26	28.04	47.96	31.12	25.7
LCNet	5.38	4.07	3.08	3.05	4.09	3.93

Table 2. Light direction estimation results of LCNet [8] trained with different inputs. Values indicate mean angular error in degree.

model input	<i>Sphere</i>	<i>Bunny</i>	<i>Dragon</i>	<i>Armadillo</i>
images	3.03	4.88	6.30	6.37
(a) attached shadows	3.50	5.07	9.78	5.22
(b) specular component	2.53	6.18	7.33	4.08
(c) shading	2.29	3.95	4.64	3.76
(a) + (b) + (c)	1.87	2.06	2.34	2.12

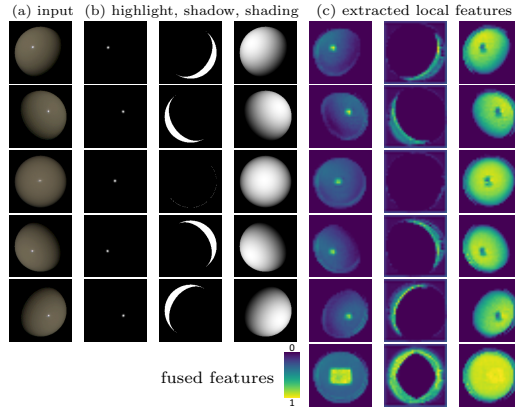
same lightings in all cases, namely the most likely lightings for the input images. The same also applies to PF14. Although LCNet’s result does not exactly equal the lightings for uniform albedo, we note that it learned from the training data that GBR-transformed surfaces are unlikely.

Although uncalibrated photometric stereo has an intrinsic GBR ambiguity for Lambertian surfaces, it was shown that GBR transformations do not preserve specularities [6, 16, 12]. Hence, specularities are helpful for ambiguity-free lighting estimation. However, traditional methods often treat non-Lambertian observations as outliers, and thus fail to make full use of specularities for disambiguation [35]. In contrast, learning-based methods can learn the relation between specular highlights and light directions through end-to-end learning. As shown in Table 1, LCNet achieves good results for non-Lambertian surfaces while PF14 completely fails when non-Lambertian observations dominate.

Feature analysis for LCNet To analyze the features learned by LCNet, we first visualize the learned local and global features. Figure 3 shows 3 representative features selected from 256 feature maps extracted by LCNet from images of a non-Lambertian *Sphere* dataset¹. Comparing Fig. 3’s Column 2 with Column

¹ Please refer to our supplemental material for more visualizations.

Fig. 3. Feature visualization of LCNNet on a non-Lambertian *Sphere*. Column 1: 5 of the 96 input images; Columns 2–4: Specular highlight centers, attached shadows, and shading rendered from ground truth; Columns 5–7: 3 of LCNNet’s 256 features maps. The last row shows the global features produced by fusing local features with max-pooling. All features are normalized to $[0, 1]$ and color coded.



5, Column 3 with Column 6, and Column 4 with Column 7, we can see that some feature maps are highly correlated with attached shadows (regions where the angle $\angle(\mathbf{n}, \mathbf{l}) \geq 90^\circ$), shadings ($\mathbf{n}^\top \mathbf{l}$), and specular highlights (regions where \mathbf{n} is close to the half angle $\mathbf{h} = \frac{\mathbf{l} + \mathbf{v}}{|\mathbf{l} + \mathbf{v}|}$ of \mathbf{l} and viewing direction \mathbf{v}). As discussed earlier, these provide strong clues for resolving the ambiguity.

To further verify our observations, we did the following. We computed (a) attached shadows, (b) the “specular components” (with a bit of concept abuse, we denote $\mathbf{h}^\top \mathbf{n}$ as specular component), and (c) shadings for the publicly available synthetic Blobby and Sculpture datasets [10] from their ground-truth light directions and normals. We then trained 4 variants of the LCNNet, taking (a), (b), (c), and (a) + (b) + (c), respectively, as input instead of regular images. We compared these 4 variant networks with LCNNet (*i.e.*, the network trained with Blobby and Sculpture images) on a synthetic test dataset introduced in Sec. 5.1. Similar to LCNNet, the variant networks also took the object mask as input. Table 2 shows that the variant models achieve results comparable to or even better than the model trained on regular images.

We can see that shadings contribute more than attached shadows and specular components for lighting estimation. This is because shading information actually includes attached shadows (*i.e.*, pixels with a zero value in the shading for synthetic data), and can be considered as an image with a uniform albedo. The uniform albedo constraint is a well-known clue for resolving the GBR ambiguity [6, 35]. In practice, attached shadows, shadings, and the specular components are not directly available as input, but this confirms our assumption that they provide strong clues for accurate lighting estimation.

4 Guided Calibration Network

In this section, we present the motivations for our guided calibration network (GCNet) and detail its structure.

4.1 Guided feature extraction

As we have seen, features like attached shadows, shadings, and specularities are important for lighting estimation, and a lighting estimation network may benefit greatly from being able to estimate them accurately. We further know that these features are completely determined by the light direction for each image as well as the inter-image shape information derived from the surface normal map. However, LCNet extracts features independently from each input image and thus cannot exploit any inter-image information during feature extraction. This observation also indicates that simply increasing the layer number of LCNet’s shared-weight feature extractor cannot produce significant improvement.

Surface normal as inter-image guidance Intuitively, if we can provide such inter-image shape information as input to the network to guide the feature extraction process, it should be able to perform better. This, however, constitutes a chicken-and-egg problem where we require normals and lightings for accurate feature extraction but at the same time we require these features for estimating accurate lightings. We therefore suggest a cyclic network structure in which we first estimate initial lightings, and then use them to estimate normals as inter-image information to guide the extraction of local (*i.e.*, per-image) features to ultimately estimate final lightings. An alternative idea might be directly estimating surface normals from images. However, previous work (UPS-FCN [10]) shows that surface normals estimated directly from images are not accurate.

Shading as intra-image guidance Another advantage of first estimating initial lighting and surface normals is that we can easily compute coarse attached shadows, shadings, or specular components as intra-image guidance for the feature extraction process (intra-image means the information is different for each image). As shading information already includes attached shadows, and not all materials exhibit specular highlights, we only compute the shading for each image as the dot-product of the estimated lighting with the surface normals, and use it as intra-image guidance. We experimentally verified that additionally including the specular component as network input does not improve results. The computed shading, on the other hand, does improve results and can assist the network to extract better features.

4.2 Network architecture

As shown in Fig. 4, the proposed GCNet consists of two lighting estimation sub-networks (L-Net) and a normal estimation sub-network (N-Net). The first L-Net, “L-Net₁”, estimates initial lightings given the input images and object masks. The N-Net then estimates surface normals from the lightings estimated by L-Net₁ and the input images. Finally, the second L-Net, “L-Net₂”, estimates more accurate lightings based on the input images, object masks, the estimated normals, and the computed shadings.

L-Net The L-Net is designed based on LCNet [8] but has less channels in the convolutional layers to reduce the model size (see Fig. 4 (a)). Compared to LCNet’s 4.4×10^6 parameters, each L-Net has only 1.78×10^6 parameters.

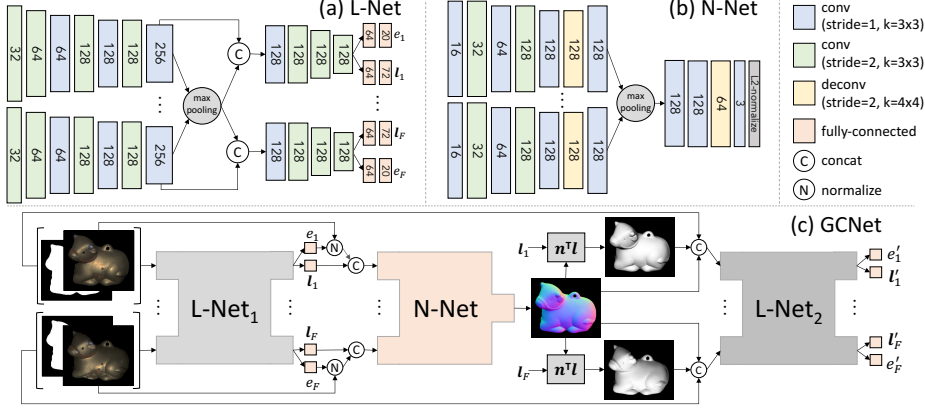


Fig. 4. Structure of (a) the lighting estimation sub-network L-Net, (b) the normal estimation sub-network N-Net, and (c) the entire GCNet. Values in layers indicate the output channel number.

Following LCNet, we discretize the lighting space and treat lighting estimation as a classification problem. Specifically, L-Net’s output light direction and intensity are in the form of softmax probability vectors (a 32-vector for elevation, a 32-vector for azimuth, and a 20-vector for intensity). Given F images, the loss function for L-Net is

$$\mathcal{L}_{\text{light}} = \frac{1}{F} \sum_f (\mathcal{L}_{l_a}^f + \mathcal{L}_{l_e}^f + \mathcal{L}_e^f), \quad (3)$$

where $\mathcal{L}_{l_a}^f$, $\mathcal{L}_{l_e}^f$, and \mathcal{L}_e^f are the cross-entropy loss for light azimuth, elevation, and intensity classifications for the f^{th} input image, respectively. For example,

$$\mathcal{L}_{l_a}^f = - \sum_{i=1}^{32} \{y_i^f = 1\} \log(p_i^f), \quad (4)$$

where $\{\cdot\}$ is a binary indicator (0 or 1) function, y_i^f is the ground-truth label (0 or 1) and p_i^f is the predicted probability for bin i (32 bins in our case) for the f^{th} image. The output probability vectors can be converted to a 3-vector light direction \mathbf{l}_f and a scalar light intensity e_f by taking the probability vector’s expectation, which is differentiable for later end-to-end fine-tuning.

L-Net₁ and L-Net₂ differ in that L-Net₁ has 4 input channels (3 for the image, 1 for the object mask) while L-Net₂ has 8 (3 additional channels for normals, 1 for shading).

N-Net The N-Net is designed based on PS-FCN [10] but with less channels, resulting in 1.1×10^6 parameters compared to PS-FCN’s 2.2×10^6 parameters

(see Fig. 4 (b) for details). Following PS-FCN, the N-Net’s loss function is

$$\mathcal{L}_{\text{normal}} = \frac{1}{P} \sum_p (1 - \mathbf{n}_p^\top \tilde{\mathbf{n}}_p), \quad (5)$$

where P is the number of pixels per image, and \mathbf{n}_p and $\tilde{\mathbf{n}}_p$ are the predicted and the ground-truth normal at pixel p , respectively.

End-to-end fine-tuning We train L-Net₁, N-Net, and L-Net₂ one after another until convergence and then fine-tune the entire network end-to-end using the following loss

$$\mathcal{L}_{\text{fine-tune}} = \mathcal{L}_{\text{light}_1} + \mathcal{L}_{\text{normal}} + \mathcal{L}_{\text{shading}} + \mathcal{L}_{\text{light}_2}, \quad (6)$$

where $\mathcal{L}_{\text{light}_1}$ and $\mathcal{L}_{\text{light}_2}$ denote the lighting estimation loss for L-Net₁ and L-Net₂. The shading loss term $\mathcal{L}_{\text{shading}} = \frac{1}{FP} \sum_f \sum_p (\mathbf{n}_p^\top \mathbf{l}_f - \tilde{\mathbf{n}}_p^\top \tilde{\mathbf{l}}_f)^2$ is included to encourage better shading estimation, and \mathbf{l}_f and $\tilde{\mathbf{l}}_f$ denote the light direction predicted by L-Net₁ and ground-truth light direction for the f^{th} image.

Training details Following LCNet [8], we trained the networks on the publicly available synthetic Blobby and Sculpture Dataset [10] which contains 85,212 surfaces, each rendered under 64 random light directions.

First, we train L-Net₁ from scratch for 20 epochs, halving the learning rate every 5 epochs. Second, we train N-Net using ground-truth lightings and input images following PS-FCN’s training procedure [10], and then retrain N-Net given the lightings estimated by L-Net₁ for 5 epochs, halving the learning rate every 2 epochs. Third, we train L-Net₂ given the input images, object masks, estimated normals, and computed shadings for 20 epochs, halving the learning rate every 5 epochs. The initial learning rate is 0.0005 for L-Net₁ and L-Net₂, and 0.0002 for retraining N-Net. End-to-end training is done for 20 epochs with an initial learning rate of 0.0001, halving it every 5 epochs.

We implemented our framework in PyTorch [36] and used the Adam optimizer [25] with default parameters. The full network has a total of 4.66×10^6 parameters which is comparable to LCNet (4.4×10^6). The batch size and the input image number for each object are fixed to 32 during training. The input images are resized to 128×128 at both training and test time.

5 Experimental Results

In this section, we evaluate our method on synthetic and real data. For measuring estimation accuracy, we used mean angular error (MAE) for light directions and surface normals, and scale-invariant relative error [8] for light intensities.

5.1 Evaluation on synthetic data

To quantitatively analyze the effects of object shapes, biased lighting distributions, and spatially-varying BRDFs (SVBRDFs) on the proposed method,

Fig. 5. (a) Lighting distribution of the synthetic test dataset. (b) Normal maps of *Sphere*, *Bunny*, *Dragon* and *Armadillo*.

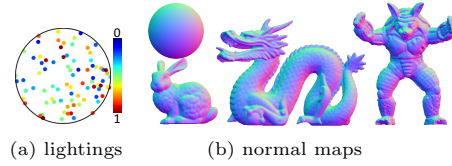


Table 3. Lighting estimation results on the synthetic test dataset. The results are averaged over 100 MERL BRDFs (bold fonts indicates the best).

ID	model	<i>Sphere</i>		<i>Bunny</i>		<i>Dragon</i>		<i>Armadillo</i>	
		direction	intensity	direction	intensity	direction	intensity	direction	intensity
0	LCNet [8]	3.03	0.064	4.88	0.066	6.30	0.072	6.37	0.074
1	L-Net ₁ + N-Net + L-Net ₂ + Finetune	2.21	0.042	2.44	0.046	3.88	0.055	3.52	0.060
2	L-Net ₁ + N-Net + L-Net ₂	2.52	0.052	2.90	0.054	4.20	0.061	3.92	0.060
3	L-Net ₁ + N-Net + L-Net ₂ ^(w/o normal)	2.45	0.050	3.35	0.051	5.82	0.070	5.25	0.059
4	L-Net ₁	3.20	0.053	4.47	0.060	5.80	0.081	5.71	0.079

Table 4. Normal estimation results on the synthetic test dataset. The estimated normals are predicted by PS-FCN [10] given the lightings estimated by LCNet and GCNet.

model	<i>Sphere</i>	<i>Bunny</i>	<i>Dragon</i>	<i>Armadillo</i>
LCNet [8] + PS-FCN [10]	2.98	4.06	5.59	6.73
GCNet + PS-FCN [10]	2.93	3.68	4.85	5.01

Table 5. Results on *Armadillo* under different lighting distributions. Direction, intensity, and normal are abbreviated to “dir.”, “int.”, and “norm.”.

model	near uniform			narrow			upward-biased		
	dir.	int.	norm.	dir.	int.	norm.	dir.	int.	norm.
LCNet [8] + PS-FCN	6.09	0.072	6.49	5.92	0.059	8.44	7.10	0.065	8.80
GCNet + PS-FCN	3.39	0.059	4.90	4.29	0.048	6.82	5.96	0.054	7.53

we rendered a synthetic test dataset using the physically-based raytracer Mitsuba [24]. We rendered 4 different shapes (*Sphere*, *Bunny*, *Dragon* and *Armadillo*) using 100 MERL BRDFs [31], resulting in 400 test objects, each illuminated under 82 randomly sampled light directions. At test time, we randomly generated relative light intensities in the range $[0.2, 2.0]$ to scale the magnitude of the images (see Fig. 5).

Ablation study To validate the design of the proposed network, we performed an ablation study and summarized the results in Table 3. The comparison between experiments with IDs 2-4 verifies that taking both the estimated normals and shading as input is beneficial for lighting estimation. The comparison between experiments with IDs 1 & 2 demonstrates that end-to-end fine-tuning further improves the performance. We can also see that L-Net₁ achieves results comparable to LCNet despite using only half of the network parameters, which indicates that simply increasing the channel number of the convolutional layers cannot guarantee better feature extraction. In the rest of the paper, we denote the results of “L-Net₁ + N-Net + L-Net₂ + Finetune” as “GCNet”.

Table 4 shows that, as expected, the calibrated photometric stereo method PS-FCN [10] can estimate more accurate normals given better estimated lighting.

Results on different lighting distributions To analyze the effect of biased lighting distributions on the proposed method, we evaluated GCNet on the *Ar-*

Table 6. Lighting estimation results on *Bunny* rendered with SVBRDFs. (a) *Bunny* with uniform BRDF. (b) and (c) show the “ramp” and “irregular” material maps and two sample images of *Bunny* with the corresponding SVBRDFs.

model	uniform		ramp		irregular	
	direction	intensity	direction	intensity	direction	intensity
LCNet [8]	4.88	0.066	6.09	0.066	6.00	0.075
GCNet	2.44	0.046	4.16	0.043	3.68	0.050

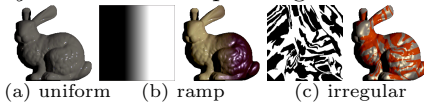
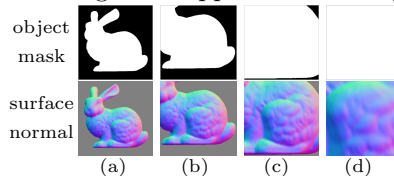


Table 7. Lighting estimation results on surface regions cropped from *Bunny*.

input	<i>alumina-oxide</i>				<i>beige-fabric</i>			
	LCNet		GCNet		LCNet		GCNet	
	dir.	int.	dir.	int.	dir.	int.	dir.	int.
(a)	4.29	0.054	1.35	0.025	4.54	0.051	2.29	0.026
(b)	3.83	0.050	1.71	0.023	4.45	0.044	2.00	0.029
(c)	3.75	0.042	2.46	0.024	4.97	0.044	3.13	0.025
(d)	4.04	0.047	2.84	0.026	4.55	0.051	3.46	0.025



madillo illuminated under three different lighting distributions: a near uniform, a narrow, and an upward-biased distribution. Table 5 shows that both GCNet and LCNet have decreased performance under biased lighting distributions (*e.g.*, the upward-biased distribution), but GCNet consistently outperforms LCNet.

Results on surfaces with SVBRDFs To analyze the effect of SVBRDFs, we used two different material maps to generate a synthetic dataset of surfaces with SVBRDFs following Goldman *et al.* [17]. Specifically, we rendered 100 test objects by randomly sampling two MERL BRDFs and blended the BRDFs for *Bunny* using “ramp” and “irregular” material maps shown in Table 6 (b) and (c). Table 6 shows that although both methods perform worse on surfaces with SVBRDFs compared to uniform BRDFs, our method is still reasonably good even though it was trained on surfaces with uniform BRDFs. This might be explained by that although SVBRDFs may affect the feature extraction of some important clues such as shading, others such as attached shadows and specular highlights are less affected and can still be extracted to estimate reliable lightings.

Effect of the object silhouette Object silhouette can provide useful information for lighting calibration (*e.g.*, normals at the occluding contour are perpendicular to the viewing direction). To investigate the effect of the silhouette, we first rendered the *Bunny* using two different types of BRDFs (*alumina-oxide* and *beige-fabric*) under 100 lightings sampled randomly from the upper hemisphere, and then cropped surface regions with different sizes for testing. Table 7 shows that both LCNet and our method perform robustly for surface regions with or without silhouette, while our method consistently outperforms LCNet. This is because the training data for both methods was generated by randomly cropping image patches from the Blobby and Sculpture datasets [10], which contains surface regions without silhouette.

Runtime The runtimes of LCNet and GCNet for processing an object (96 images in total) from the DiLiGenT benchmark are ~ 0.25 s and ~ 0.5 s including

Table 8. Lighting estimation results on DiLiGenT benchmark.

















																					average	
model	dir.	int.	dir.	int.	dir.	int.	dir.	int.	dir.	int.	dir.	int.	dir.	int.	dir.	int.	dir.	int.	dir.	int.	dir.	int.
PF14 [35]	4.90	0.036	5.31	0.059	2.43	0.017	5.24	0.098	13.52	0.044	9.76	0.053	33.22	0.223	21.77	0.122	16.34	0.074	24.99	0.156	13.75	0.088
LCNet [8]	3.27	0.039	4.08	0.095	5.44	0.058	3.47	0.061	2.87	0.048	4.34	0.048	10.36	0.067	4.50	0.105	4.52	0.073	6.32	0.082	4.92	0.068
GCNet	1.75	0.027	4.58	0.075	1.41	0.039	2.44	0.101	2.81	0.059	2.86	0.032	2.98	0.042	5.47	0.048	3.15	0.031	5.74	0.065	3.32	0.052

Table 9. Lighting estimation results on Light Stage Data Gallery.

													average	
model	dir.	int.	dir.	int.	dir.	int.	dir.	int.	dir.	int.	dir.	int.	dir.	int.
PF14 [35]	25.40	0.576	20.56	0.227	69.50	1.137	46.69	9.805	33.81	1.311	81.60	0.133	46.26	2.198
LCNet [8]	6.57	0.212	16.06	0.170	15.95	0.214	19.84	0.199	11.60	0.286	11.62	0.248	13.61	0.221
GCNet	5.33	0.096	10.49	0.154	13.42	0.168	14.41	0.181	5.31	0.198	6.22	0.183	9.20	0.163

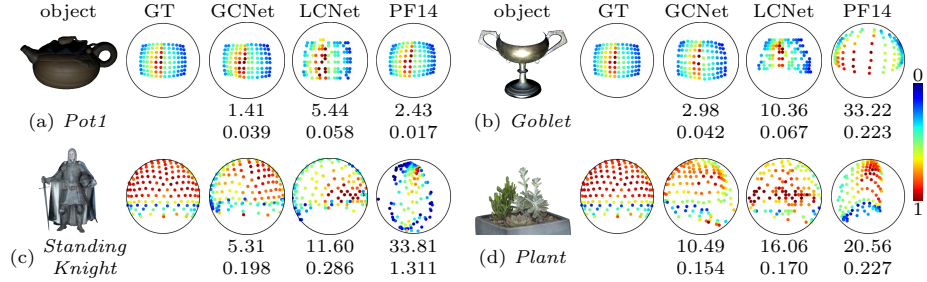


Fig. 6. Visualization of the ground-truth and estimated lighting distribution for the DiLiGenT benchmark and Light Stage Data Gallery.

data loading and network feed-forward time, measured on a single 1080 Ti GPU. Even though LCNet runs slightly faster, both methods are very fast and run within a second.

5.2 Evaluation on real data

To demonstrate the proposed method's capability to handle real-world non-Lambertian objects, we evaluated our method on the challenging *DiLiGenT benchmark* [45] and the *Light Stage Data Gallery* [14].

Results on lighting estimation We first compared our method's lighting estimation results with the state-of-the-art learning-based method LCNet [8] and non-learning method PF14 [35]. Table 8 shows that GCNet achieves the best average results on the DiLiGenT benchmark with an MAE of 3.32 for light directions and a relative error of 0.052 for light intensities. Although our method does not achieve the best results for all objects, it exhibits the most robust performance with a maximum MAE of 5.77 and a maximum relative error of 0.101 compared with LCNet (MAE: 10.36, relative error: 0.105) and

Table 10. Normal estimation results on DiLiGenT benchmark. (* indicates the results of the calibrated method using ground-truth lightings as input.)

model	Ball	Cat	Pot1	Bear	Pot2	Buddha	Goblet	Reading	Cow	Harvest	average
AM07 [4]	7.3	31.5	18.4	16.8	49.2	32.8	46.5	53.7	54.7	61.7	37.3
SM10 [44]	8.9	19.8	16.7	12.0	50.7	15.5	48.8	26.9	22.7	73.9	29.6
WT13 [51]	4.4	36.6	9.4	6.4	14.5	13.2	20.6	59.0	19.8	55.5	23.9
LM13 [29]	22.4	25.0	32.8	15.4	20.6	25.8	29.2	48.2	22.5	34.5	27.6
PF14 [35]	4.8	9.5	9.5	9.1	15.9	14.9	29.9	24.2	19.5	29.2	16.7
LC18 [28]	9.3	12.6	12.4	10.9	15.7	19.0	18.3	22.3	15.0	28.0	16.3
LCNet + ST14	4.1	8.2	8.8	8.4	9.7	11.6	13.5	15.2	13.4	27.7	12.1
GCNet + ST14	2.0	7.7	7.5	5.7	9.3	10.9	10.0	14.8	13.5	26.9	10.8
ST14* [46]	1.7	6.1	6.5	6.1	8.8	10.6	10.1	13.6	13.9	25.4	10.3
LCNet + PS-FCN	3.2	7.6	8.4	11.4	7.0	8.3	11.6	14.6	7.8	17.5	9.7
GCNet + PS-FCN	2.5	7.9	7.2	5.6	7.1	8.6	9.6	14.9	7.8	16.2	8.7
PS-FCN* [10]	2.8	6.2	7.1	7.6	7.3	7.9	8.6	13.3	7.3	15.9	8.4
LCNet + IS18	6.4	15.6	10.6	8.5	12.2	13.9	18.5	23.8	29.3	25.7	16.5
GCNet + IS18	3.1	6.9	7.3	5.7	7.1	8.9	7.0	15.9	8.8	15.6	8.6
IS18* [22]	2.2	4.6	5.4	8.3	6.0	7.9	7.3	12.6	8.0	14.0	7.6

PF14 (MAE: 33.22, relative error: 0.223). Figures 6 (a)-(b) visualize the lighting estimation results for the *Pot1* and the *Goblet*. The non-learning method PF14 works well for near-diffuse surfaces (*e.g.*, *Pot1*), but quickly degenerates on highly specular surfaces (*e.g.*, *Goblet*). Compared with LCNet, our method is more robust to surfaces with different reflectances and shapes.

Table 9 shows lighting estimation results on the Light Stage Data Gallery. Our method significantly outperforms LCNet and PF14, and achieves an average MAE of 9.20 for light directions and a relative error of 0.163 for light intensities, improving the results of LCNet by 32.4% and 26.4% for light directions and light intensities respectively. Figures 6 (c)-(d) visualize lighting estimation results for the Light Stage Data Gallery’s *Standing Knight* and *Plant*.

Results on surface normal estimation We then verified that the proposed GCNet can be seamlessly integrated with existing calibrated methods to handle uncalibrated photometric stereo. Specifically, we integrated the GCNet with a state-of-the-art non-learning calibrated method ST14 [46] and two learning-based methods PS-FCN [10] and IS18 [22]. Table 10 shows that these integrations can outperform existing state-of-the-art uncalibrated methods [4, 44, 51, 29, 35, 28] by a large margin on the DiLiGenT benchmark. We can further see that ST14, PS-FCN, as well as IS18 perform better with ours instead of LCNet’s predicted lightings: 10.8 vs. 12.1 for ST14, 8.7 vs. 9.7 for PS-FCN, and 8.6 vs. 16.5 for IS18. Figure 7 presents a visual comparisons on the *Goblet* from the DiLiGenT benchmark. Please refer to our supplemental material for more results.

5.3 Failure cases

As discussed in Sec. 3, LCNet [8] relies on features like attached shadows, shading, and specular highlights, which is also true for our method. For piecewise planar surfaces with a sparse normal distribution such as the one in Fig. 8 (a), few useful features can be extracted and as a result our method cannot predict reliable lightings for such surfaces. For highly-concave shapes under directional

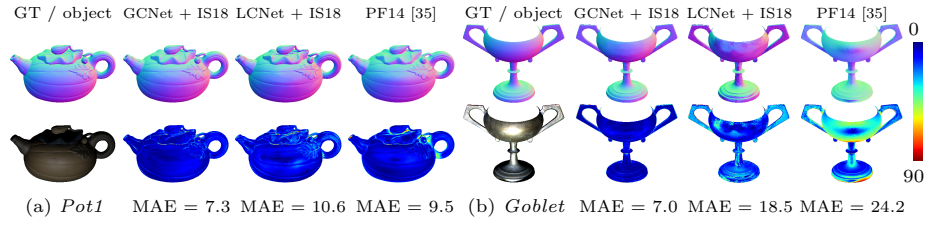
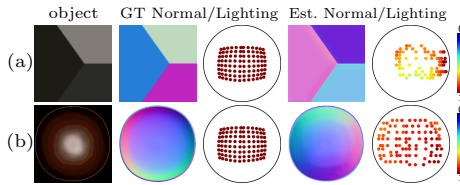


Fig. 7. Visual comparisons of normal estimation for *Pot1* and *Goblet* in the DiLi-GenT benchmark. We compared the normal estimation results of a calibrated method IS18 [22] given lightings estimated by our method and LCNet [8].

Fig. 8. Failure cases. (a) Results on a piecewise planar surface with sparse normal distribution. (b) Results on a highly-concave bowl. The estimated normals are predicted by PS-FCN [8] given our method’s estimated lightings.



lightings, strong cast shadows largely affect the extraction of useful features. Figure 8 (b) shows that GCNet erroneously estimates a highly-concave bowl to be convex. Note that LCNet [8] and PF14 [35] also have similar problems.

6 Conclusions

This paper targeted discovering what is learned in deep uncalibrated photometric stereo to resolve the ambiguity. Specifically, we analyzed and discussed the behavior of the recent deep uncalibrated photometric stereo method LCNet. Based on our findings, we then introduced the guided calibration network (GCNet) that explicitly leverages inter-image information of object shape and intra-image information of shading to estimate more reliable lightings. Experiments on both synthetic and real datasets showed that our method significantly outperforms the state-of-the-art LCNet in lighting estimation, and demonstrated that our method can be integrated with existing calibrated photometric stereo methods to handle uncalibrated setups. Since strong cast shadows affect our method’s feature extraction process and lead to unsatisfactory results, we will explore better methods to handle cast shadows in the future.

Acknowledgments Michael Waechter was supported through a JSPS Postdoctoral Fellowship (JP17F17350). Boxin Shi is supported by the National Natural Science Foundation of China under Grant No. 61872012, National Key R&D Program of China (2019YFF0302902), and Beijing Academy of Artificial Intelligence (BAAI). Kwan-Yee K. Wong is supported by the Research Grant Council of Hong Kong (SAR), China, under the project HKU 17203119. Yasuyuki Matsushita is supported by JSPS KAKENHI Grant Number JP19H01123.

References

1. Ackermann, J., Fuhrmann, S., Goesele, M.: Geometric Point Light Source Calibration. In: *Vision, Modeling & Visualization* (2013)
2. Ackermann, J., Goesele, M.: A survey of photometric stereo techniques. *Foundations and Trends in Computer Graphics and Vision* (2015)
3. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building Rome in a day. *Communications of the ACM* (2011)
4. Alldrin, N.G., Mallick, S.P., Kriegman, D.J.: Resolving the generalized bas-relief ambiguity by entropy minimization. In: *CVPR* (2007)
5. Basri, R., Jacobs, D., Kemelmacher, I.: Photometric stereo with general, unknown lighting. *IJCV* (2007)
6. Belhumeur, P.N., Kriegman, D.J., Yuille, A.L.: The bas-relief ambiguity. *IJCV* (1999)
7. Chandraker, M.K., Kahl, F., Kriegman, D.J.: Reflections on the generalized bas-relief ambiguity. In: *CVPR* (2005)
8. Chen, G., Han, K., Shi, B., Matsushita, Y., Wong, K.Y.K.: Self-calibrating deep photometric stereo networks. In: *CVPR* (2019)
9. Chen, G., Han, K., Shi, B., Matsushita, Y., Wong, K.Y.K.: Deep photometric stereo for non-Lambertian surfaces. *TPAMI* (2020)
10. Chen, G., Han, K., Wong, K.Y.K.: PS-FCN: A flexible learning framework for photometric stereo. In: *ECCV* (2018)
11. Cho, D., Matsushita, Y., Tai, Y.W., Kweon, I.S.: Semi-calibrated photometric stereo. *TPAMI* (2018)
12. Drbohlav, O., Chantler, M.: Can two specular pixels calibrate photometric stereo? In: *ICCV* (2005)
13. Drbohlav, O., Šára, R.: Specularities reduce ambiguity of uncalibrated photometric stereo. In: *ECCV* (2002)
14. Einarsson, P., Chabert, C.F., Jones, A., Ma, W.C., Lamond, B., Hawkins, T., Bolas, M., Sylwan, S., Debevec, P.: Relighting human locomotion with flowed reflectance fields. In: *EGSR* (2006)
15. Epstein, R., Yuille, A.L., Belhumeur, P.N.: Learning object representations from lighting variations. In: *International Workshop on Object Representation in Computer Vision* (1996)
16. Georgiades, A.S.: Incorporating the Torrance and Sparrow model of reflectance in uncalibrated photometric stereo. In: *ICCV* (2003)
17. Goldman, D.B., Curless, B., Hertzmann, A., Seitz, S.M.: Shape and spatially-varying BRDFs from photometric stereo. *TPAMI* (2010)
18. Haefner, B., Ye, Z., Gao, M., Wu, T., Quéau, Y., Cremers, D.: Variational uncalibrated photometric stereo under general lighting. In: *ICCV* (2019)
19. Hayakawa, H.: Photometric stereo under a light source with arbitrary motion. *JOSA A* (1994)
20. Herbort, S., Wöhler, C.: An introduction to image-based 3D surface reconstruction and a survey of photometric stereo methods. *3D Research* (2011)
21. Hertzmann, A., Seitz, S.M.: Example-based photometric stereo: Shape reconstruction with general, varying BRDFs. *TPAMI* (2005)
22. Ikehata, S.: CNN-PS: CNN-based photometric stereo for general non-convex surfaces. In: *ECCV* (2018)
23. Ikehata, S., Aizawa, K.: Photometric stereo using constrained bivariate regression for general isotropic surfaces. In: *CVPR* (2014)

24. Jakob, W.: Mitsuba renderer (2010)
25. Kingma, D., Ba, J.: ADAM: A method for stochastic optimization. In: ICLR (2015)
26. Kriegman, D.J., Belhumeur, P.N.: What shadows reveal about object structure. *JOSA A* (2001)
27. Li, J., Robles-Kelly, A., You, S., Matsushita, Y.: Learning to minify photometric stereo. In: CVPR (2019)
28. Lu, F., Chen, X., Sato, I., Sato, Y.: SymPS: BRDF symmetry guided photometric stereo for shape and light source estimation. *TPAMI* (2018)
29. Lu, F., Matsushita, Y., Sato, I., Okabe, T., Sato, Y.: Uncalibrated photometric stereo for unknown isotropic reflectances. In: CVPR (2013)
30. Lu, F., Matsushita, Y., Sato, I., Okabe, T., Sato, Y.: From intensity profile to surface normal: Photometric stereo for unknown light sources and isotropic reflectances. *TPAMI* (2015)
31. Matusik, W., Pfister, H., Brand, M., McMillan, L.: A data-driven reflectance model. In: SIGGRAPH (2003)
32. Midorikawa, K., Yamasaki, T., Aizawa, K.: Uncalibrated photometric stereo by stepwise optimization using principal components of isotropic BRDFs. In: CVPR (2016)
33. Mo, Z., Shi, B., Lu, F., Yeung, S.K., Matsushita, Y.: Uncalibrated photometric stereo under natural illumination. In: CVPR (2018)
34. Okabe, T., Sato, I., Sato, Y.: Attached shadow coding: Estimating surface normals from shadows under unknown reflectance and lighting conditions. In: ICCV (2009)
35. Papadhimetri, T., Favaro, P.: A closed-form, consistent and robust solution to uncalibrated photometric stereo via local diffuse reflectance maxima. *IJCV* (2014)
36. Paszke, A., Gross, S., Chintala, S., Chanan, G.: PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration (2017)
37. Quéau, Y., Wu, T., Lauze, F., Durou, J.D., Cremers, D.: A non-convex variational approach to photometric stereo under inaccurate lighting. In: CVPR (2017)
38. Santo, H., Samejima, M., Sugano, Y., Shi, B., Matsushita, Y.: Deep photometric stereo network. In: ICCV Workshops (2017)
39. Santo, H., Waechter, M., Samejima, M., Sugano, Y., Matsushita, Y.: Light structure from pin motion: Simple and accurate point light calibration for physics-based modeling. In: ECCV (2018)
40. Sato, I., Okabe, T., Yu, Q., Sato, Y.: Shape reconstruction based on similarity in radiance changes under varying illumination. In: ICCV (2007)
41. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: CVPR (2006)
42. Shen, H.L., Cheng, Y.: Calibrating light sources by using a planar mirror. *Journal of Electronic Imaging* **20**(1), 013002-1–013002-6 (2011)
43. Shi, B., Inose, K., Matsushita, Y., Tan, P., Yeung, S.K., Ikeuchi, K.: Photometric stereo using internet images. In: 3DV (2014)
44. Shi, B., Matsushita, Y., Wei, Y., Xu, C., Tan, P.: Self-calibrating photometric stereo. In: CVPR (2010)
45. 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. *TPAMI* (2019)
46. Shi, B., Tan, P., Matsushita, Y., Ikeuchi, K.: Bi-polynomial modeling of low-frequency reflectances. *TPAMI* (2014)
47. Silver, W.M.: Determining shape and reflectance using multiple images. Ph.D. thesis, Massachusetts Institute of Technology (1980)

48. Tan, P., Mallick, S.P., Quan, L., Kriegman, D.J., Zickler, T.: Isotropy, reciprocity and the generalized bas-relief ambiguity. In: CVPR (2007)
49. Tani, T., Maehara, T.: Neural inverse rendering for general reflectance photometric stereo. In: ICML (2018)
50. Woodham, R.J.: Photometric method for determining surface orientation from multiple images. *Optical Engineering* (1980)
51. Wu, Z., Tan, P.: Calibrating photometric stereo by holistic reflectance symmetry analysis. In: CVPR (2013)
52. Yuille, A.L., Snow, D., Epstein, R., Belhumeur, P.N.: Determining generative models of objects under varying illumination: Shape and albedo from multiple images using SVD and integrability. *IJCV* (1999)
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: ICCV (2019)
54. Zhou, Z., Tan, P.: Ring-light photometric stereo. In: ECCV (2010)