[Supplementary Material] Uncalibrated Neural Inverse Rendering for Photometric Stereo of General Surfaces

Berk Kaya[1] Suryansh Kumar[1] Carlos Oliveira[1] Vittorio Ferrari[2] Luc Van Gool[1,3]
Computer Vision Lab, ETH Zürich[1], Google Research[2], KU Leuven[3]

Abstract

In our supplementary material, we first present a few case studies to analyze our method's effectiveness. Next, we give a detailed description of our coding implementation for training and testing the neural network outlined in the main paper. Formally, this report includes the coding platform details —both hardware and software, with train and test time observed across different datasets. Further, mathematical derivations of our robust initialization and specular-reflectance map formulations are supplied. Finally, we analyze the light estimation performance and discuss the possible future extensions of our method. Besides, our supplementary material includes a short video clip that illustrates the image acquisition setup and visual results.

1. Case Study

This section provides the observation on the case study that we conducted for our proposed method. It is done to analyze the behavior of our method under different possible variations in our experimental setup. Such a study can help us understand the behavior, pros, and cons of our approach.

Case Study 1: *What if we use ground-truth light as input to inverse rendering network instead of relying on light estimation network?*

This case study investigates the reliability of our method. To conduct this experiment, we supplied ground-truth light source directions and intensities as input to the inverse rendering network and robust initialization. The goal is to study the expected deviation in the accuracy of surface normals when ground-truth light sources information is used, compared to the light calibration network. Table (1) compares our method's performance with recent deep calibrated photometric stereo methods on our proposed dataset. The results show that our inverse rendering method achieves the best performance in the calibrated setting, although it does not use a training dataset like other deep-learning-based methods. Additionally, we observed that the CNNPS model proposed by Ikehata [28] which performs per-pixel estimation using observation maps, may not provide accurate surface normals for interreflecting surfaces such as the *Vase* and the *Broken Pot*. Hence, we conclude that extracting information by utilizing the surface geometry is crucial for solving photometric stereo since all surface points affect each other.

Moreover, in Table (1), we show the comparison of our method's performance under calibrated and uncalibrated settings. Our method achieves 12.68° MAE on average, using ground-truth light as input. At the same time, it reaches an average MAE of 14.74° utilizing the information of the light source obtained from the light estimation network. The difference between these two scores is 2.06 degrees, which indicates that the gap between the calibrated and uncalibrated settings is not substantial. Accordingly, we can conclude that our method is robust to the variations in the estimated lighting. Further, we observed that our method performs better with the network estimated light sources information in the categories like *Golf-ball*, *Face*. Hence, based on that observation, we can conclude that the availability of ground-truth calibration data is not a strict requirement for achieving better surface normals estimates in photometric stereo for all kinds of surface geometry.

Case Study 2: *What if we use noisy images?*

Photometric stereo uses a camera acquisition setup, and this implies that noise due to imaging is inevitable. This case study aims to investigate the behavior of our method on different noise levels. To study such a behavior, we synthesized images by adding noise to the images of our proposed dataset. Fig.2 compares the performance of our method under different noise levels. For this case study, we used zero-mean Gaussian noise with different standard deviations (σ=0.05, σ=0.1, σ=0.2). The quantitative results indicate that increasing the noise generally degrades the performance. We observed that the behavior under different noise levels varies among the subjects.

Case Study 3: *Photometric stereo on concentric surfaces with deep concavities and large surface discontinuity.*

To study our photometric stereo method's boundary-condition, we took a complex geometric structure with concentric surfaces, deep-concavities, and large discontinuities for investigation. Accordingly, we synthesized the *Rose*

Type	G.T. Normal	Methods↓ | Dataset →	Vase	Golf-ball	Face	Tablet 1	Tablet 2	Broken Pot	Average Performance
NN-based	✓	Ikehata (2018)[28]	34.00	14.96	16.61	16.64	12.32	18.31	18.81
NN-based	✓	Chen et al.(PS-FCN)(2018)[12]	27.11	15.99	16.17	**10.23**	5.79	**8.68**	14.00
NN-based	✗	Ours (Ground-truth light/ calibrated)	**16.40**	14.23	14.24	10.77	**4.49**	15.92	**12.68**
NN-based	✗	Ours (Estimated light/ uncalibrated)	19.91	**11.04**	**13.43**	12.37	13.12	18.55	14.74
		Diff. in MAE (Ours(Est)-Ours(GT))	+3.51	-3.19	-0.81	+1.60	+8.63	+2.63	+2.06

Table 1: Comparison of recent deep **calibrated** photometric stereo methods Ikehata [28] and Chen *et al*. [12] (PS-FCN) against our method under **uncalibrated** and **calibrated** setting. For testing our method under the calibrated setting, we evaluate the performances assuming that ground-truth light source directions and intensities are available. Note that Chen *et al*. [12] and Ikehata [28] additionally uses ground-truth surface normals for training, in contrast to our method. The last row shows the difference between our method results when used under uncalibrated and calibrated setting respectively. We can see that the average difference in MAE between the two settings of our method is not significant.

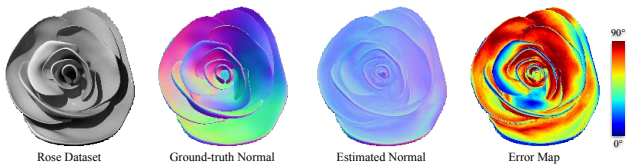


Figure 1: **Failure case:** Qualitative results on the Rose dataset.

dataset using the same dome-settings outlined in the main paper. Fig.1 shows the qualitative results obtained on this dataset. Our method achieves 60.82 degrees of MAE on this particular example. We observed that our approach could not handle this complex geometry because the surface is highly discontinuous with excessive gaps between the leaves. The scene is also affected by occlusions and cast shadows, and therefore, modeling the interreflections for this case seems very difficult.

Though our method applies to a broad range of objects, our interreflection modeling is inspired by Nayar *et al*. [49] formulation, which may not hold for all kinds of surfaces. The interreflection modeling computes depth from the normal map under the continuous surface assumption, which fails in this case study. Furthermore, it models continuous surfaces with discrete facets. Due to such limitations, our method may not be suitable for concentric surfaces with deep concavities and large discontinuities. In such cases, the interreflection effect is very complicated, and our approach may disappoint to model such complex light phenomena.

2. Coding Details

This section provides a detailed description of our source code implementation. We start by introducing the light estimation network's training phase. Then we focus on the testing phase, where the inverse rendering network is optimized to estimate the surface normals, depth, and BRDF values. Finally, we present details on training and testing run-times.

2.1. Training Details

As our inverse rendering network optimizes its learnable parameters at the test time, we apply a training stage only to the light estimation network. For training the network, we

Noise Std (σ)	Vase	Golf-ball	Face	Tablet1	Tablet2	Broken Pot	Average
0.0	19.91	11.01	13.43	12.37	13.12	18.55	14.73
0.05	21.96	11.54	12.94	17.25	11.22	17.22	15.36
0.1	25.01	11.83	15.12	18.80	11.55	19.06	16.90
0.2	24.41	14.25	19.62	21.27	10.07	18.16	17.96

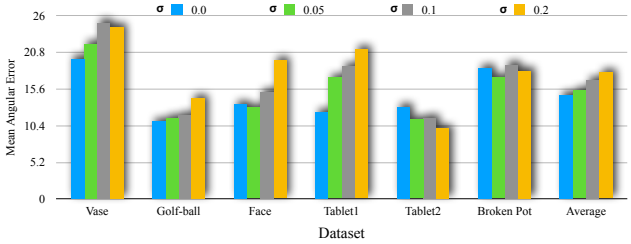


Figure 2: The performance of our method against different noise levels. We used zero-mean Gaussian noise (μ = 0) with different standard deviations (σ). We observed that increasing the noise level generally degrades the performance. Still, the behavior under different noise levels varies among the subjects as the performance depends on the signal-to-noise ratio of the images.

used Blobby and Sculpture datasets that are introduced by Chen *et al*. [12]. This dataset is created by using 3D geometries of Blobby [33], and Sculpture [68] shape datasets and combining them with different material BRDFs taken from MERL dataset [46]. In total, the complete dataset contains 85212 subjects. For each subject, there exist 64 renderings with different light source directions. The intensity of the light sources is kept constant during the whole data generation process. To simulate different intensities during training, image intensity values are randomly generated in the range of $[0.2, 2]$, and these intensity values are used to scale the image data linearly. In each training iteration, the input data is perturbed in the range of $[-0.025, 0.025]$ for augmentation.

The light estimation network is a multiple-input multiple-output (MIMO) system which requires images of the same object captured under different illumination conditions (see Fig.3). The core idea is that all input images have the same surface, and having more images helps the network extract better global features. During training, we use 32 images of the same object for global feature extraction. Note that all of the images are used for feature extraction at test time to achieve the best performance from the network.

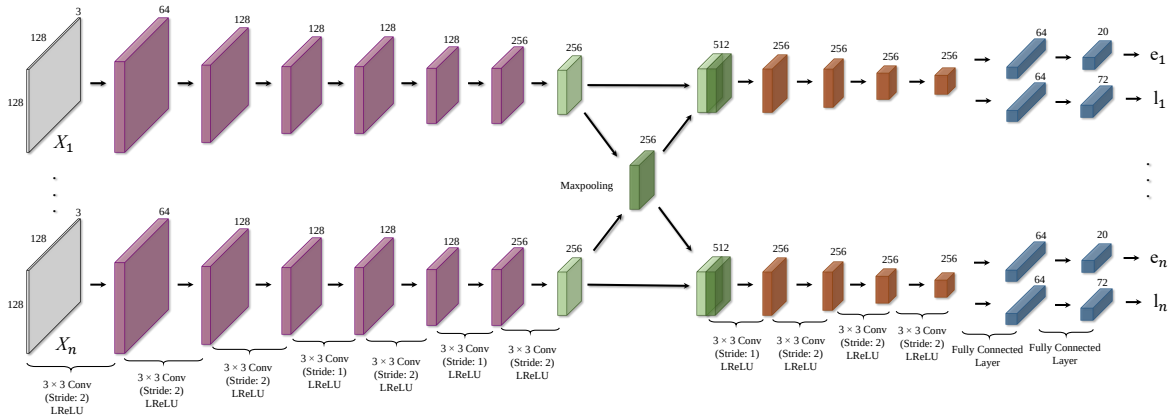


Figure 3: **Architecture of the light estimation network.** The network first extracts features from the input images separately using feature extraction layers (purple). Then, the extracted image-specific features (light-green) are fused with max-pooling operation to obtain a global representation of the entire scene (dark-green). Finally, all image-specific and global features are used in classifier network where convolution (brown) and fully-connected (blue) layers are used to predict light intensity values (e_i's) and direction vectors ($\mathbf{l}_i$'s).

2.2. Testing Details

Given a set of test images $\mathbf{X}$ and object mask $\mathbf{O}$, we first use the light estimation network to have light source directions and intensities. However, the light estimation network operates on 128×128 images because it uses fully connected layers for classification, and these layers process only fixed-length vectors. Consequently, we scale the input images into the resolution of 128×128 before feeding them to the network. We apply this pre-processing step only for the light estimation network and use the original image size for all other operations during testing.

Once we obtain the light source directions and intensities, we apply the robust initialization algorithm to get an initial surface normal matrix $\mathbf{N}_{init}$. It also provides an albedo map that is transformed into $\mathbf{P} \in \mathbb{R}^{m \times m}$ which is required for interreflection modeling. Details about the robust initialization method are explained and derived in §3.1.

After the robust initialization process, we start the optimization of our inverse rendering framework. First, we initialize all the network parameters (Θ_f, Θ_{n1}, Θ_{sp}, Θ_{lg} Θ_{ri}) which correspond to the weights of the convolution operations. In this step, we initialize the weights randomly by sampling from a Gaussian distribution with zero mean and 0.02 variance. We perform 1000 iterations in total using Adam optimizer [34] with an initial learning rate of 8×10^{-4}. The learning rate is reduced by a factor of 10 after 900 iterations for fine-tuning. We observed that setting these hyperparameters may result in convergence problems in our dataset. For this reason, we set the initial learning rate of the estimation branch (ξ_f and ξ_{n1}) to 8×10^{-5} while experimenting on our dataset. We also inject Gaussian noise with zero mean and 0.1 variance to the images before feeding them to f_{sp} for image reconstruction. We observed that this prohibits the network from generating degenerate solutions. At every 100 iterations, we update the depth and the interreflection kernel matrix entries using the normal estimation $\mathbf{N}_o$.

(a) Depth: To compute the depth from normals, we use a gradient-based method with surface orientation constraint [3]. Given the surface normals, we first compute a gradient field $\hat{\mathbf{G}} \in \mathbb{R}^{h \times w \times 2}$ where h and w are the spatial dimensions. The idea is that the gradient field computed from surface normal map and the estimated depth $\mathbf{D} \in \mathbb{R}^{h \times w}$ should be consistent, *i.e.*, $\nabla \mathbf{D} \approx \hat{\mathbf{G}}$. That corresponds to an overdetermined system of linear equations and is solved by minimizing the following objective function *i.e.*, Eq:(1) using the least-squares approach

$$\min_{\mathbf{D}} \|\nabla \mathbf{D} - \hat{\mathbf{G}}\|^2 \tag{1}$$

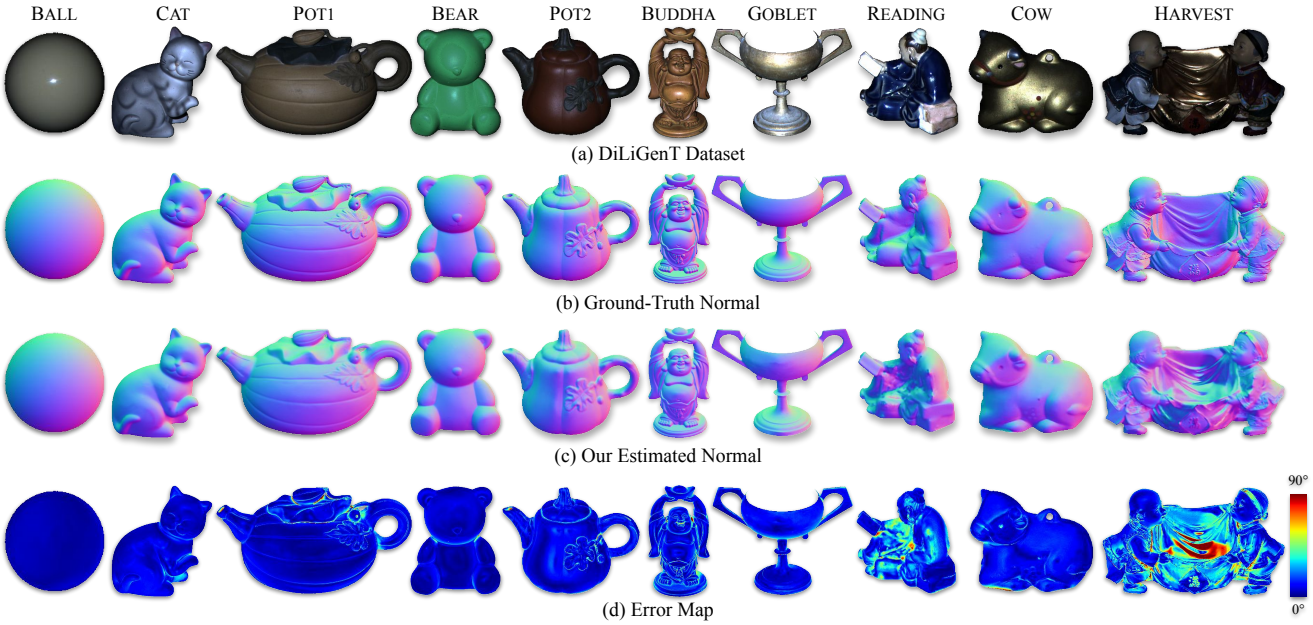
(b) Interreflection Modeling: To consider the effect of interreflection during the image reconstruction process, we define the function ξ_{n2} which uses the estimated normal $\mathbf{N}_o \in \mathbb{R}^{3 \times m}$, albedo matrix $\mathbf{P} \in \mathbb{R}^{m \times m}$ and the interreflection kernel $\mathbf{K} \in \mathbb{R}^{m \times m}$. Given all these components, Nayar *et al.* [49] relates the observed radiance ($\mathbf{X}$) and the radiance due to primary light source ($\mathbf{X_s}$) as follows:

$$\mathbf{X} = (\mathbf{I} - \mathbf{PK})^{-1} \mathbf{X_s} \tag{2}$$

Assuming the surface shows Lambertian reflectance property, we model the radiance in terms of facet matrices as follows:

$$\mathbf{X} = \mathbf{F}_{ny}\mathbf{L}, \;\; \mathbf{X_s} = \mathbf{FL}, \;\; \Rightarrow \mathbf{F_{ny}} = (\mathbf{I} - \mathbf{PK})^{-1}\mathbf{F} \tag{3}$$

Here $\mathbf{F}_{ny} \in \mathbb{R}^{m \times 3}$ and $\mathbf{F} \in \mathbb{R}^{m \times 3}$ are the facet matrices which contain surface normals $\mathbf{N}_{ny}$ and $\mathbf{N}_o$ scaled with local reflectance value. We use Eq:(3) to obtain $\mathbf{F}_{ny}$ and normalize each row to unit vector to obtain $\mathbf{N}_{ny}$.



(a) DiLiGenT Dataset

(b) Ground-Truth Normal

(c) Our Estimated Normal

(d) Error Map

Figure 4: We present visual results of our method on all of the DiLiGenT categories. The bottom row demonstrates the angular error maps obtained form our estimations and ground-truth normals.

The computation of the interreflection kernel $\mathbf{K}$ has the complexity of $\mathcal{O}(n^2)$ where n is the number of facets. Therefore, treating each pixel as a facet limits the application of our method. To approximate the effect of interreflections, we downsample the normal maps with the factor of 4 and calculated the kernel values accordingly. After the normal is updated, we scale it to the original size managing the image details appropriately.

2.3. Timing Details

Our framework is implemented in Python using PyTorch version 1.1.0. Table (2) provides the light estimation network's training time and the inference time of neural inverse rendering network on two datasets separately.

	GPU	Time
Training of Light Estimation Network	Titan X Pascal (12GB)	$\approx$ 22 hours
Inference on DiLiGenT	GeForce GTX TITAN X (12GB)	53.41 $\pm$ 41.57 min per subject
Inference on our Dataset	GeForce GTX TITAN X (12GB)	29.08 $\pm$ 15.99 min per subject

Table 2: Measured training and testing time with respect to the utilized hardware. For our dataset, we have 100 to 260 images per subject and the DiLiGenT dataset has 96 images per subject. Note: Deep photometric stereo method processes a set of images rather than one image for estimating normals.

3. Mathematical Derivations

Here, we supply the mathematical derivation pertaining to the initialization of the surface normals to the inverse rendering network. For completion, we also supplied the well-known deviation of reflection vector §3.2.

3.1. Robust Initialization

Our surface normals initialization procedure aims at recovering the low rank matrix $\mathbf{Z} \in \mathbb{R}^{m\times n}$ from the image matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ such that $\mathbf{X} = \mathbf{Z} + \mathbf{E}$ where $\mathbf{E} \in \mathbb{R}^{m\times n}$ is the matrix of outliers. Here, we assume that the low-rank matrix follows the classical photometric stereo model ($\mathbf{Z} = \mathbf{N}^{\mathrm{T}}\mathbf{L}$) and the outlier matrix $\mathbf{E}$ is sparse in its distribution. Since it is known by definition that $\mathbf{Z}$ spans a rank-3 space, it can be formulated as a standard RPCA problem [71]. However, we know that RPCA formulation performs the nuclear norm minimization of $\mathbf{Z}$ matrix which not only minimizes the rank but also the variance of $\mathbf{Z}$ within the target rank. Now, for the photometric stereo model, it is easy to infer that $\mathbf{N}$ lies in a rank 3 space. As the true rank for $\mathbf{Z}$ is known from its mathematical construction, we do not want to minimize the subspace variance within the target range. Nevertheless, this strict constraint is difficult to meet due to the complex imaging model, and therefore, we encourage to preserve the variance of information within the target range while minimizing the other singular values outside the target rank (K). So, we minimize the partial sum of the singular values which are outside the target rank with the following optimization as follows:

$$\underset{\mathbf{Z},\mathbf{E}}{\text{minimize}} \ \|\mathbf{Z}\|_{r=K} + \lambda\|\mathbf{E}\|_1, \ \text{ subject to: } \mathbf{X} = \mathbf{Z} + \mathbf{E} \tag{4}$$

The Augmented Lagrangian function of Eq:(4) can be written as follows:

$$\mathcal{L}(\mathbf{Z},\mathbf{E},\mathbf{Y}) = \|\mathbf{Z}\|_{r=K} + \lambda\|\mathbf{E}\|_1 + \frac{\mu}{2}\|\mathbf{X}-\mathbf{Z}-\mathbf{E}\|_F^2 + \\ < \mathbf{Y}, \mathbf{X}-\mathbf{Z}-\mathbf{E} > \tag{5}$$

Here, μ is a positive scalar and $\mathbf{Y} \in \mathbb{R}^{m\times n}$ is the estimate of the Lagrange multiplier. As minimizing this function is challenging, we solve it by utilizing the alternating direction method of multipliers (ADMM)[8, 50, 40]. Accordingly, the optimization problem in Eq:(5) can be divided into sub-problems, where $\mathbf{Z}$, $\mathbf{E}$ and $\mathbf{Y}$ are updated alternatively while keeping the other variables fixed.

1. Solution to Z:

$$\mathbf{Z}^* = \underset{\mathbf{Z}}{\text{argmin}}\ \|\mathbf{Z}\|_{r=K} + \frac{\mu_k}{2}\|\mathbf{Z} - (\mathbf{X} - \mathbf{E}_k + \mu_k^{-1}\mathbf{Y}_k)\|_F^2 \tag{6}$$

The solution to Eq:(6) sub-problem at k^{th} iteration is given by $\mathbf{Z}_k = \mathcal{P}_{K,\mu_k^{-1}}[\mathbf{X} - \mathbf{E}_k + \mu_k^{-1}\mathbf{Y}_k]$ where, $\mathcal{P}_{K,\tau}[\mathbf{M}] = \mathbf{U}_\mathbf{M}(\Sigma_{\mathbf{M_1}} + \mathcal{S}_\tau[\Sigma_{\mathbf{M_2}}])\mathbf{V}_\mathbf{M}^T$ is the partial singular value thresholding operator [50] and $\mathcal{S}_\tau[x] = \text{sign}(x)\max(|x|-\tau, 0)$ is the soft-thresholding operator [23]. Here, $\mathbf{U}_\mathbf{M}, \mathbf{V}_\mathbf{M}$ are the singular vector of matrix $\mathbf{M}$ and $\Sigma_{\mathbf{M_1}} = \textbf{diag}(\sigma_1, \sigma_2, ...\sigma_K, 0, 0)$, $\Sigma_{\mathbf{M_2}} = \textbf{diag}(0, 0, .., \sigma_{K+1}, .., \sigma_N)$.

2. Solution to E:

$$\mathbf{E}^* = \underset{\mathbf{E}}{\text{argmin}}\ \lambda\|\mathbf{E}\|_1 + \frac{\mu_k}{2}\|\mathbf{E} - (\mathbf{X} - \mathbf{Z}_{k+1} + \mu_k^{-1}\mathbf{Y}_k)\|_F^2 \tag{7}$$

The solution to Eq:(7) sub-problem at k^{th} iteration is given by $\mathbf{E}_k = \mathcal{S}_{\lambda\mu_k^{-1}}[\mathbf{X} - \mathbf{Z}_{k+1} + \mu_k^{-1}\mathbf{Y}_k]$ where, $\mathcal{S}_\tau[x] = \text{sign}(x)\max(|x|-\tau, 0)$ is a soft-thresholding operator [23]. For proof of convergence and theoretical analysis of partial singular value thresholding operator kindly refer to Oh *et al*. [50] work. We solve for $\mathbf{Z}$, $\mathbf{E}$ using ADMM until convergence for $K = 3$ and use the obtained surface normals for initializing the loss function of inverse rendering network.

3. Solution to Y: The variable $\mathbf{Y}$ is updated as follows over the iteration:

$$\mathbf{Y}_{k+1} = \mathbf{Y}_k + \mu_k(\mathbf{X} - \mathbf{Z}_{k+1} - \mathbf{E}_{k+1}) \tag{8}$$

For more details on the implementation kindly refer to Oh *et al*. [50] method.

3.2. Derivation of Specular-Reflection Equation 11 in the Main Paper

For completion, we derive Equation 11 of the main paper that is used to compute the specular-reflection map $R_i \in \mathbb{R}^{h\times w\times 1}$ for each image. To compute it, we first compute $\mathbf{r}_{\mathbf{x}i}$ for each point $\mathbf{x}$ that is the direction vector with the highest specular component using the following well-known relation; assuming $\mathbf{l}_i$, and $\mathbf{n}_o$ as unit length vectors:

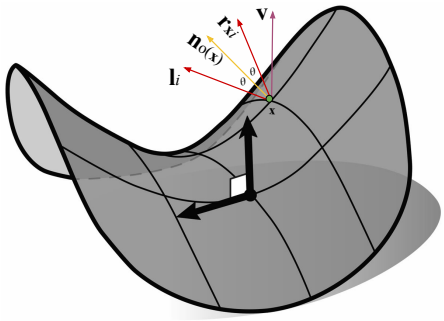


Figure 5: Illustration of surface reflectance. When light ray $\mathbf{l}_i$ hits a surface element, the specular component along the view-direction of the point x due to i^{th} source is given by $\mathbf{r}_{\mathbf{x}i}$. This presentation of 3D geometry is inspired by Keenan work [17].

$$\begin{aligned} &\mathbf{r}_{\mathbf{x}i} + \mathbf{l}_i = 2cos(\theta).\mathbf{n}_o(\mathbf{x});\ \ \mathbf{n}_o(\mathbf{x})^T\mathbf{l}_i = cos(\theta) \\ &\mathbf{r}_{\mathbf{x}i} = 2\big(\mathbf{n}_o(\mathbf{x})^T\mathbf{l}_i\big)\mathbf{n}_o(\mathbf{x}) - \mathbf{l}_i \end{aligned} \tag{9}$$

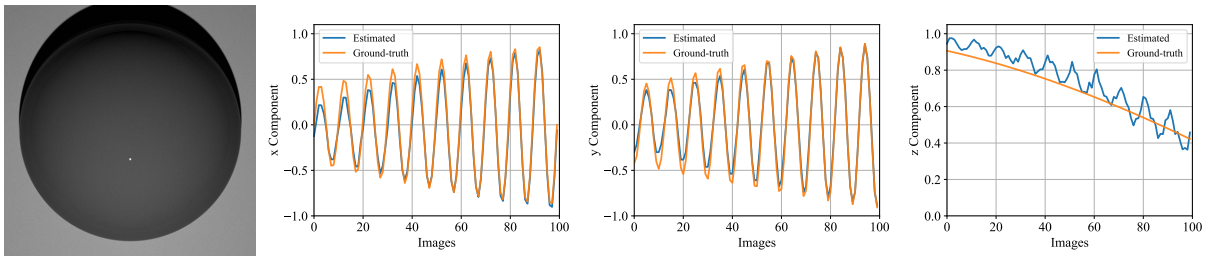
Here, $\mathbf{r}_{\mathbf{x}i}$ is also a unit length vector (see Fig.5). The component of specular reflection in the view-direction $\mathbf{v} = (0,0,1)^T$ of the point $\mathbf{x}$ due to i^{th} light is computed as:

$$\mathbf{r}_{\mathbf{x}i} = \mathbf{v}^T\Big(2\big(\mathbf{n}_o(\mathbf{x})^T\mathbf{l}_i\big)\mathbf{n}_o(\mathbf{x}) - \mathbf{l}_i\Big) \tag{10}$$

The above relation show that the specular highlights are strongest if the normal $\mathbf{n}_o(\mathbf{x})$ is closest to $\mathbf{r}_{\mathbf{x}i}$. Performing this operation for each point gives us the specular-reflection map $\mathbf{R}_i$.

4. Statistical Analysis of Estimated Light Source Directions

We aim to investigate the source directions' behavior predicted by the light estimation network (Fig.3). For that purpose, we use a well-known setup used for light calibration, *i.e*., a calibration sphere. Our renderings from the calibration sphere (see Fig.6(a)) has specular highlights and attached shadows, which provide useful cues for the light estimation network. Figures 6(b)-6(d) illustrate the x, y and z components of the estimated light source direction and ground-truth with respect to the images. We measured the MAE between these vectors as 6.31 degrees. We also observed that the x and y components match well with the ground-truth values. On the other hand, we observed fluctuations on the z component where the values slightly deviate from the ground-truth in a specific pattern. One possible explanation for this observation is that the network has a bias such that its behavior changes in the different regions of the lighting space. Since we generated the data by moving the light source on a circular pattern around z-axis, Fig. 6(d)



(a) Calibration Sphere (b) x Component (c) y Component (d) z Component

Figure 6: Light source directions obtained from the calibration sphere (a) using the light estimation network. We demonstrate the x,y and z components of the light direction vectors (b-d). The mean angular error between the ground-truth and estimated light directions is 6.31 degrees.

also follows a similar pattern with the same frequency with x and y components' curves.

5. More Qualitative Results Comparison on our Dataset

Here, we present qualitative results on all of the categories of our proposed dataset. Figure 7 to Figure 12 compares the output normal maps of our method with other baselines. Note that our implementation of Nayar *et al.* [49] uses Woodham's classical photometric stereo [69] to calculate the pseudo surface and updates the normals with the interreflection modeling for 15 iterations. Even though the Nayar *et al.* [49] interreflection algorithm is not theoretically guaranteed to converge for all surfaces, it gives a stable response on our dataset. We initialized Nayar's algorithm using the same predicted light sources of our method for a fair comparison.

The results show that our method achieves the best results overall, both qualitatively and quantitatively. We observed that other deep learning networks [12, 10] may fail to remove the surface ambiguity in challenging subjects. This is because these networks require supervised training with ground-truth normals, and their performance depends on the content of the training dataset. On the other hand, the results show that Nayar *et al.* [49] performs much better on challenging concave shapes. However, it cannot model specularities and cast shadows. On the other hand, our method can model these non-Lambertian effects with the reflectance mapping, and therefore, it performs better than Nayar *et al.* in all the tested categories.

Lastly, we provide the reflectance map obtained using our method on the proposed dataset. Figure 13 and Figure 14 show the reflectance map obtained using our method on the synthetic and real sequence respectively.

6. Some General Comments

Q1*: Influence of complex texture on the light estimation.* Indeed, surface texture is can be important for light estimation. However, the present benchmark datasets *i.e.*, DiLiGenT is composed of textureless subjects, and therefore, our focus was to perform surface reconstruction on textureless objects.

Ball	Cat	Pot1	Bear	Pot2	Buddha	Goblet	Reading
985	2808	3601	2585	2193	2787	1636	1723
Cow	Harvest	Vase	Golf-ball	Face	Tablet 1	Tablet 2	Broken Pot
1651	3582	1280	468	435	1610	437	1046

Table 3: Number of facets per subject used for our experiments.

Q2*: Nayar interreflection model vs. Monte Carlo:* Monte Carlo method can provide more photo-realistic renderings. However, such an approach is again expensive, requires analytic BRDF models, and a sophisticated sampling strategy for computation, which can make the pipeline better, but more involved. So, we favored Nayar's method and used reflectance maps to handle non-Lambertian effects.

Q3*: Number of parameters for the normal estimation network and interreflection kernel computation:* The inverse rendering network has ≈3.7 million parameters (12.3 MB). The interreflection kernel is generally sparse, and efficient software are available to handle large-sized sparse matrices. Table (3) provides the number of facets used for our experiments to calculate the interreflection kernel.

7. Other Possible Future Extension

Our proposed method enables the application of photometric stereo on a broader range of objects. Yet, we think that there are possible future directions to extend it. Firstly, our method is generally a two-stage framework that utilizes a light estimation network and inverse rendering network in separate phases during inference. As an extension of our work, we aim to combine those stages in an end-to-end framework where light, surface normals, and reflectance values are estimated simultaneously. Secondly, our method uses a physical rendering equation for image reconstruction that is not sufficient for modeling all physical interactions between the object and the light. We believe that an improved rendering equation with additional physical constraints will allow better normal estimates. In addition to that, our method utilizes a specular-reflectance map inspired by the Phong reflectance model. Using other sophisticated variants of specular-reflectance map such as

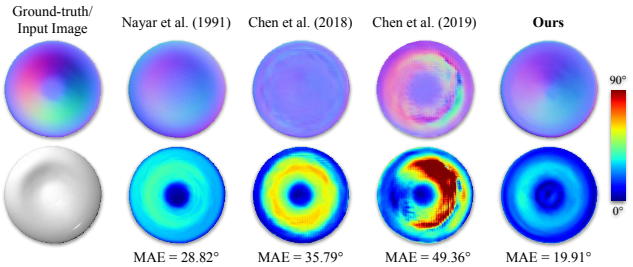


Figure 7: Qualitative comparison on the **Vase** scene. Here, it is obvious that previous deep learning based methods fail to handle the concavity of the subject. In contrast, our method works reasonably well showing the competence of our modeling procedure.

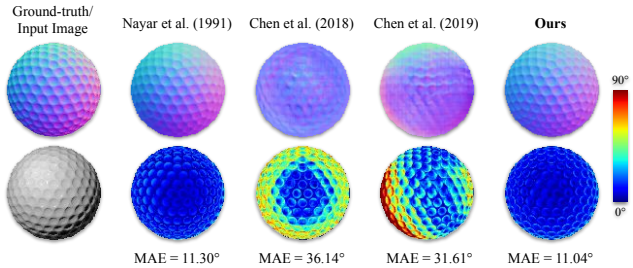


Figure 8: Qualitative comparison on the **Golf-ball** scene. Although deep learning based methods perform well smooth objects, they cannot handle fine structures and indentations.

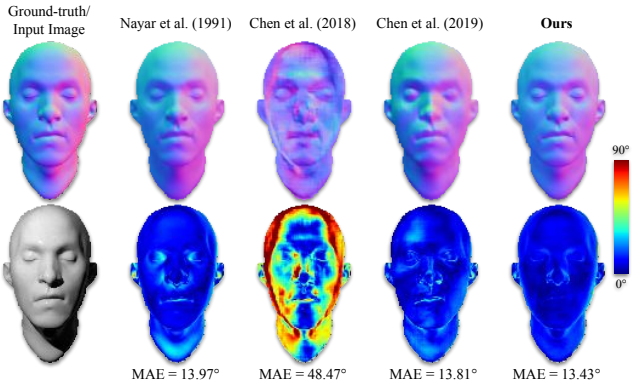


Figure 9: Qualitative comparison on the **Face** scene. Although Nayar *et al*. [49] models interreflections, it cannot handle cast shadows. Therefore, it performs poorly on regions surrounding the eyes and the nose where cast shadows are effective. Here, we also observe that Chen *et al*. [12] cannot estimate accurately for higher slant angles.

the Blinn-Phong reflection model [7] may further advance our approach. Finally, we observed that our method is very convenient for practical usage as it doesn't require ground-truth normals for supervised training. However, it could be possible to improve performance by utilizing training data in a similar framework.

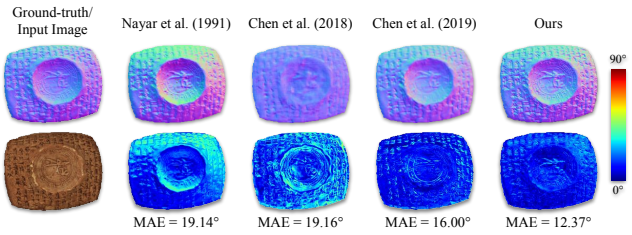


Figure 10: Qualitative comparison on the **Tablet1** scene. This subject has a complicated geometry involving cuneiform and reliefs. Apart from these fine structures, the object can be treated as a composite surface which has a large concavity in the middle part.

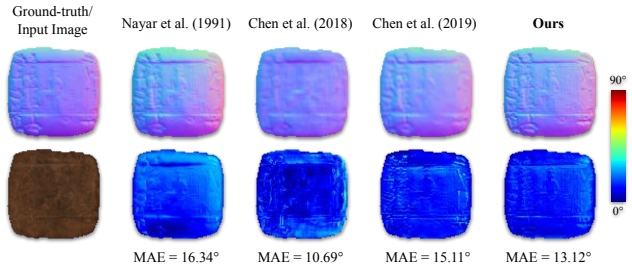


Figure 11: Qualitative comparison on the **Tablet2** scene. Similar to *Tablet1*, this subject also contains reliefs and cuneiform scripts. Since the overall geometry is approximately flat, all methods perform comparable on this category.

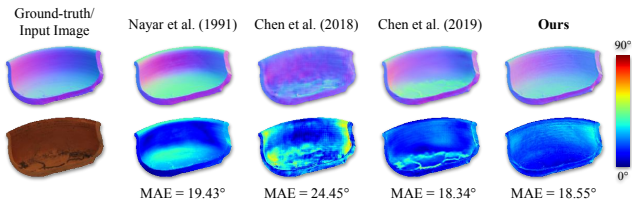


Figure 12: Qualitative comparison on the **Broken Pot** scene.

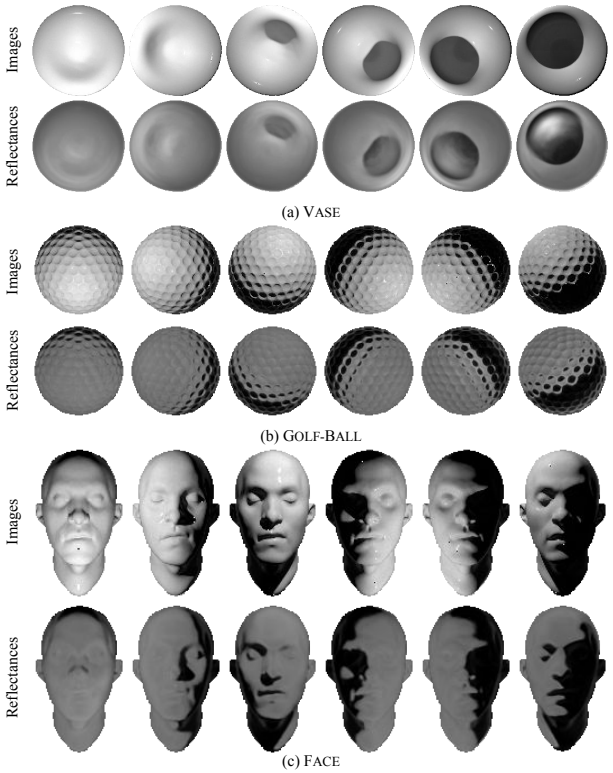


Figure 13: Reflectance maps obtained with our method from **Vase**, **Golf-ball** and **Face** categories.

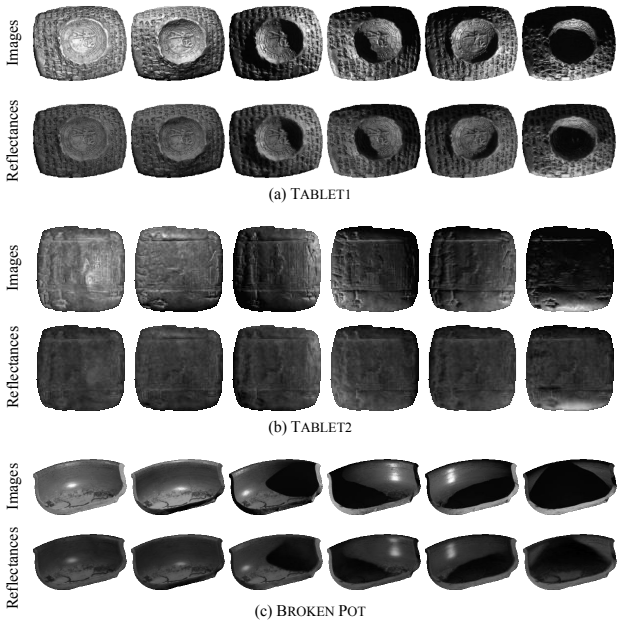


Figure 14: Reflectance maps obtained with our method from **Tablet1**, **Tablet2** and **Broken Pot** categories.

References

[1] Neil Alldrin, Todd Zickler, and David Kriegman. Photometric stereo with non-parametric and spatially-varying reflectance. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[2] Neil G Alldrin, Satya P Mallick, and David J Kriegman. Resolving the generalized bas-relief ambiguity by entropy minimization. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–7. IEEE, 2007.

[3] Doris Antensteiner, Svorad Štolc, and Thomas Pock. A review of depth and normal fusion algorithms. *Sensors*, 18(2):431, 2018.

[4] Louis-Philippe Asselin, Denis Laurendeau, and Jean-Francois Lalonde. Deep SVBRDF estimation on real materials. In *2020 International Conference on 3D Vision (3DV)*. IEEE, nov 2020.

[5] Peter N Belhumeur, David J Kriegman, and Alan L Yuille. The bas-relief ambiguity. *International journal of computer vision*, 35(1):33–44, 1999.

[6] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, David Kriegman, and Ravi Ramamoorthi. Deep 3d capture: Geometry and reflectance from sparse multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5960–5969, 2020.

[7] James F Blinn. Models of light reflection for computer synthesized pictures. In *Proceedings of the 4th annual conference on Computer graphics and interactive techniques*, pages 192–198, 1977.

[8] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[9] Manmohan Krishna Chandraker, Fredrik Kahl, and David J Kriegman. Reflections on the generalized bas-relief ambiguity. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 788–795. IEEE, 2005.

[10] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K Wong. Self-calibrating deep photometric stereo networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8747, 2019.

[11] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee Kenneth Wong. Deep photometric stereo for non-lambertian surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[12] Guanying Chen, Kai Han, and Kwan-Yee K Wong. Ps-fcn: A flexible learning framework for photometric stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18, 2018.

[13] Guanying Chen, Michael Waechter, Boxin Shi, Kwan-Yee K Wong, and Yasuyuki Matsushita. What is learned in deep uncalibrated photometric stereo? In *European Conference on Computer Vision*, 2020.

[14] Lixiong Chen, Yinqiang Zheng, Boxin Shi, Art Subpa-asa, and Imari Sato. A microfacet-based model for photometric stereo with general isotropic reflectance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[15] Zhang Chen, Anpei Chen, Guli Zhang, Chengyuan Wang, Yu Ji, Kiriakos N Kutulakos, and Jingyi Yu. A neural rendering framework for free-viewpoint relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5599–5610, 2020.

[16] Hin-Shun Chung and Jiaya Jia. Efficient photometric stereo on glossy surfaces with wide specular lobes. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[17] Keenan Crane. *Conformal Geometry Processing*. PhD thesis, Caltech, June 2013.

[18] Ondrej Drbohlav and M Chaniler. Can two specular pixels calibrate photometric stereo? In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1850–1857. IEEE, 2005.

[19] Kenji Enomoto, Michael Waechter, Kiriakos N Kutulakos, and Yasuyuki Matsushita. Photometric stereo via discrete hypothesis-and-test search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2311–2319, 2020.

[20] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2009.

[21] Athinodoros S Georghiades. Incorporating the torrance and sparrow model of reflectance in uncalibrated photometric stereo. In *ICCV*, pages 816–823. IEEE, 2003.

[22] Dan B Goldman, Brian Curless, Aaron Hertzmann, and Steven M Seitz. Shape and spatially-varying brdfs from photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(6):1060–1071, 2009.

[23] Elaine T Hale, Wotao Yin, and Yin Zhang. Fixed-point continuation for l_1-minimization: Methodology and convergence. *SIAM Journal on Optimization*, 19(3):1107–1130, 2008.

[24] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[25] Steffen Herbort and Christian Wöhler. An introduction to image-based 3d surface reconstruction and a survey of photometric stereo methods. *3D Research*, 2(3):4, 2011.

[26] Tomoaki Higo, Yasuyuki Matsushita, and Katsushi Ikeuchi. Consensus photometric stereo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1157–1164. IEEE, 2010.

[27] Santo Hiroaki, Michael Waechter, and Yasuyuki Matsushita. Deep near-light photometric stereo for spatially varying reflectances. In *European Conference on Computer Vision*, 2020.

[28] Satoshi Ikehata. Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18, 2018.

[29] Satoshi Ikehata and Kiyoharu Aizawa. Photometric stereo using constrained bivariate regression for general isotropic

surfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2179–2186, 2014.

[30] Satoshi Ikehata, David Wipf, Yasuyuki Matsushita, and Kiyoharu Aizawa. Robust photometric stereo using sparse regression. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 318–325. IEEE, 2012.

[31] Satoshi Ikehata, David Wipf, Yasuyuki Matsushita, and Kiyoharu Aizawa. Photometric stereo using sparse bayesian regression for general diffuse surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(9):1816–1831, 2014.

[32] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[33] Micah K Johnson and Edward H Adelson. Shape estimation in natural illumination. In *CVPR 2011*, pages 2553–2560. IEEE, 2011.

[34] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[35] Suryansh Kumar. Jumping manifolds: Geometry aware dense non-rigid structure from motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2019.

[36] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Monocular dense 3d reconstruction of a complex dynamic scene from two perspective frames. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4649–4657, 2017.

[37] Suryansh Kumar, Yuchao Dai, and Hongdong Li. Superpixel soup: Monocular dense 3d reconstruction of a complex dynamic scene. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[38] Junxuan Li, Antonio Robles-Kelly, Shaodi You, and Yasuyuki Matsushita. Learning to minify photometric stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7568–7576, 2019.

[39] Min Li, Zhenglong Zhou, Zhe Wu, Boxin Shi, Changyu Diao, and Ping Tan. Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials. *IEEE Transactions on Image Processing*, 29:4159–4173, 2020.

[40] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[41] Fotios Logothetis, Ignas Budvytis, Roberto Mecca, and Roberto Cipolla. A CNN based approach for the near-field photometric stereo problem. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

[42] Fotios Logothetis, Ignas Budvytis, Roberto Mecca, and Roberto Cipolla. Px-net: Simple, efficient pixel-wise training of photometric stereo networks. *arXiv preprint arXiv:2008.04933*, 2020.

[43] Fotios Logothetis, Roberto Mecca, and Roberto Cipolla. A differential volumetric approach to multi-view photometric stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1052–1061, 2019.

[44] Feng Lu, Xiaowu Chen, Imari Sato, and Yoichi Sato. 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.

[45] Feng Lu, Yasuyuki Matsushita, Imari Sato, Takahiro Okabe, and Yoichi Sato. Uncalibrated photometric stereo for unknown isotropic reflectances. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1490–1497, 2013.

[46] Wojciech Matusik. *A data-driven reflectance model*. PhD thesis, Massachusetts Institute of Technology, 2003.

[47] Daisuke Miyazaki, Kenji Hara, and Katsushi Ikeuchi. Median photometric stereo as applied to the segonko tumulus and museum objects. *International Journal of Computer Vision*, 86(2-3):229, 2010.

[48] Yasuhiro Mukaigawa, Yasunori Ishii, and Takeshi Shakunaga. Analysis of photometric factors based on photometric linearization. *JOSA A*, 24(10):3326–3334, 2007.

[49] Shree K Nayar, Katsushi Ikeuchi, and Takeo Kanade. Shape from interreflections. *International Journal of Computer Vision*, 6(3):173–195, 1991.

[50] Tae-Hyun Oh, Hyeongwoo Kim, Yu-Wing Tai, Jean-Charles Bazin, and In So Kweon. Partial sum minimization of singular values in rpca for low-level vision. In *Proceedings of the IEEE international conference on computer vision*, pages 145–152, 2013.

[51] Thoma Papadhimitri and Paolo Favaro. A new perspective on uncalibrated photometric stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1474–1481, 2013.

[52] Thoma Papadhimitri and Paolo Favaro. 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.

[53] Jaesik Park, Sudipta N Sinha, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. Multiview photometric stereo using planar mesh parameterization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1161–1168, 2013.

[54] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[55] Yvain Quéau, Tao Wu, François Lauze, Jean-Denis Durou, and Daniel Cremers. A non-convex variational approach to photometric stereo under inaccurate lighting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2017.

[56] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita. Deep photometric stereo network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 501–509, 2017.

[57] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.

[58] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6296–6305, 2018.

[59] Soumyadip Sengupta, Hao Zhou, Walter Forkel, Ronen Basri, Tom Goldstein, and David Jacobs. Solving uncalibrated photometric stereo using fewer images by jointly optimizing low-rank matrix completion and integrability. *Journal of Mathematical Imaging and Vision*, 60(4):563–575, 2018.

[60] Boxin Shi, Yasuyuki Matsushita, Yichen Wei, Chao Xu, and Ping Tan. Self-calibrating photometric stereo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1118–1125. IEEE, 2010.

[61] Boxin Shi, Ping Tan, Yasuyuki Matsushita, and Katsushi Ikeuchi. Bi-polynomial modeling of low-frequency reflectances. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1078–1091, 2013.

[62] Boxin Shi, Zhe Wu, Zhipeng Mo, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3707–3716, 2016.

[63] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[64] Ping Tan, Satya P Mallick, Long Quan, David J Kriegman, and Todd Zickler. Isotropy, reciprocity and the generalized bas-relief ambiguity. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[65] Tatsunori Taniai and Takanori Maehara. Neural inverse rendering for general reflectance photometric stereo. In *International Conference on Machine Learning (ICML)*, pages 4857–4866, 2018.

[66] Xueying Wang, Yudong Guo, Bailin Deng, and Juyong Zhang. Lightweight photometric stereo for facial details recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 740–749, 2020.

[67] Xi Wang, Zhenxiong Jian, and Mingjun Ren. Non-lambertian photometric stereo network based on inverse reflectance model with collocated light. *IEEE Transactions on Image Processing*, 29:6032–6042, 2020.

[68] Olivia Wiles and Andrew Zisserman. SilNet : Single- and multi-view reconstruction by learning from silhouettes. In *Procedings of the British Machine Vision Conference 2017*. British Machine Vision Association, 2017.

[69] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):191139, 1980.

[70] Changchang Wu, Sameer Agarwal, Brian Curless, and Steven M Seitz. Schematic surface reconstruction. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1498–1505. IEEE, 2012.

[71] Lun Wu, Arvind Ganesh, Boxin Shi, Yasuyuki Matsushita, Yongtian Wang, and Yi Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *Asian Conference on Computer Vision*, pages 703–717. Springer, 2010.

[72] Tai-Pang Wu and Chi-Keung Tang. Photometric stereo via expectation maximization. *IEEE transactions on pattern analysis and machine intelligence*, 32(3):546–560, 2009.

[73] Zhe Wu and Ping Tan. Calibrating photometric stereo by holistic reflectance symmetry analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1498–1505, 2013.

[74] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[75] Zhuokun Yao, Kun Li, Ying Fu, Haofeng Hu, and Boxin Shi. Gps-net: Graph-based photometric stereo network. *Advances in Neural Information Processing Systems*, 33, 2020.

[76] Enliang Zheng and Changchang Wu. Structure from motion using structure-less resection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2075–2083, 2015.

[77] Qian Zheng, Yiming Jia, Boxin Shi, Xudong Jiang, Ling-Yu Duan, and Alex C Kot. Spline-net: Sparse photometric stereo through lighting interpolation and normal estimation networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8549–8558, 2019.