NeIF: Representing General Reflectance as Neural Intrinsics Fields for Uncalibrated Photometric Stereo

Zongrui Li¹, Qian Zheng², Feishi Wang³, Boxin Shi³, Xudong Jiang¹

¹School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
²College of Computer Science and Technology, Zhejiang University, Hangzhou, China
³School of Computer Science, Peking University, Beijing, China
{zongrui001, EXDJiang}@ntu.edu.sg, {qianzheng, gpan}@zju.edu.cn,
{wangfeishi, shiboxin}@pku.edu.cn

Abstract

Uncalibrated photometric stereo (UPS) is challenging due to the inherent ambiguity brought by unknown light. Existing solutions alleviate the ambiguity by either explicitly associating reflectance to light conditions or resolving light conditions in a supervised manner. This paper establishes an implicit relation between light clues and light estimation and solves UPS in an unsupervised manner. The key idea is to represent the reflectance as four neural intrinsics fields, i.e., position, light, specular, and shadow, based on which the neural light field is implicitly associated with light clues of specular reflectance and cast shadow. The unsupervised, joint optimization of neural intrinsics fields can be free from training data bias as well as accumulating error, and fully exploits all observed pixel values for UPS. Our method achieves a superior performance advantage over state-of-the-art UPS methods on public and self-collected datasets, under regular and challenging setups. The code will be released soon.

1 Introduction

Photometric stereo (PS) [60] aims at recovering the surface normal from several light-varying images captured at a fixed viewpoint. As compared with other approaches (e.g., multi-view stereo [48], active sensor-based solutions [70]), photometric stereo is excellent at recovering fine-detailed surfaces and has been widely used for Hollywood movies [10], industrial quality inspection [59], and biometrics [65]. Calibrating accurate lighting directions is crucial to the performance of photometric stereo methods [64]. However, lighting calibration is often tedious, dramatically restricting the applicability in the real-world. To this end, researchers develop uncalibrated photometric stereo (UPS) methods that estimate surface normal with unknown lights.

Uncalibrated photometric stereo suffers from General Bas-Relief (GBR) ambiguity [6] for an integrable surface. Early solutions address the ambiguity by explicitly associating reflectance to light, i.e., adopting analytic reflectance models (e.g., Lambertian reflectance [4], parametric specular reflection [21], specular spikes [67], inter-reflection [11]) or imposing priors from reflectance properties [3, 27, 26, 52]. Thus, due to the strong reliance on reflectance assumption, these methods can be less effective for unknown reflectance. Besides, these methods ignore the clue of cast shadow and even fail in shadow regions due to the shadow-free surface assumption. Further, most of them require the

¹In Lambertian objects, there is an inherent 9-parameter ambiguity in the surface normals and light source directions. In the case when the integrability constraint is applied to the surface normals, the ambiguity is reduced to the 3-parameter generalized bas-relief (GBR) ambiguity [21].

²Lambertian reflectance is the property that defines an ideal diffusely reflecting surface [34].
light intensity to be identical for robust estimation. Recently, deep learning-based approaches address the ambiguity by estimating light direction and intensity before recovering surface normal \([12, 14, 33, 47]\). They train a light estimation network using a large-scale amount of rendered data in a supervised manner. However, the training data bias \([37]\) can hardly be eliminated because rendered training data inevitably contains the domain gap from the real ones and scarcely covers all surfaces with different geometry and reflectance in the real-world. Besides, such a two-step solution can bring accumulating errors for surface normal estimation. Further, all these methods assume the light intensity to distribute in a pre-defined range (i.e., \([0, 2]\)). Training data bias, accumulating errors, and pre-defined light intensity range make previous PS methods sensitive to the lab setup of light, i.e., if the setup slightly violates the light assumption (e.g., non-ideal directional light, with little ambient light, out the range of pre-defined intensity), results of previous supervised PS methods degrade significantly.

To this end, we propose NeIF, representing general reflectance as Neural Intrinsics Fields for uncalibrated photometric stereo. Our method differs from previous methods in three aspects: 1) it fully considers clues of specular reflectance and cast shadow from each observed pixel for light estimation so that it is expected to produce accurate estimation for both light conditions and surface normal; 2) it does not make explicit assumptions about the absolute range of reflectance or light so that it works with general surface reflectance and flexible light settings; 3) it infers light and surface normal in an unsupervised and joint manner. Therefore, it can better balance the accuracy between estimated light and surface normal even if the light assumption is not strictly satisfied, which is expected to facilitate the PS method to casual users. Our key idea is to represent the general reflectance as four neural intrinsics fields (i.e., position, light, specular, and shadow, see Fig. 1), implemented by four multi-layer perceptrons (MLPs). These four fields are connected based on the implicit relation (or dependence) of these intrinsics, optimized by reconstructing each pixel value from observed images which fully exploit mutual information across different dimensions, as shown in Fig. 1. The reconstruction error is backpropagated to the neural light field through neural specular and shadow fields so that clues of specular and shadow can be implicitly and fully considered for light estimation. Our contributions are summarized as:

- We represent general reflectance as four intrinsic neural fields to implicitly associate per-pixel reflectance to light, which solves uncalibrated photometric stereo by fully considering clues of specular reflectance and cast shadow for light estimation.
- We propose the NeIF, an uncalibrated photometric stereo method trained in an unsupervised and joint optimization manner, which works with general surface reflectance and flexible light settings, and is free from training data bias and accumulating errors.
- We show that our method achieves superior performance over uncalibrated and unsupervised methods. We also demonstrate its excellent generalization capacity to data from different sources and robust performance with the challenging setup of sparse uncalibrated photometric stereo and casual scenarios.

2 Related Work

This section mainly reviews the latest works in neural reflectance representation and related works on unsupervised PS methods, UPS methods, and intrinsic decomposition. Readers may refer
Neural Reflectance Representation in Multi-view Stereo. Recently, Neural Radiance Fields (NeRFs) \cite{Mildenhall19NeRF} implicitly represent the full 3D geometries and the reflectance of the objects through MLPs. Many subsequent works explore its application in various computer vision problems, such as 3D reconstruction \cite{Chen20Nerf,Mildenhall19NeRF,Chan21Magi,Cao20NeMo,Schonberger19Ours} and relighting \cite{Chen19NeRF,Mildenhall19NeRF,Schonberger19Ours}. While NeRF \cite{Mildenhall19NeRF} applies volume rendering \cite{Lafortune13SDF} to optimize the MLPs through reconstruction loss, they often end up with ‘fake reflectance’ \cite{Schonberger19Ours} (i.e., the emitters inside the objects contribute to specific reflectance effects) and inaccurate geometries due to the lack of an implicit surface representation \cite{Tancik19NeRF}. To solve this, works like \cite{Liu19NeMo,Li20NeMo} combine surface rendering \cite{Lafortune13SDF} with volume rendering \cite{Lafortune13SDF}. Other works like \cite{Li20NeMo,Schonberger19Ours,Cao20NeMo} adopt explicit representation on surface’s reflectance to mimic realistic reflectance, which improve the utility and accuracy of NeRF \cite{Mildenhall19NeRF}. However, all mentioned methods work in multi-view stereo, which varies in the view direction with the static light. Unlike those methods, we reconstruct the surface normal of an object, instead of the full 3D structures, under a setup of photometric stereo that varies in the light directions with a fixed view direction. We also apply surface rendering \cite{Lafortune13SDF} to explicitly model the interaction between light and objects, producing a more realistic approximation of the physical-based intrinsics of an object with different materials.

Unsupervised Photometric Stereo. Classical methods solve the calibrated photometric stereo problem without knowing the ground truth surface normal. Therefore, we classify them as unsupervised methods. The least square-based algorithm \cite{Shen78LS} provides the simplest solution, which assumes the object to be Lambertian. It is generally served as a baseline method due to its stability, but its strong assumption on the reflectance model makes it fail for the non-Lambertian surface. The following works either regard the non-Lambertian reflectance components as the outliers \cite{Li17Unsupervised,Mildenhall19NeRF,Chan21Magi,Schonberger19Ours,Cao20NeMo} or apply analytic reflectance models including Torrance-Sparrow \cite{Torrance86}, the Ward model \cite{Ward80}, specular spike \cite{Cahill94}, etc. to consider the non-Lambertian effects. Those analytic models explicitly consider different reflectance effects based on the microfacet theories \cite{Munz11}. However, the performance of those methods can only deal with limited types of materials. There are also more advanced methods that utilize the general reflectance features such as isotropy \cite{Shen78LS} and monotonicity \cite{Chang76}. Those methods give a reliable estimation for objects with a broad range of materials.

With the progress of deep learning, many learning-based frameworks have been proposed for calibrated photometric stereo. Tanai \textit{et al.} \cite{Tanai19Unsupervised} proposed the first unsupervised learning-based photometric stereo method through a rendering equation. However, their reflectance model does not separately consider shadow, specular highlights, and diffuse components. Recently, Li \textit{et al.} \cite{Li20NeMo} proposed a framework that estimates the spatially varying BRDF (svBRDF) and the surface normal simultaneously with known lights. Considering the ambiguity in UPS, we handle cast shadow and specularly differently by exploiting the clues in those effects. Particularly, the reconstruction loss between the rendered images and the ground truth images can backpropagate from the shadow field to the light field for robust light estimation, while their shadow handling module cannot be differentiated at the light’s input. Moreover, their estimated shadow map is sensitive to the normal map at a specific epoch, which may introduce accumulation errors to the normal map estimation, while ours does not have that concern due to our light-normal joint optimization.

Uncalibrated Photometric Stereo. Previous works hold the Lambertian assumption and exploit extra clues from reflectance, such as half-vector symmetry \cite{Cho16HalfVector,Ullman80}, albedo clustering \cite{Li19ICCV}, specular spikes \cite{Cahill94}, or make additional assumptions of light source distribution, such as ring light \cite{Wang19RingLight}, symmetry light \cite{Li19ICCV}, or uniform-distributed light sources \cite{Wang19DistributedLight,Li19ICCV,Cho15DistributedLight,Shen10DistributedLight}, to resolve three parameters GBR ambiguity \cite{Ross10GPR}. However, all these methods require the light intensity to be identical, which is inapplicable in the real-world datasets such as DILIGENT \cite{Bi20DILIGENT}, APPLE & GOURD \cite{AppleGourd}, and LIGHT STAGE DATA GALLERY \cite{Cho16HalfVector}. Cho \textit{et al.} \cite{Cho16HalfVector} put up a semi-calibrated method to deal with non-uniform light intensity, but they assume the light directions to be known. Quéau \textit{et al.} \cite{Queau19Ours} address the photometric stereo problem under inaccurate lighting calibration, while the accuracy can significantly drop when non-Lambertian components become dominant.

Recently, many deep learning methods have been proposed for uncalibrated photometric stereo. Chen \textit{et al.} \cite{Chen20SDPS} propose a supervised uncalibrated framework, SDPS-Net, which can simultaneously estimate the light conditions (intensity and direction) and the surface normal. They suggest treating light estimation as a classification problem and separating the normal and light prediction to reduce the complexity. Their following work, GC-Net \cite{Chen21GC}, improves the performance of SDPS-Net by
adding shading as an extra channel to the input of the light estimation network. However, as a common problem for all supervised methods, an over-fitting problem may occur due to the training data bias. In contrast, unsupervised methods do not have such a concern. Another benefit is that there is no need to synthesize training sets for unsupervised network training. To utilize the advantage of unsupervised methods, Kaya et al. propose a compromised method that trains the light estimation network in a supervised manner (similar to [42]) but estimates the surface normal in an unsupervised way (similar to [55]). However, they still suffer from training data bias during light estimation. Besides, all these methods make a strict assumption that the light intensity is distributed in a pre-defined range (i.e., \([0.2, 2]\)) and suffer from accumulating error during normal estimation due to their two-step frameworks. In contrast, our method neither makes a strict assumption on reflectance nor needs special light source distribution and jointly solves light conditions and surface normal in an unsupervised manner.

**Intrinsics Decomposition.** Intrinsics decomposition includes topics in a wide range of setting [20]. Works involved are implemented under different types of light sources (e.g., directional light [29], point light [42], or arbitrary lights [68, 49, 75]), targeting on multifarious types of objects (e.g., single arbitrary object [24], faces [77], or human body [32]) described in multiple material models (i.e., Blinn-Phong [38, 71], general sVBRDF [18, 19], etc.). We solved a sub-topic in intrinsics decomposition targeting on the single arbitrary object with the highest complexity in image formation models [20], where the surface normal, the object’s material, and the light are unknown. We also assume the light source to be directional and distant.

## 3 Method

Given a set of observations \( I \triangleq (I_0, I_1, ..., I_f) \) of a static surface, illuminated by \( f \) unknown directional illuminations distributing on the upper-hemisphere, uncalibrated photometric stereo aims at recovering light directions \( L \triangleq (l_0, l_1, ..., l_f) \), light intensities \( E \triangleq (e_0, e_1, ..., e_f) \), and surface normal \( N \triangleq \{ n_i | i \in \mathbb{P} \} \), \( \mathbb{P} \) is the set of all positions on the surface. The solution is achieved by solving the optimization problem,

\[
\arg\min_{L,E,N} \sum_{i=1}^{f} \sum_{j=1}^{f} D(\bar{m}_{ij}, m_{ij}),
\]

(1)

where \( \bar{m}_{ij} \in I_j \) is the observed pixel intensity at position \( i \), \( \mathbb{P} \) is the number of elements in \( \mathbb{P} \), \( m_{ij} \) is the corresponding rendered pixel intensity. \( D(\cdot, \cdot) \) is a metric describing their difference. Under an orthographic camera with the linear radiometric response, \( m_{ij} \) is formulated as (simplified in a per-pixel form),

\[
m_{ij} = e_j \rho_j(n_i^\top l_j, v) \max(n_i^\top l_j, 0) = e_j \rho_{ij} \max(n_i^\top l_j, 0),
\]

(2)

where \( v = [0, 0, 1] \) is the view direction pointing toward the viewer, \( \rho_{ij} \) describes the general reflectance, \( \max(n_i^\top l_j, 0) \) represents the attach shadow.

Unknown light brings two ambiguities when solving Eq. (1), i.e., shape-light ambiguity\(^3\) which is denoted as an invertible matrix \( G \in \mathbb{R}^{3 \times 3} \), and reflectance-light ambiguity, which is denoted as a non-zero scalar \( c_j \in \mathbb{R} \),

\[
m_{ij} = e_j (c_j c_j^{-1}) \rho_{ij} \max(n_i^\top (GG^{-1})l_j, 0).
\]

(3)

### 3.1 Neural Intrinsics Fields

**General reflectance decomposition.** To exploit clues of specular reflectance and cast shadow for light estimation, we decompose the general reflectance as the cast shadow term \( s_{ij} \) multiplying the bidirectional reflectance term,

\[
\rho_{ij} = s_{ij} (k^d_{ij} \rho^d_i + k^s_{ij} \rho^s_j),
\]

(4)

where subscript ‘i’ and ‘j’ indicate position- (or normal-) and light-dependent factors, respectively; the cast shadow term \( s_{ij} \) is either 0 or 1; \( \rho^d_i, \rho^s_j \) represent the diffuse and specular reflectance, and \( k^d_{ij} \) and \( k^s_{ij} \) are coefficients that balance out the effects of specular and diffuse reflectance\(^4\).

\(^3\)Without loss of generality, we put ‘#’ before a set symbol to represent its number of elements in this paper.

\(^4\)Shape-light ambiguity is a more general form of GBR ambiguity in arbitrary non-Lambertian objects.

\(^5\)We think these coefficients of a point will not change under different lights, while they can be different at different positions.
Figure 2: The framework of the proposed NeIF. PositionNet takes the input of positional code and outputs diffuse reflectance $\rho_i^d$, surface normal $n_i$, and coefficients $k_i^d$, $k_i^s$. LightNet takes the observed image $I_j$ as the input and outputs light intensity $e_j$ and direction $l_j$. SpecularNet takes $v_j^\top h_j$ and $n_j^\top h_j$ as the input and outputs specular reflectance $\rho_{ij}^s$. ShadowNet takes inputs of positional code and light direction $l_j$ and output shadow indicator $s_{ij}$. All the intrinsics are used to render the observed pixel value $m_{ij}$ using Eq. (2) and Eq. (4).

Neural fields of specular reflectance and cast shadow. Previous methods identify specular features from specific pixels and associate an explicit relation to light for light estimation. However, this scheme fails for the surface where the specular features are invisible, or the explicit relation is violated. Besides, leaving out the clue of cast shadow can obstruct producing competitive performance.

In contrast, we leverage both clues of specular and shadow for light estimation, which is achieved by building the neural specular field and neural shadow field and associating the fields to light conditions. These neural fields make the utmost of all observed pixels and exploit the mutual information across different normal-light (specular) and position-light (shadow) for light estimation. Two MLPs, namely SpecularNet and ShadowNet, implement these neural fields, respectively. Instead of explicitly estimating light conditions from SpecularNet or ShadowNet, we take the estimated light direction as the input of these networks to achieve their implicit association to light conditions. Since the specular reflectance and cast shadow are normal- and position-dependent, we also feed the estimated surface normal and the positional code to them, respectively. SpecularNet and ShadowNet output $\rho_{ij}^s$ and $s_{ij}$, respectively, as shown in Fig. 2.

Neural light field. There is mutual information across different observed images, i.e., observed images with similar appearances are expected to be illuminated by similar lights. To fully consider and exploit this mutual information, we build the neural light field by concatenating a CNN encoder to an MLP, namely LightNet. The encoder extracts a light code from each observed image, and the LightNet infers the corresponding light conditions (i.e., $e_j$, $l_j$) from the light code, as shown in Fig. 2.

Neural position field. There is mutual information across different positions on a surface, i.e., the consistency of shape and diffuse reflectance in the spatial domain. To fully consider and exploit this mutual information, we establish a neural position field, namely PositionNet, implemented by an MLP. The neural position field outputs position-dependent, light-independent factors, i.e., $n_i$, $\rho_i^d$, $k_i^d$, $k_i^s$. Particularly, factors like $k_i^d$, $k_i^s$ can generate svBRDF by assigning position-dependent values to different surface points even our specular field is not position-dependent (i.e., $k_i^s$ controls the specularity and $k_i^d$ controls the diffuse reflectance of a certain point). The PositionNet takes the positional code as the input, as shown in Fig. 2.

---

6 Unlike [35], our SpecularNet cannot estimate objects with multiple specular lobes. Due to the complexity of the UPS problem, a reflectance model with a smaller number of unknowns helps stabilize the training. Please refer to the Sec. A in the appendix.

7 We use light code instead of positional code because we experimentally find that the light code contains discriminative features of light intensity and direction.

8 Since specular reflectance and cast shadow are both position-dependent and light-dependent, predicting them requires the input of light conditions, which increases the complexity of the neural position field. Therefore, we do not estimate them in the PositionNet, but predict them using SpecularNet and ShadowNet, respectively.

9 Please refer to Sec. C in the appendix for a visual illustration of the estimated svBRDF.
3.2 Optimizing Neural Intrinsics Fields

We adopt the reconstruction loss function with the $\ell_1$ metric to optimize our NeIF,

$$L_{\text{rec}} = \frac{1}{\#P \times f} \sum_{i=1}^{\#P} \sum_{j=1}^{f} \left| \hat{m}_{ij} - e_j s_{ij} (k_i^d \rho_i^d + k_i^s \rho_i^s) \max(\mathbf{n}_i^T \mathbf{l}_j, 0) \right|,$$

(5)

Silhouette constraint. $L_{\text{rec}}$ cannot resolve the shape-light ambiguity in an unsupervised manner due to the inherently severe ill-posedness. Therefore, we introduce the silhouette constraint (similar to those in [25][14]) to stabilize the training of PositionNet. To be specific, we use polynomial fitting with a moving window block to traverse and pre-compute the contour’s normal of the given objects, represented as $\hat{N}^s \triangleq \{ \hat{n}_k^s \in \mathbb{R}^{2 \times 1}, k \in \mathbb{S} \}$. $\mathbb{S}$ is the point set of the contour. We consider $\hat{N}^s$ can guide the prediction of the azimuth of boundary normal (at the same positions) and introduce the silhouette loss function,

$$L_{\text{si}} = \sum_{k=1}^{\#\mathbb{S}} \left| \text{Nor}(C(n_k)) - \hat{n}_k^s \right|,$$

(6)

where $n_k \in \mathbb{R}^{1 \times 3}$ represents the estimated surface normal at the positions of silhouette from PositionNet, $C(\cdot)$ cuts off the 3rd dimension of $n_k$ (i.e., $C(n_k) \in \mathbb{R}^{1 \times 2}$), and Nor(•) is the vector normalization operation.

Warm-up loss functions. To avoid local minimum and achieve faster convergence, we warm up the NeIF in early-stage during training with three additional loss functions. We use the azimuth of lighting direction estimated by YS97 [69] to initialize the weights$^{17}$ of LightNet,

$$L_{\text{az}} = \frac{1}{f} \sum_{j=1}^{f} \left| \text{Nor}(C(l_j)) - \text{Nor}(C(l^w_j)) \right|_2,$$

(7)

where $l^w_j$ are estimated light directions by [69]. We adopt the gradient penalty $^{23}L_{\text{gp}}$ to stabilize the training of SpecularNet,

$$L_{\text{gp}} = \frac{1}{\#P \times f} \sum_{i=1}^{\#P} \sum_{j=1}^{f} \left| \max(-\nabla_{n_i^s}^T h_j, \rho_i^s, 0) \right|_2,$$

(8)

The intuition is from Blinn-Phong model [7], where the specular reflectance is monotonically increasing w.r.t the $n_i^s h_j$, i.e., $\nabla_{n_i^s}^T h_j, \rho_i^s > 0$. We also supervise the training of ShadowNet using pseudo shadow maps $\hat{S}_j, j = 1, 2, ..., f$,

$$L_{\text{shadow}} = \frac{1}{\#P} \sum_{j=1}^{f} \left| S_j - \hat{S}_j \right|_2,$$

(9)

The pseudo shadow maps are obtained by binarizing the observed images, i.e., considering an observed pixel to be cast shadow if its intensity value is smaller than 0.2 times the mean intensity value of this image. After early-stage training, we discard these loss functions for a broad range of reflectance.

NeIF training. We train NeIF with the warm-up loss function in first 10 epochs,

$$L_{\text{warmup}} = L_{\text{rec}} + \lambda_{\text{si}} L_{\text{si}} + \lambda_{\text{az}} L_{\text{az}} + \lambda_{\text{gp}} L_{\text{gp}} + \lambda_{\text{shadow}} L_{\text{shadow}},$$

(10)

where $\lambda_{\text{si}} = 5$, $\lambda_{\text{az}} = 0.1$, $\lambda_{\text{gp}} = 10$, $\lambda_{\text{shadow}} = 10$. We then train NeIF until 500 epochs or converging with the loss function,

$$L_{\text{NeIF}} = L_{\text{rec}} + \lambda_{\text{si}} L_{\text{si}} + \lambda_{\text{shadow}} L_{\text{recShadow}}.$$

(11)

$^{17}$Please refer to Sec. D in the appendix for an explanation about how silhouette constraint solves shape-light ambiguity.

$^{18}$We use YS97 [69] as a weight initialization method on LightNet at the very beginning, with limited influence for subsequent training epochs due to its inaccurate result. To highlight our unsupervised learning manner, a better initialization (e.g., light estimation results of CW20 [14]) is not applied.

$^{15}$We present the details in Sec. E of the appendix about how we calculate the $\hat{S}_j$ by [9].
Table 1: Quantitative comparison in terms of mean angular error for surface normal on DILIGENT benchmark [53]. This table summarizes comparison methods. We report the mean of five random tests. Please refer to Sec. F of the appendix for the standard deviation of our method. 'N.A.' represents not applicable as calibrated PS is with known $\ell$ and $c$. 'Semi' indicates certain methods leverage partial information of light. '✓' (or '✗') represents that certain methods (do not) adopt supervised learning for the estimation of surface normal $n$, light direction $l$, or light intensity $c$. 'Identical' means certain methods require the light intensity of different illuminating to be identical.

| Method | PS | UPS | $\eta$ | $\ell$ | $c$ | BALL | BEAR | BUDDHA | CAT | COW | GOBLEY | HARVEST | POT1 | POT2 | READING |
|--------|----|-----|--------|-------|-----|------|------|--------|-----|-----|--------|---------|------|------|---------|
| LS [60] | PS | ✓   | N.A.   | N.A.  |     | 6.40 | 5.50 | 6.04   | 5.79 | 7.20 | 6.32   | 11.47   | 22.59| 6.09  | 7.76    | 11.03  | 8.83 |
| TM18 PS | PS | ✓   | N.A.   | N.A.  |     | 2.43 | 3.64 | 8.04   | 4.86 | 6.72 | 4.72   | 6.08    | 14.90| 5.99  | 4.97    | 8.75   | 6.50 |
| L12.2 [56] | PS | ✓   | N.A.   | N.A.  |     | 1.47 | 5.79 | 10.36  | 5.44 | 6.32 | 11.47  | 22.59   | 6.09 | 7.76  | 11.03   | 8.83   | 6.50 |
| CHY9 [17] | UPS | ✓   | ✓      | ✓     | ✓   | 2.77 | 6.88 | 7.94   | 8.16 | 11.91| 7.48   | 8.12    | 14.90| 4.97  | 8.75    | 13.85  | 9.51 |
| CM20 [15] | UPS | ✓   | ✓      | ✓     | ✓   | 2.70 | 6.11 | 7.00   | 7.54 | 9.60 | 16.20  | 7.20    | 7.10 | 14.90 | 8.71    |        |      |
| K21 [37] | UPS | ✓   | ✓      | ✓     | ✓   | 2.30 | 5.60 | 8.60   | 9.70 | 8.48 | 10.95  | 8.20    | 7.78 | 7.85  | 10.57   | 11.62  | 9.15 |
| SR22 [27] | UPS | ✓   | ✓      | ✓     | ✓   | 2.68 | 4.78 | 6.04   | 4.86 | 9.78 | 17.97  | 7.97    | 7.76 | 7.00  | 16.10   | 10.01  | 8.88 |
| Ours   | UPS | ✓   | ✓      | ✓     | ✓   | 1.17 | 4.49 | 7.83   | 6.89 | 8.27 | 8.97   | 14.90   | 18.31| 7.00  | 5.85    | 12.02  | 7.83 |

where $\mathcal{L}_{rec\text{Shadow}}$ is another shadow map supervision loss to train the ShadowNet. The loss function is the same to $\mathcal{L}_{\text{shadow}}$. The only difference is the calculation of $\tilde{S}_j$. For $\mathcal{L}_{rec\text{Shadow}}$, we calculate $\tilde{S}_j$ by rendering a depth map with the estimated $I_j$. The depth map is reconstructed from the estimated surface normal map $N$ by method [9]. $\mathcal{L}_{rec\text{Shadow}}$ is used to align outputs of ShadowNet to those of PositionNet.

**Implementation details.** We generate the positional code from the coordinate (in the image plane) by the same method in [39]. With the assumption of isotropic reflectance, we simplify the input of SpecularNet from \{v, $I_j, v \cdot n_i, n_i$\} to \{v, $h_j, n_i$\} [36] for easier training. Similar to the Cook-Torrance reflectance model [17], we assume $k_d + k_s = 1$ to reduce the number of unknowns. The CNN decoder with declining channels processes the down-sampled images with a dimension of $256 \times 256$. The LightNet takes flatten features to predict $I_j, e_j$ in two different branches. For the ShadowNet, we also generate the positional code for $I_j$ and concatenate it to the feature in the 9th layer. The output of ShadowNet is either 0 or 1, which is realized by a step function with a similar implementation in [45].

## 4 Experiments

For network structure, device specification, and hyperparameters, reader can refer to Sec. G of the appendix for more details. For evaluation, we adopt the same metric in [12], the scale-invariant relative error, to measure the accuracy of recovered light intensity as,

$$E_{int} = \frac{1}{f} \sum_{j=1}^{f} \left( \frac{|\eta e_j - \bar{e}_j|}{\bar{e}_j} \right).$$

where, $\eta$ is calculated by solving $\arg \min_{\eta} \sum_{j=1}^{f} (\eta e_j - \bar{e}_j)^2$ by least squares minimization. The metric to measure the accuracy of the predicted light directions and surface normal is the widely used mean angle error (MAE) in degree.

### 4.1 Evaluation on Public Datasets

Since the proposed NeIF is an unsupervised uncalibrated photometric stereo method, we compare its performance with state-of-the-art uncalibrated and unsupervised photometric stereo methods. Three real-world datasets, including the DILIGENT benchmark dataset [53], APPLE & GOURD dataset [2], and LIGHT STAGE DATA GALLERY dataset [10], are used for evaluation.

**Comparison for normal map.** Table 1 lists relevant works for a comprehensive surface normal estimation comparison. As summarized in Table 1, our method is the only method that addresses UPS without the supervision of $N$, $L$, or $E$. However, our method achieves the best performance

---

\footnote{\(h_j\) is the bisection of $I_j$ and $v$, $h_j = \frac{I_j + v}{||I_j + v||}$.

\footnote{More ablation studies could be found in Sec. H of the appendix. Comprehensive estimation results on the public datasets can be found in Sec. I of the appendix.}
Figure 3: Left-right: the reference image, estimated surface normal map from our method, the lighting direction error, light intensity error, the estimated shadow map, diffuse reflectance map, specular reflectance map, diffuse scaling coefficient map, and specular coefficient map. Two objects are GOURD1 from APPLE & GOURD dataset [2], and STANDING KNIGHT from LIGHT STAGE DATA GALLERY dataset [10].

Table 2: Quantitative comparison in terms of mean angular error for light direction and scale-invariant error for intensity on DiLiGENT benchmark [53].

| Model | BALL | BEAR | BUDDHA | CAT | COW | GOBLET | HARVEST | POTT1 | POTT2 | READING | AVG |
|-------|------|------|--------|-----|-----|--------|---------|-------|-------|---------|-----|
|       | dr.  | int. | dr.    | int. | dr. | int.    | dr.     | int.  | int.  | int.    | dr.|
| YS97  | 12.41| 0.194| 14.06  | 0.290| 11.68| 0.300   | 13.75   | 0.318 | 15.79 | 0.251   | 15.24|
| PF14  | 4.90 | 0.036| 5.24   | 0.098| 9.76 | 0.053   | 16.34   | 0.074 | 31.22 | 0.223   | 24.99|
| CH19  | 1.27 | 0.039| 3.47   | 0.061| 4.34 | 0.048   | 0.088   | 0.095 | 0.452 | 0.071   | 0.386|
| CW20  | 1.75 | 0.027| 2.44   | 0.101| 2.86 | 0.032   | 4.58    | 0.075 | 3.15  | 0.031   | 2.98 |
| Ours  | 1.79 | 0.014| 3.54   | 0.031| 2.35 | 0.004   | 2.60    | 0.023 | 5.81  | 0.136   | 8.45 |

Comparison for light directions and intensities. As can be observed in Table 2, our method achieves a superior performance advantage over unsupervised methods (YS97 [69] and PF14 [44]) while maintaining competitive performance as compared with supervised methods (CH19 [12] and CW20 [14]). These supervised methods adopt two-step solutions and pre-defined intensities range, which suffer from accumulating error and data bias. Therefore, even though they achieve similar or better performance in terms of light conditions accuracy on published datasets, they are less effective in estimating surface normal or even failed in the intensities distribution deviating from the pre-defined range, as compared with our method (see Table 1). Readers can refer to Sec. J of the appendix for more details about the light estimation results on APPLE & GOURD dataset [2] and LIGHT STAGE DATA GALLERY dataset [10], and Sec. K of the appendix for the randomly scaled light intensities experiments.

4.2 Evaluation on Challenging Setup

Casual scenarios. To further highlight the advantage of our method regarding unsupervised learning and light-normal joint optimization, we perform an experimental comparison on data collected in casual environments (i.e., the environment with the ambient light and non-ideal directional light source. Please refer to Sec. L of the appendix for more details about the experiment setting and data collecting process). We compare our method with the state-of-the-art UPS method, CW20 [14].

in concurrent UPS methods and maintains a considerable advantage over other competitors [12, 14]. Handling the objects in DiLiGENT dataset [53] with different shapes and reflectance, numbers from our method are either best or competitive (about 1° as compared with the best performing UPS method), which shows its good generalization capacity to general reflectance and various shapes. This is because our NeIF fully considers mutual information by building up intrinsics fields, and the implicit modeling facilitates general reflectance modeling. We also show the reconstructed normal map and other estimated factors for APPLE and STANDING KNIGHT in Fig. 3. Our method produces reliable estimation for most regions, thanks to the full exploitation of mutual information across different dimensions. However, there are inconsistent boundaries on the predicted shadow map, mainly due to the per-pixel estimation manner of ShadowNet. Those artifacts are hard to remove because ShadowNet functions in such a per-pixel way providing important clues for light estimation and stabilizing the training process.\[15\]

\[15\]We replace the ShadowNet with pseudo shadow maps, and the results show the necessity of the ShadowNet. Please refer to Sec I of the appendix for more details.
Figure 4: Visual quality comparison in terms of normal map on BUNNY (row 1) and MOUSE (row 2). Left-right: mean of the observed images, coarse normal map download in https://sketchfab.com/ that has similar shape with the objects for reference, normal map from our method, CW20 [14], light estimation result from our method, CW20 [14]. The predicted lights are projected onto the XY-plane. The color of the subfigures (column 5-6) indicates the value of the light intensity.

The estimated normal map for BUNNY and MOUSE are shown in Fig. 4. According to the result, CW20 [14] are sensitive to the data bias in supervised learning of the light estimation model. The error of estimated light dramatically degrades the performance of normal estimation. In contrast, we have a more reasonable estimation of the selected objects thanks to our unsupervised and light-normal joint optimization manner. The results indicate that our method can better balance the accuracy of the surface normal and light in general cases.

Sparse Uncalibrated Photometric Stereo. To further illustrate the superiority and robustness of our method, we investigate NeIF under sparse lights. We randomly select 10 or 16 images illuminated by different lights from DILGENT dataset [53] and test our method using these images. We report the mean results over 30 trials in Sec. M of the appendix. According to the results, our method achieves competitive performance on normal estimation as compared with state-of-the-art sparse PS methods [74, 13, 44] with known light (i.e., 11.62 for NeIF vs. 9.82 for ZJ19 [74] under 10 lights, and 10.38 for NeIF vs. 9.00 for CH18 [13] under 16 lights). Kindly note that we do not require any ground truth normal for supervision. Compared with the unsupervised UPS method (PF14 [44]), we achieve superior performance. Please refer to Sec. M of the appendix for more details on quantity and quality comparison.

5 Conclusion

This paper proposes NeIF for UPS. By representing the general reflectance as four neural intrinsics fields, it implicitly imposes the light clues of specular reflectance and cast shadow for light estimation, which facilitates solving UPS with general reflectance. The proposed NeIF can fully exploit mutual information from all observed pixel values so that it produces stable estimation for UPS under challenging settings. The unsupervised training manner of NeIF is beneficial to the generalization capacity of data from different sources.

Limitations. Although our method produces promising results for light conditions and surface normal estimation, it has several limitations. First, due to the challenging UPS task, we simplify the image formation model for stable training (i.e., we do not consider the inter-reflection and anisotropic materials; we assume the specular reflectance to be single lobe and bivariate; we only predict the gray-level intrinsics). Second, as shown in Fig. 3 there are artifacts in predicted material intrinsics and shadow map due to the shape-light and reflectance-light ambiguities. Third, the training process of our method is more time-consuming compared to concurrent inverse rendering methods [45, 55] due to the joint optimization manner and the per-pixel estimation, which is inapplicable for real-time applications. At last, the image’s noise will degrade our model’s performance and other inverse rendering methods. Therefore, the observed images should be neither overexposure nor underexposure.

16 Another object, VENUS, is made up of spatially varying BRDF with anisotropic material (i.e., glass). The results are shown in Sec. L of the appendix.
17 Please refer to Sec. N of the appendix for further analysis about the time consumption.
References

[1] Jens Ackermann, Fabian Langguth, Simon Fuhrmann, and Michael Goesele. “Photometric stereo for outdoor webcams”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2012.

[2] Neil Alldrin, Todd Zickler, and David Kriegman. “Photometric stereo with non-parametric and spatially-varying reflectance”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2008.

[3] Neil G Alldrin and David J Kriegman. “Toward reconstructing surfaces with arbitrary isotropic reflectance: A stratified photometric stereo approach”. In: Proc. International Conference on Computer Vision (ICCV). 2007.

[4] Neil G Alldrin, Satya P Mallick, and David J Kriegman. “Resolving the generalized bas-relief ambiguity by entropy minimization”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2007.

[5] Svetlana Barsky and Maria Petrou. “The 4-source photometric stereo technique for three-dimensional surfaces in the presence of highlights and shadows”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2003.

[6] Peter N Belhumeur, David J Kriegman, and Alan L Yuille. “The bas-relief ambiguity”. In: International Journal of Computer Vision (IJCV). 1999.

[7] James F Blinn. “Models of light reflection for computer synthesized pictures”. In: International Conference on Computer Graphics and Interactive Techniques. 1977.

[8] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. “NeRD: Neural Reflectance Decomposition From Image Collections”. In: Proc. International Conference on Computer Vision (ICCV). 2021.

[9] Xu Cao, Boxin Shi, Fumio Okura, and Yasuyuki Matsushita. “Normal Integration via Inverse Plane Fitting With Minimum Point-to-Plane Distance”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2021.

[10] Charles-Félix Chabert, Per Einarsson, Andrew Jones, Bruce Lamond, Wan-Chun Ma, Sebastian Sylwan, Tim Hawkins, and Paul Debevec. “Relighting human locomotion with flowed reflectance fields”. In: EGSR. 2006.

[11] Manmohan Krishna Chandraker, Fredrik Kahl, and David J Kriegman. “Reflections on the generalized bas-relief ambiguity”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2005.

[12] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K Wong. “Self-calibrating deep photometric stereo networks”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2019.

[13] Guanying Chen, Kai Han, and Kwan-Yee K. Wong. “PS-FCN: A Flexible Learning Framework for Photometric Stereo”. In: Proc. European Conference on Computer Vision (ECCV). 2018.

[14] Guanying Chen, Michael Waechter, Boxin Shi, Kwan-Yee K Wong, and Yasuyuki Matsushita. “What is learned in deep uncalibrated photometric stereo?” In: Proc. European Conference on Computer Vision (ECCV). 2020.

[15] Donghyeon Cho, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. “Semi-calibrated photometric stereo”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2018.

[16] Hin-Shun Chung and Jiaya Jia. “Efficient photometric stereo on glossy surfaces with wide specular lobes”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2008.

[17] Robert L Cook and Kenneth E. Torrance. “A reflectance model for computer graphics”. In: ACM TOG. 1982.

[18] Valentin Deschaïntrie, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. “Single-image svbrdf capture with a rendering-aware deep network”. In: ACM TOG. 2018.

[19] Valentin Deschaïntrie, Miika Aittala, Frédéric Durand, George Drettakis, and Adrien Bousseau. “Flexible svbrdf capture with a multi-image deep network”. In: Computer Graphics Forum. 2019.

[20] Elena Garces, Carlos Rodriguez-Pardo, Dan Casas, and Jorge Lopez-Moreno. “A Survey on Intrinsic Images: Delving Deep into Lambert and Beyond”. In: International Journal of Computer Vision. 2022.
[21] Athinodoros S Georghiades. “Incorporating the torrance and sparrow model of reflectance in uncalibrated photometric stereo”. In: Proc. International Conference on Computer Vision (ICCV). 2003.
[22] Dan B Goldman, Brian Curless, Aaron Hertzmann, and Steven M Seitz. “Shape and spatially-varying brdfs from photometric stereo”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2009.
[23] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. “Improved training of wasserstein gans”. In: Proc. Conference on Neural Information Processing Systems (NeurIPS). 2017.
[24] Xian-Feng Han, Hamid Laga, and Mohammed Bennamoun. “Image-Based 3D Object Reconstruction: State-of-the-Art and Trends in the Deep Learning Era”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2021.
[25] Shuhei Hashimoto, Daisuke Miyazaki, and Shinsaku Hiura. “Uncalibrated photometric stereo constrained by intrinsic reflectance image and shape from silhouette”. In: International Conference on Machine Vision Applications (MVA). 2019.
[26] Tomoaki Higo, Yasuyuki Matsushita, and Katsushi Ikeuchi. “Consensus photometric stereo”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2010.
[27] Michael Holroyd, Jason Lawrence, Greg Humphreys, and Todd Zickler. “A photometric approach for estimating normals and tangents”. In: ACM TOG. 2008.
[28] Satoshi Ikehata, David Wipf, Yasuyuki Matsushita, and Kiyoharu Aizawa. “Robust photometric stereo using sparse regression”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2012.
[29] Michael Janner, Jiajun Wu, Tejas D Kulkarni, Ilker Yildirim, and Josh Tenenbaum. “Self-supervised intrinsic image decomposition”. In: Proc. Conference on Neural Information Processing Systems (NeurIPS). 2017.
[30] James T Kajiya. “The rendering equation”. In: Proceedings of the 13th annual conference on Computer graphics and interactive techniques. 1986, pp. 143–150.
[31] James T Kajiya and Brian P Von Herzen. “Ray tracing volume densities”. In: ACM SIGGRAPH computer graphics. 1984.
[32] Yoshihiro Kanamori and Yuki Endo. “Relighting humans: occlusion-aware inverse rendering for full-body human images”. In: ACM TOG. 2019.
[33] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. “Uncalibrated neural inverse rendering for photometric stereo of general surfaces”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2021.
[34] Sanjeev J. Koppal. “Lambertian Reflectance”. In: Computer Vision: A Reference Guide. 2020.
[35] Junxuan Li and Hongdong Li. “Neural Reflectance for Shape Recovery with Shadow Handling”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2022.
[36] Feng Lu, Xiaowu Chen, Imari Sato, and Yoichi Sato. “SymPS: BRDF symmetry guided photometric stereo for shape and light source estimation”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2017.
[37] Brian Mac Namee, Padraig Cunningham, Stephen Byrne, and Owen I Corrigan. “The problem of bias in training data in regression problems in medical decision support”. In: Artificial Intelligence in Medicine. 2002.
[38] Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. “Lime: Live intrinsic material estimation”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2018.
[39] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. “Nerf: Representing scenes as neural radiance fields for view synthesis”. In: Proc. European Conference on Computer Vision (ECCV). 2020.
[40] Kazuma Minami, Hiroaki Santo, Fumio Okura, and Yasuyuki Matsushita. “Symmetric-Light Photometric Stereo”. In: Proc. Winter Conference on Application of Computer Vision (WACV). 2022.
[41] Daisuke Miyazaki, Kenji Hara, and Katsushi Ikeuchi. “Median photometric stereo as applied to the segonko tumulus and museum objects”. In: International Journal of Computer Vision (IJCV). 2010.
[42] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. “Learning physics-guided face relighting under directional light”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2020.

[43] Michael Oechsle, Songyou Peng, and Andreas Geiger. “Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2021.

[44] Thoma Papadhimitri and Paolo Favaro. “A closed-form, consistent and robust solution to uncalibrated photometric stereo via local diffuse reflectance maxima”. In: International Journal of Computer Vision (IJCV). 2014.

[45] Haotong Qin, Ruihao Gong, Xianglong Liu, Xiao Bai, Jingkuan Song, and Nicu Sebe. “Binary neural networks: A survey”. In: Pattern Recognition. 2020.

[46] Yvain Quéau, Tao Wu, François Lauze, Jean-Denis Durou, and Daniel Cremers. “A non-convex variational approach to photometric stereo under inaccurate lighting”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2017.

[47] Francesco Sarno, Suryansh Kumar, Berk Kaya, Zhiwu Huang, Vittorio Ferrari, and Luc Van Gool. “Neural Architecture Search for Efficient Uncalibrated Deep Photometric Stereo”. In: Proc. Winter Conference on Application of Computer Vision (WACV). 2022.

[48] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. “A comparison and evaluation of multi-view stereo reconstruction algorithms”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2006.

[49] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W Jacobs, and Jan Kautz. “Neural inverse rendering of an indoor scene from a single image”. In: Proc. International Conference on Computer Vision (ICCV). 2019.

[50] Boxin Shi, Yasuyuki Matsushita, Yichen Wei, Chao Xu, and Ping Tan. “Self-calibrating photometric stereo”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2010.

[51] Boxin Shi, Ping Tan, Yasuyuki Matsushita, and Katsushi Ikeuchi. “Bi-polynomial modeling of low-frequency reflectances”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2013.

[52] Boxin Shi, Ping Tan, Yasuyuki Matsushita, and Katsushi Ikeuchi. “Elevation angle from reflectance monotonicity: Photometric stereo for general isotropic reflectances”. In: Proc. European Conference on Computer Vision (ECCV). 2012.

[53] 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: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2016.

[54] Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. “NeRV: Neural Reflectance and Visibility Fields for Relighting and View Synthesis”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2021.

[55] Tatsunori Tanai and Takanori Maehara. “Neural inverse rendering for general reflectance photometric stereo”. In: Proc. International Conference on Machine Learning (ICML). 2018.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. “Attention is all you need”. In: Proc. Conference on Neural Information Processing Systems (NeurIPS). 2017.

[57] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. “Ref-nerf: Structured view-dependent appearance for neural radiance fields”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2021.

[58] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. “Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction”. In: Proc. Conference on Neural Information Processing Systems (NeurIPS). 2021.

[59] Eva Weigl, Sebastian Zambal, Matthias Stöger, and Christian Eitzinger. “Photometric stereo sensor for robot-assisted industrial quality inspection of coated composite material surfaces”. In: International Conference on Quality Control by Artificial Vision (QCAV). 2015.

[60] Robert J Woodham. “Photometric method for determining surface orientation from multiple images”. In: Optical Engineering. 1980.

[61] Lun Wu, Arvind Ganesh, Boxin Shi, Yasuyuki Matsushita, Yongtian Wang, and Yi Ma. “Robust photometric stereo via low-rank matrix completion and recovery”. In: Proc. Asian Conference on Computer Vision (ACCV). 2010.
[62] Tai-Pang Wu, Kam-Lun Tang, Chi-Keung Tang, and Tien-Tsin Wong. “Dense photometric stereo: A markov random field approach”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2006.

[63] Zhe Wu and Ping Tan. “Calibrating photometric stereo by holistic reflectance symmetry analysis”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2013.

[64] Limin Xie, Zhan Song, Guohua Jiao, Xinhan Huang, and Kui Jia. “A practical means for calibrating an LED-based photometric stereo system”. In: Optics and Lasers in Engineering. 2015.

[65] Wuyuan Xie, Zhan Song, and Ronald C Chung. “Real-time three-dimensional fingerprint acquisition via a new photometric stereo means”. In: Optical Engineering. 2013.

[66] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. “Neural fields in visual computing and beyond”. In: Computer Graphics Forum. 2022.

[67] Sai-Kit Yeung, Tai-Pang Wu, Chi-Keung Tang, Tony F Chan, and Stanley J Osher. “Normal estimation of a transparent object using a video”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2014.

[68] Ye Yu and William AP Smith. “Inverserendernet: Learning single image inverse rendering”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2019.

[69] Alan Yuille and Daniel Snow. “Shape and albedo from multiple images using integrability”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 1997.

[70] Guangjun Zhang and Zhenzhong Wei. “A novel calibration approach to structured light 3D vision inspection”. In: Optics & Laser Technology. 2002.

[71] Jason Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. “NeRS: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild”. In: Proc. Conference on Neural Information Processing Systems (NeurIPS). 2021.

[72] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. “PhySG: Inverse Rendering With Spherical Gaussians for Physics-Based Material Editing and Relighting”. In: Proc. Computer Vision and Pattern Recognition (CVPR). 2021.

[73] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. “Nerfactor: Neural factorization of shape and reflectance under an unknown illumination”. In: ACM TOG. 2021.

[74] 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: Proc. International Conference on Computer Vision (ICCV). 2019.

[75] Hao Zhou, Xiang Yu, and David W Jacobs. “Glosh: Global-local spherical harmonics for intrinsic image decomposition”. In: Proc. International Conference on Computer Vision (ICCV). 2019.

[76] Zhenglong Zhou and Ping Tan. “Ring-light photometric stereo”. In: Proc. European Conference on Computer Vision (ECCV). 2010.

[77] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. “State of the art on monocular 3D face reconstruction, tracking, and applications”. In: Computer Graphics Forum. 2018.
6 Appendix

In this appendix,

1. Sec. A we switch our rendering model to be identical to \[35\], where we use 9 specular basis instead of 1. (footnote 6 in the main paper).
2. Sec. B we conduct correlation analysis to validate the effectiveness of the CNN extractor. (footnote 7 in the main paper).
3. Sec. C we visualize the BRDFs of different points to demonstrate that our method could generate spatially varying BRDF. (footnote 9 in the main paper).
4. Sec. D we provide an intuitive explanation of how silhouette constraints work in our method. (footnote 10 in the main paper).
5. Sec. E we present details of calculating $\hat{S}_j$. (footnote 12 in the main paper).
6. Sec. F we provide the standard deviation for 10 objects on the DILIGENT benchmark \[53\] under 5 random tests. (table 1’s caption in the main paper).
7. Sec. G we provide details of the network structure, device specification, and training procedure. (Sec. 4 in the main paper).
8. Sec. H we conduct ablation studies to validate effectiveness of our loss functions and network structures. (footnote 14 in the main paper).
9. Sec. I we compare our method with those that remove the ShadowNet and use the shadow maps obtained from simple binarization of observed images (as in \[35\], footnote 15 in the main paper).
10. Sec. J we show complete qualitative comparison regarding 10 objects from DILIGENT dataset \[53\], 3 objects from LIGHT STAGE DATA GALLERY \[10\], and 3 objects from APPLE & GOURD \[2\] in Sec. J. (footnote 14 and Sec. 4.1 in the main paper).
11. Sec. K we randomly scale the intensity of input images and provide an experimental comparison. (Sec. 4.1 in the main paper).
12. Sec. L we collect data in casual scenarios to validate the effectiveness of our method on unsupervised learning and light-normal joint optimization. (Sec. 4.2 in the main paper).
13. Sec. M we conduct additional experiments to validate the effectiveness of our method for sparse uncalibrated photometric stereo. (Sec. 4.2 in the main paper).
14. Sec. N we list the testing time of our method for objects in DILIGENT benchmark \[53\]. We also explain why our model takes a long time to train. (footnote 17 in the main paper).

A Multiple Specular Basis

We extend the number of the basis used in our SpecularNet from 1 to 9 and switch our rendering model to be identical to \[35\]. Specifically,

$$m_{ij} = e_j (\rho^d_i + c_i^T D(h_{ij}, n_i)) \max (n_i^T l_j, 0).$$

Where, $m_{ij}$ is the pixel value, $e_j$ is the light intensity of light $j$, $\rho^d_i$ is the diffuse reflectance at point $i$, $c_i \triangleq (c_0, \ldots, c_k)$ is the specular weights generated by the PositionNet, $k$ is the number of basis, $h_{ij}$ is the half-vector between the surface normal at point $i$ and the light direction $l_j$, $n_i$ is the surface normal of point $i$). The results are shown in Table 3. Although there are improvements in objects like BEAR, POT1, and READING, the overall performance drops significantly. This further illustrates that our rendering model is a necessary compromise for the complexity of the UPS problem.

|          | BALL | BEAR | BUDDHA | CAT  | COW  | GOBLET | HARVEST | POT1 | POT2 | READING | AVG  |
|----------|------|------|--------|------|------|--------|---------|------|------|---------|------|
| Ours     |      |      |        |      |      |        |         |      |      |         |      |
| norm.    | 1.17 | 4.49 | 8.73   | 4.89 | 6.27 | 9.53   | 18.31   | 7.08 | 5.85 | 12.02   | 7.83 |
| dirs.    | 1.79 | 3.54 | 2.33   | 2.60 | 5.81 | 8.45   | 7.40    | 3.73 | 2.10 | 7.91    | 4.57 |
| int.     | 0.01 | 0.01 | 0.03   | 0.02 | 0.20 | 0.04   | 0.03    | 0.07 | 0.04 | 0.05    | 0.05 |
| Ours (9 basis) |      |      |        |      |      |        |         |      |      |         |      |
| norm.    | 1.69 | 4.03 | 12.84  | 5.90 | 8.89 | 14.45  | 20.02   | 6.82 | 9.64 | 11.34   | 9.56 |
| dirs.    | 2.50 | 3.01 | 7.34   | 4.94 | 8.69 | 21.53  | 12.60   | 3.88 | 6.97 | 6.41    | 7.79 |
| int.     | 0.011 | 0.019 | 0.025 | 0.036 | 0.074 | 0.072 | 0.059 | 0.026 | 0.030 | 0.051 | 0.040 |

Table 3: Quantitative results in terms of mean angular error for surface normal, light direction, and scale-invariant error for intensity on DILIGENT benchmark \[53\]. ‘Ours (9 basis)’ indicates using 9 specular basis in SpecularNet.
B Correlation Analysis

While the positional code of the light index seems to be a good candidate as the input of LightNet, it is less effective to estimate light conditions than the observed image as it provides observable clues of intensity. Fig. [5] reveal that the observed images provides important clues. Therefore, we take the observed image as the input of LightNet.

![Correlation Analysis Diagram](image)

Figure 5: Correlation analysis between the pixel value and the light intensity on each object from DiLiGenT dataset [53]. The X-axis refers to the pixel value of the 90th percentile maximum of each 96 image. Y-axis indicates the corresponding light intensity of the corresponding image. The values are scaled into [0, 1] by min-max normalization. The mean correlation of the 10 objects is 0.98.

**Qualitative validation.** We extract feature maps from the last layer of our CNN encoder. As shown in Fig. [6] those feature maps contain observable specular features, indicating both the direction and the intensity of the light sources. [14]

![Feature Visualization Diagram](image)

Figure 6: Feature visualization of CNN encoder on BALL and COW. Left-right: the input images of the CNN encoder, four feature maps from the last layer of the CNN encoder. Similar to [14], the feature maps are normalized into [0, 1] for better illustration.

**Quantitative validation.** We conduct correlation analysis to illustrate the correlation between the learned feature map and the lighting information. First, we use Euclidean Distance (i.e., $L_2$) to measure the distance between feature maps under two different lights, and we order these distances to get
a $\frac{96 \times 95}{2}$-dimension vector $V_i^f$, $i = 1, 2, 3, 4$. Similarly, we get two $\frac{96 \times 95}{2}$-dimension vectors by measuring the distance between every two lights, i.e., $V_i^{ld}$ for light directions with the metric of cosine distance, and $V_i^l$ for light intensities with the metric of $\ell_1$ difference. The cosine similarity between these vectors is shown in Table 4. We find that there are at least one feature map that is highly correlated to the ground truth light direction or light intensity, especially for objects like BALL and READING that have sharp specular lope.

Table 4: The cosine similarity between the distance matrix of the four extracted feature maps ($V_i^f, i = 1, 2, 3, 4$) and the light directions/intensities. "max" indicates the maximum similarity among the feature maps.

|       | BALL  | BEAR  | BUDDHA | CAT   | COW   | GOBLET | HARVEST | POT1  | POT2  | READING |
|-------|-------|-------|--------|-------|-------|--------|---------|-------|-------|----------|
| dir. int. | dir. int. | dir. int. | dir. int. | dir. int. | dir. int. | dir. int. | dir. int. | dir. int. | dir. int. | dir. int. |
| $V_1$ | 0.84 0.93 0.83 0.69 0.96 0.81 0.93 0.81 0.94 0.76 0.87 0.80 0.82 0.89 0.92 0.67 0.90 0.77 0.89 0.91 |
| $V_2$ | 0.96 0.81 0.91 0.82 0.82 0.95 0.71 0.94 0.75 0.94 0.76 0.87 0.80 0.94 0.70 0.89 0.66 0.89 0.89 |
| $V_3$ | 0.85 0.90 0.86 0.74 0.90 0.89 0.94 0.79 0.94 0.76 0.92 0.80 0.86 0.86 0.89 0.62 0.89 0.71 0.90 0.85 |
| $V_4$ | 0.96 0.78 0.95 0.81 0.90 0.67 0.96 0.74 0.94 0.77 0.88 0.86 0.82 0.92 0.72 0.90 0.76 0.89 0.87 |
| max  | 0.96 0.93 0.95 0.82 0.96 0.89 0.96 0.81 0.94 0.77 0.94 0.86 0.88 0.88 0.95 0.72 0.90 0.77 0.90 0.91 |

C Spatially Varying BRDF

We visualize the BRDF in different position of KNIGHT STANDING in Fig. 7. Although we are restricted by our specular model, we can still generate spatially varying BRDF by assigning different scaling factors $k_d \in [0, 1]$ to different spatial points. That is, $k_d$ controls the specularity of a certain point. The reflectance could be Lambertian if $k_d$ is 1 and non-Lambertian otherwise. For instance, we have a larger $k_d$ on the knight’s face and cloak, but a smaller $k_d$ on the armor, generating different BRDF and shading effects shown in Fig. 7.

![Figure 7: Visual illustration of BRDF at different positions. On the left is the observed image of STANDING KNIGHT. We select three different points on the object and show our predicted BRDF spheres of those points on the right side. We normalized the image and the predicted BRDF from 0 to 1 for a better illustration. For the right, sub-figures top-bottom are: BRDF for the point on the face, BRDF for the point on the armor, and BRDF for the point on the cloak, respectively.](image-url)
D Silhouette Constraint

For Lambertian objects, the GBR ambiguity is represented as:

\[
\begin{align*}
I^k &= B^\top S^k, \\
B^k &= \rho^d G^{-\top} N^s, \\
S^k &= e^k G l^k,
\end{align*}
\] (14)

Where, \( G \) is the \( 3 \times 3 \) ambiguity matrix, \( \rho^d \) is the albedo, \( N \) is the surface normal, \( e^k \) is the light intensity, \( l^k \) is the light direction. We denote the object’s silhouette normal as \( N^s \), the fitted silhouette normal as \( \bar{N}^s \), and the predicted silhouette normal by PositionNet as \( \hat{N}^s \) that contains GBR ambiguity, i.e, \( \bar{N}^s = G^{-\top} N^s \).

During training, the output of the LightNet is \( \bar{e}^k \) and \( \bar{l}^k \). The GBR ambiguity is solved by:

\[
\arg \min_{G} \sum_{k=1}^{f} (|I^k - \bar{B}^S \bar{S}^k| + |\text{Nor} (C(\bar{N}^s)) - \hat{N}^s|).
\] (15)

Where, \( f \) is the number of the light source, \( C(\cdot) \) cuts off the 3rd dimension of \( n_k \) (i.e, \( C(n_k) \in \mathbb{R}^{1\times2} \)), and \( \text{Nor}(\cdot) \) is the vector normalization operation. This can be easily extended to the non-Lambertian case with the shape-light ambiguity.

We also intuitively illustrate the effect of fitted silhouette normal in Fig. 8. The expression of the shape-light ambiguity during training is the synchronous rotation of estimated surface normal and light direction. While we fix the silhouette normal’s azimuth angle, these constraints help mitigate the ambiguity by aligning the predicted surface normal to the fitted silhouette normal’s azimuth, which suppresses the rotation of surface normal.

![Figure 8](image-url)

Figure 8: Visual qualitative comparison of the surface normal for different training epochs. Top: the surface normal maps. Bottom: the error maps. Note that the silhouette normal takes effects from the boundary.
Calculating $\hat{S}_j$

The ShadowNet implicitly records the interaction between the depth map and the light direction. To align the estimated surface normal (output of PositionNet) to the estimated shadow (output of ShadowNet), we establish the relationship between the normal map and shadow map. We achieve this by reconstructing the depth map from the predicted normal map and rendering the shadow map from the depth map and estimated light direction.

**Depth map reconstruction.** Reconstructing the depth map is also called normal integration, and many algorithms are proposed to deal with it. We apply the orthogonal five-point plane fitting method in [9] for depth map reconstruction. Because the normal map from our method might contain outliers due to our per-pixel manner solution, while method in [9] is more robust to outliers and sharp features than previous ones. Specifically, this method formulates normal integration as an inverse problem of local plane fitting in the camera coordinates by minimizing the perpendicular point-to-plane distance.

The method formulates the plane as

$$ p^\top n + d = 0 \quad (16) $$

where $n$ is the unit normal vector perpendicular to the plane, and $d$ is the distance from the coordinate origin to the plane. And the goal is to minimize the sum of the squared distance of all points to their nearby plane.

$$ \min_{p,d} \sum_{p \in \Omega} \sum_{p \in N(p(u))} (p^\top n(u) + d(u))^2 \quad (17) $$

where each point $p$ is parameterized by the known projection of the 3D point on the image plane $u$, and the 3D point’s unknown position on the camera ray $z$. And $N(p(u))$ is defined as five nearest neighbor points, including $p$ itself and its left / right / upper / lower counterparts.

**Shadow rendering.** We set out a light ray at the query point $x = (u, v, w)$ from the depth map, $r = x + tl$, where $l$ is the estimated light direction and $t$ is a parameter for sampling points. Then, a series of points (we sampled 32 pixel points) $X = \{(u_i, v_i, w_i) | 1 \leq i \leq 32\}$ on the ray can be sampled. We then get 32 depth values $\{\hat{w}_i | 1 \leq i \leq 32\}$ with the same $(u_i, v_i)$ in X from the depth map. The query point $x$ is determined to be either in the cast shadow or not through the equation:

$$ s = H(\min\{\hat{w}_i - \hat{w}_i | 1 \leq i \leq 32\}) \quad (18) $$

Where $H(\cdot)$ is the Heaviside step function.

**F Standard Deviation of Random Tests on Diligent [53] Benchmark**

By changing the random seeds, we provide the standard deviation for 10 objects on the Diligent [53] benchmark. The results are shown in Table 5.

Table 5: Quantitative results in terms of mean angular error for surface normal, light direction, and scale-invariant error for intensity on Diligent benchmark [53]. This table summarizes two versions. Rows 2-4 are the reported results in the main paper. Row 5-7 are the mean of 5 random tests.

| BALL | BEAR | BUDDHA | CAT | COW | GOBLET | HARVEST | POTT1 | POTT2 | READING | AVG | STD |
|------|------|--------|-----|-----|--------|----------|-------|-------|----------|-----|-----|
| Ours |      |        |     |     |        |          |       |       |          |     |     |
| norm.| 1.17 | 4.49   | 8.73| 4.89| 6.27   | 9.53      | 18.31 | 7.08  | 5.85     | 12.02| 7.83| 0.44 |
| dirs.| 1.79 | 3.54   | 2.33| 2.60| 5.81   | 8.45      | 7.40  | 3.73  | 2.10     | 7.91 | 4.57| 0.77 |
| ints.| 0.014| 0.011  | 0.034| 0.023| 0.196  | 0.045     | 0.032 | 0.071| 0.037    | 0.047| 0.051| 0.033|
G Implementation Details

In Fig. 9, we display the network architectures of PositionNet, LightNet, ShadowNet, and SpecularNet.

**Positional Encoding.** The positional encoding module is identical to NeRF \cite{mildenhall2020nerf}, which is based on the following equation \cite{mildenhall2020nerf}:

\[
E(p) = (\sin(2^0 \pi p), \cos(2^0 \pi p), \cdots, \sin(2^{L-1} \pi p), \cos(2^{L-1} \pi p)),
\]

where \( L \) is the dimension of the positional code, \( p \) is the encoded features. The \( p \) in our model is a vector, i.e., \((\cos \langle h_j, n_i \rangle, \cos \langle h_j, v \rangle)\) for SpecularNet, \((x, y)\) for PositionNet and ShadowNet, and \( l \) for ShadowNet. We compute each dimension’s positional code \( E(p) \) and concatenate these codes to get our positional code.

**Binarization Layer.** The binarization layer follows a similar design in \cite{yu2020ups}. Considering a step function \( B(\cdot) \),

\[
B(x) = \begin{cases} 
1 & x > 0.5 \\
0 & x \leq 0.5 
\end{cases}
\]

During training, the gradient of the binarization layer is defined as \( \nabla_x B(x) = 1 \) to make the ShadowNet trainable, which is similar to that in \cite{yu2020ups}.

![Network architectures of SpecularNet, PositionNet, ShadowNet, and LightNet.](image)

**Training details.** Our main framework is implemented in PyTorch, while the pre-calculation (method \cite{yu2020ups} and silhouette fitting) are implemented in MATLAB. We use Adam as the optimizer with a learning rate \( \alpha = 5 \times 10^{-4} \) to train our framework in 500 epochs for each scene separately, and the warm-up stage takes up 10 epochs. The last 100 epochs use a lower learning rate \( \alpha = 5 \times 10^{-5} \) for fine-tuning. The batch size is 32 for lighting and 256 for spatially random sampling. After every epoch, the depth map is reconstructed according to the predicted normal map by method \cite{yu2020ups}. Each scene takes from 2 hours to 6 hours on one RTX 2080Ti 12GB GPU, depending on the resolution of the objects. For Sparse UPS, due to the fewer reconstruction terms to train our NeIF, we slightly increase the dimension of the positional encoding module from 4 to 6 that increases the frequency \cite{yu2020ups} to stabilize the training, i.e., strengthening the role of positional code to reduce the variance of estimated intrinsics. Besides, we drop \( L_{ux} \) during early stage warm-up because YS97 \cite{yu2020ups} fails for the sparse inputs.

\[\text{https://github.com/facebookresearch/pytorch3d/tree/main/projects/nerf}\]
H Ablation Study

In this section, we evaluate the effectiveness of the loss functions in Eq. (10) and Eq. (11) and learning strategies. To be specific, we consider the ablation studies including:

- (1) without back-propagation from ShadowNet to LightNet;
- (2) without back-propagation from SpecularNet to LightNet;
- (3) without both mentioned back-propagation paths;
- (4) without early-stage supervised by YS97 [69];
- (5) without gradient penalty in the early-stage.

The quantitative results comparison is displayed in Table 6. The comparison with cases (1), (2), and (3) indicates the important roles of shadow and specular for uncalibrated photometric stereo as they provide clues of light estimation. The comparison with case (4) shows that early-stage weak supervision is necessary to prevent the neural networks from poor initialization. The comparison with case (5) illustrates the effectiveness of the gradient penalty. Without that, the prediction of $\rho_s$ could be less stable.

Table 6: Quantitative comparison in terms of normal, light direction, and light intensity errors on DiLiGenT dataset [53]. w/o $\rho_s$-back indicates without back-propagation from ShadowNet to LightNet; w/o $s$-back indicates without back-propagation from SpecularNet to LightNet; w/o YS97 [69] indicates no YS97 weak-supervision (i.e., $L_{az}$ in Eq. (10)) in the early-stage warm-up; w/o GP indicates no gradient penalty loss function (i.e., $L_{gp}$ in Eq. (10)) in the early stage warm-up.

|       | BALL | BEAR | BUDDHA | CAT | COW | GOBLET | HARVEST | POT1 | POT2 | READING | AVG  |
|-------|------|------|--------|-----|-----|--------|---------|------|------|---------|------|
| Ours  |      |      |        |     |     |        |         |      |      |         |      |
|       | norm. | 1.17 | 4.49   | 8.73 | 4.89 | 6.27   | 9.53    | 18.31 | 7.08 | 5.85    | 12.02|
|       | dirs. | 1.79 | 3.54   | 2.33 | 2.60 | 5.81   | 8.45    | 7.40  | 3.73 | 2.50    | 7.91 |
|       | ints. | 0.014| 0.011  | 0.034| 0.023| 0.196  | 0.045   | 0.032 | 0.071| 0.037   | 0.047|
| (1) w/o $\rho_s$-back | norm. | 1.07 | 6.66   | 13.26| 4.94 | 9.71   | 11.44   | 31.35 | 6.88 | 6.19    | 16.29|
|       | dirs. | 4.74 | 7.14   | 2.70 | 12.79| 13.10  | 13.10   | 11.07 | 6.67 | 2.13    | 19.54|
|       | ints. | 0.011| 0.011  | 0.099| 0.030| 0.148  | 0.043   | 0.477 | 0.044| 0.034   | 0.054|
|       | norm. | 1.04 | 5.32   | 12.59| 5.28 | 6.41   | 12.03   | 17.20 | 7.36 | 6.22    | 11.20|
|       | dir.  | 1.95 | 5.98   | 11.24| 4.00 | 6.25   | 16.18   | 5.74  | 7.12 | 2.01    | 7.04 |
|       | int.  | 0.011| 0.013  | 0.045| 0.024| 0.048  | 0.050   | 0.033 | 0.041| 0.037   | 0.043|
| (2) w/o s-back | norm. | 1.26 | 8.76   | 13.66| 7.17 | 13.56  | 11.68   | 12.35 | 7.84 | 6.37    | 16.67|
|       | dirs. | 2.28 | 13.66 | 7.41 | 8.23 | 19.50  | 15.12   | 11.57 | 7.47 | 2.82    | 20.66|
|       | ints. | 0.011| 0.028  | 0.106| 0.027| 0.200  | 0.048   | 0.277 | 0.017| 0.042   | 0.071|
| (3) w/o $\rho_s$-back | norm. | 1.29 | 9.31   | 10.39| 7.74 | 39.74  | 27.65   | 34.03 | 6.99 | 57.82   | 20.29|
|       | dirs. | 1.42 | 13.36 | 9.41 | 7.5  | 29.08  | 66.78   | 71.31 | 6.52 | 50.15   | 27.08|
|       | ints. | 0.013| 0.039  | 0.031| 0.224| 0.045  | 0.389   | 0.270 | 0.056| 0.448   | 0.124|
| (4) w/o YS97 [69] | norm. | 1.10 | 4.11   | 8.45 | 6.65 | 7.29   | 9.65    | 30.32 | 12.35| 6.83    | 16.08|
|       | dirs. | 1.65 | 3.36   | 2.39 | 7.94 | 9.24   | 11.05   | 6.02  | 11.35| 1.99    | 9.79 |
|       | ints. | 0.028| 0.011  | 0.030| 0.197| 0.042  | 0.041   | 0.079 | 0.146| 0.055   | 0.235|
| (5) w/o GP | norm. | 1.17 | 4.49   | 8.73 | 4.89 | 6.27   | 9.53    | 18.31 | 7.08 | 5.85    | 12.02|
|       | dirs. | 1.79 | 3.54   | 2.33 | 2.60 | 5.81   | 8.45    | 7.40  | 3.73 | 2.10    | 7.91 |
|       | ints. | 0.014| 0.011  | 0.034| 0.023| 0.196  | 0.045   | 0.032 | 0.071| 0.037   | 0.047|

I Statistical Shadow Handling

We remove the ShadowNet and train our method using the pseudo shadow maps for reconstruction loss. The pseudo shadow maps are obtained by binarizing the observed images, i.e., considering an observed pixel to be cast shadow if its intensity value is smaller than $0.2 \times$ the mean intensity values of this image. The result is shown in Table 7. Although pseudo shadow map works well on objects like BALL, BUDDHA, and CAT, the performance, the accuracy of the light directions, dropped significantly on objects like HARVEST, GOBLET, and READING. This is because pseudo shadow maps do not provide any extra clues for the LightNet through the back-propagation.

Table 7: Quantitative comparison in terms of normal error, light direction error, and light intensity error on DiLiGenT dataset [53]. ‘Ours w $s$’ indicates using pseudo shadow map instead of the ShadowNet for shadow handling.

|       | BALL | BEAR | BUDDHA | CAT | COW | GOBLET | HARVEST | POT1 | POT2 | READING | AVG  |
|-------|------|------|--------|-----|-----|--------|---------|------|------|---------|------|
| Ours  |      |      |        |     |     |        |         |      |      |         |      |
|       | norm. | 1.17 | 4.49   | 8.73 | 4.89 | 6.27   | 9.53    | 18.31 | 7.08 | 5.85    | 12.02|
|       | dirs. | 1.79 | 3.54   | 2.33 | 2.60 | 5.81   | 8.45    | 7.40  | 3.73 | 2.10    | 7.91 |
|       | ints. | 0.014| 0.011  | 0.034| 0.023| 0.196  | 0.045   | 0.032 | 0.071| 0.037   | 0.047|

20
J Qualitative and Quantitative Comparison

J.1 Results on DiLiGenT dataset [53]

Surface normal. Fig. 10 to Fig. 12 show the normal and error maps comparison with CW20 [14], CH19 [12], and TM18 [55]. As TM18 [55] did not provide any pre-trained models, we retrained the whole network and obtained a result that is slightly different from the original paper. Here we use the retrained models for visual quality comparison. The difference between the retrained and reported versions is shown in Table 8.

Figure 10: Error maps and surface normal of BALL, BEAR, BUDDHA
Figure 11: Error maps and surface normal of Cat, Cow, Goblet and Harvest
Figure 12: Error maps and surface normal of POT1, POT2, READING

Table 8: Quantitative comparison in terms of mean angular error for estimated surface normal map on DILIGENT dataset [53].

|          | BALL | BEAR | BUDDHA | CAT | COW | GOBLET | HARVEST | POT1 | POT2 | READING | AVG |
|----------|------|------|--------|-----|-----|--------|---------|------|------|---------|-----|
| TM18 [53] (retrained) | 1.31 | 7.74 | 11.75  | 5.52| 6.11| 11.21  | 22.46   | 6.62 | 7.69 | 10.87   | 9.35|
| TM18 [53] (reported)  | 1.47 | 5.79 | 10.36  | 5.44| 6.32| 11.47  | 22.59   | 6.09 | 7.76 | 11.03   | 8.83|
| Ours     | 1.17 | 4.49 | 8.73   | 4.89| 6.27| 9.53   | 18.31   | 7.08 | 5.85 | 12.02   | 7.83|
Light direction and intensity. Fig. 13 shows the results of light estimation for 10 objects.

| Object       | Ours (96) | CW20 | CH19 | Ours (96) | CW20 | CH19 |
|--------------|-----------|------|------|-----------|------|------|
| DILL         | 1.69      | 1.75 | 3.27 | 0.030     | 0.027| 0.039|
| BEAK         | 3.96      | 2.44 | 3.47 | 0.010     | 0.101| 0.061|
| BUDDHA       | 1.73      | 2.86 | 4.34 | 0.032     | 0.032| 0.048|
| CLEF         | 2.92      | 4.58 | 4.08 | 0.021     | 0.075| 0.095|
| COW          | 4.98      | 3.15 | 4.52 | 0.050     | 0.031| 0.073|
| GOBLET       | 6.82      | 2.98 | 10.36| 0.040     | 0.042| 0.067|
| HARVEST      | 7.06      | 5.74 | 6.32 | 0.032     | 0.065| 0.082|
| PORT1        | 3.33      | 1.41 | 5.44 | 0.134     | 0.039| 0.058|
| PORT2        | 3.71      | 2.81 | 2.87 | 0.028     | 0.059| 0.048|
| READING      | 7.45      | 5.47 | 4.50 | 0.042     | 0.048| 0.105|

Figure 13: Visual quality comparison of light direction and intensities’ error maps on DiLIGENT dataset [53].
Intrinsics visualization. Fig. 14 shows the results of estimated intrinsics for all 10 objects in DiLiGENT dataset [53].

Figure 14: Visual illustration of estimated intrinsics of the 10 objects from DiLiGENT dataset [53]. From left to right: images of objects, estimated shadow maps, diffuse reflectance maps, specular reflectance maps, diffuse coefficient maps, and specular coefficient maps. The diffuse reflectance maps are magnified 3 times, and all of the intrinsic maps are scaled to [0, 1] for better visualization.
J.2 Results on Apple & Gourd dataset [2]

In this section, we show the intrinsics obtained by NeIF in Apple, Gourd1 and Gourd2 in Fig. 15. We also report the MAE of light directions in Table 9.

Table 9: Quantitative comparison in terms of mean angular error for light direction and scale-invariant error for intensity on Apple & Gourd [2].

| Model  | Apple dirs. | Apple ints. | Gourd1 dirs. | Gourd1 ints. | Gourd2 dirs. | Gourd2 ints. | Avg dirs. | Avg ints. |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|----------|----------|
| PF14 [4] | 6.68        | 0.109       | 21.23       | 0.096       | 25.87       | 0.329       | 17.92    | 0.178    |
| CH19 [12] | 9.31        | 0.106       | 4.07        | 0.048       | 7.11        | 0.186       | 6.83     | 0.113    |
| CW20 [14] | 10.91       | 0.094       | 4.29        | 0.042       | 7.13        | 0.199       | 7.44     | 0.112    |
| Ours    | 2.65        | 0.011       | 1.76        | 0.029       | 3.21        | 0.230       | 2.54     | 0.090    |

J.3 Results on Light Stage Data Gallery [10]

In this section, we show the intrinsics obtained by NeIF in Standing Knight, Helmet Side Left and Plant. Light Stage Data Gallery is challenging because of the complex details for each object. Although the given hyper-parameters work well in most cases, we do observe that for some cases, such as Plant, which has complicated shapes and shadows, while with limited specular effects, the current setting may not be optimal. In such cases, we need to increase the update frequency of the reconstructed shadow maps from per epoch to per batch.

Table 10: Quantitative comparison in terms of mean angular error for light direction and scale-invariant error for intensity on Standing Knight, Helmet Side Left and Plant [10].

| Model  | Standing Knight dirs. | Standing Knight ints. | Helmet dirs. | Helmet ints. | Plant dirs. | Plant ints. | Avg dirs. | Avg ints. |
|--------|------------------------|------------------------|--------------|--------------|-------------|-------------|----------|----------|
| Ours   | 2.65                   | 0.011                  | 1.76         | 0.029        | 3.21        | 0.230       | 2.54     | 0.090    |
Table 10: Quantitative comparison in terms of light direction error and light intensity error on STANDING KNIGHT, HELMET SIDE LEFT and PLANT from LIGHT STAGE DATA GALLERY [10].

| Model | STANDING KNIGHT | HELMET SIDE LEFT | PLANT | AVG |
|-------|----------------|------------------|-------|-----|
|       | dirs. | ints. | dirs. | ints. | dirs. | ints. | dirs. | ints. |
| PF14 [14] | 33.81 | 1.31 | 25.4 | 0.576 | 20.56 | 0.227 | 26.59 | 0.705 |
| CH19 [12] | 11.6 | 0.286 | 6.57 | 0.212 | 16.06 | 0.17 | 11.41 | 0.223 |
| CW20 [14] | 5.31 | 0.198 | 5.33 | 0.096 | 10.49 | 0.154 | 7.04 | 0.149 |
| Ours | 13.38 | 0.189 | 8.12 | 0.082 | 18.04 | 0.464 | 13.18 | 0.245 |

K. Scale on the Input Images

The intensity of 96 images from READING are scaled by 96 variables generated by the the uniform distribution \( \mathcal{U}(0.01, 1) \), respectively. The new images are used as the training data to train our method. The results on READING is shown in Fig. 17, which further illustrates that our method can handle the varying light intensity and is free from data bias. While CW20 [14] (15.84 v.s. ours 11.05) fails because they have a pre-defined range on the intensities.

![Figure 17: Visual illustration of error maps of the normal. From left to right: error maps of READING, normal map of READING. Top-bottom: our result and CW20 [14]'s result.](image-url)
L Comparison on Data Collected under Casual Environments

To further highlight the advantage of our method regarding unsupervised learning and light-normal joint optimization, we perform an experimental comparison on data collected in casual environments.

**Captured objects.** We use three objects for this experiment, including BUNNY, VENUS, and MOUSE. BUNNY contains lots of fine details, with a broad specular lobe on a uniform material (phenolic resin). VENUS is made up of glass for the pearl on the tray and gypsum for the body. MOUSE has lots of defects on the shell, made up of a spatially varying but mostly diffuse material. The illustration of these objects can be found in Fig. 19.

**Data capture.** We use the iPhone 13 Pro Max camera fixed on a phone tripod for data capture. We turn the HDR mode on and set the exposure compensation to -2 to avoid overexposure. We record a video with a raw resolution at $1920 \times 1080$. During video recording, we move the light source slowly to illuminate the object at different angles and keep the trajectory parallel to the image plane. Fig. 18 shows the capture equipment. BUNNY and VENUS are shot at about 1 meter away with the light at about 1-1.5 meters away. MOUSE is shot at about 30cm away with light at about 40cm away.

**Ambient light and light sources.** All data are captured with ambient light because the full-dark chamber is also expensive for casual users. For the data capture of BUNNY and VENUS, we control the impact of the ambient light by changing the intensity of the electric torch, i.e., weak impact (or strong light source intensity) for BUNNY, and strong impact (or weak light source intensity) for VENUS. The capture environment for these objects is relatively controlled, and the ambient light is from a small window (see Fig. 18). For the data capture of MOUSE, we consider a more casual and challenging scenario (see Fig. 18). That is, the ambient light is more dominant and uncontrolled. We use a flashlight from the mobile phone to further relax the assumption of directional light in photometric stereo.

**Data processing.** We use MATLAB 2020 to extract 100 frames uniformly from each video and downsample them to $960 \times 540$. We extract the mask of each object via PhotoShop 2020.

Figure 18: Data collection process. From top to the bottom, Row 1: the equipment we use in scene 1 to capture the VENUS and BUNNY; Row 2: the equipment we capture MOUSE in scene 2.

**Comparison methods.** We compare our method with CW20 [14]. As LL22 [35] is not for UPS, we feed it by the estimated light from the state-of-the-art method CW20 [14] and denoted it as CW20 [14]+LL22 [35].

L.1 Comparison of Surface Normal

The estimated normal map for each object is shown in Fig. 4. However, CW20 [14] and CW20 [14]+LL22 [35] is sensitive to the data bias in supervised learning of light estimation model.
The error of estimated light dramatically degrades the performance of normal estimation. In contrast, our method is robust to casual environments. We have a more reasonable estimation on BUNNY with the ambient light and MOUSE in the challenging scene and much better results on VENUS as compared with CW20 [14] and CW20 [14]+LL22 [35]. Since our model adopts the directional light assumption, it produces inaccurate light. The superior performance advantage on surface normal estimation indicates that our method could better balance the accuracy of the surface normal and light due to its light-normal joint optimization manner.

![Visual quality comparison in terms of normal map on BUNNY (row 1), MOUSE (row 2), and VENUS (row 3) from DILIGENT [53]. For each subfig, from left to right: mean of the observed images, coarse normal map download in https://sketchfab.com/, that has similar shape with the objects for reference, normal map from our method, CW20 [14], LL22 [35]+ CW20 [14].](image)

**L.2 Comparison of Light Estimation**

We further make a comparison on the estimation of light. The predicted light direction projected on the XY-plane is shown in Fig. 20 and represents the estimated light intensity by color (note that the light intensity is normalized for better visualization). Since the intensity of our light source during data collection is unchanged, our method produces much more accurate light intensity than the comparison methods. Besides, we change the light direction by moving the light source at a large distance. The trajectory recovered by our method is more reasonable than the comparison methods, especially for objects of MOUSE and VENUS. These results indicate that our method is more robust to training data bias due to its unsupervised manner.

![Visual comparison in terms of the estimated light trajectory on BUNNY(column 1), MOUSE (column 2), and VENUS (column 3). For each sub-figure, top-bottom: Ours, CW20 [14]’s predicted light trajectory that projected onto the XY-plane, the color indicates the value of the light intensity.](image)
M Sparse Uncalibrated Photometric Stereo

M.1 Quantitative Results on DILiGENT dataset [53]

We evaluate the proposed NeIF for sparse uncalibrated photometric stereo. We use images under sparse lights (10 and 16 randomly selected lights with different seeds) for normal and light estimation. The results of 30 random trials are shown in Fig. 21 and Table 11. As can be found, there is a large fluctuation of NeIF’s performance for normal estimation on HARVEST that contains a large region of cast shadow under most lights, which has the most difficult geometries to estimate due to the limited information (i.e., the impact of the cast shadow is more significant under sparse lights). Compared to the calibrated and supervised method [74] and uncalibrated method [44], NeIF shows a promising performance on average for 10 lights (11.62 for normal estimation, 19.29 for light direction, and 0.199 for intensities) and 16 lights (10.38 for normal estimation, 7.21 for light direction, and 0.047 for intensities).

Figure 21: Estimation results of NeIF given only 10 lights ((a), (c), (e)) and 16 lights ((b), (d), (f)) on DILiGENT [53] benchmark regarding surface normal ((a), (b)), light direction ((c), (d)), and light intensity ((e), (f)). The X-axis indicates object names. The Y-axis is the MAE (mean angular error, in degree) of the surface normal, light direction, or scale-invariant relative error of intensities, displayed using the box-and-whisker plot. The red line indicates the median value, the upper and the lower bound of the blue box are the first and third quartile values, respectively, and the top and bottom end of the black lines indicate the minimum and maximum.

Table 11: Quantitative comparison in terms of normal map error, light direction error, and light intensity error on DILiGENT dataset [53]. The numbers in brackets indicate the number of lights. For example, Ours (10) indicates that our method takes 10 images under different lights as the input.

|          | BALL | BEAR | BUDDHA | CAT | COW | GOBLET | HARVEST | POT1 | POT2 | READING | AVG |
|----------|------|------|--------|-----|-----|--------|---------|------|------|----------|-----|
| ZJ19 (10) |      |      |        |     |     |        |         |      |      |          |     |
| norm.    | 4.96 | 5.99 | 10.07  | 7.52| 8.80| 10.43  | 19.05   | 8.77 | 11.79| 16.13    | 10.35|
| dirs.    | 14.31| 13.38| 18.32  | 18.37| 16.48| 24.43  | 25.99   | 22.80| 17.14| 26.60    | 19.74|
| ints.    | 0.145| 0.121| 0.150  | 0.232| 0.269| 0.244  | 0.272   | 0.182| 0.139| 0.299    | 0.205|
| ZJ19 (10) |      |      |        |     |     |        |         |      |      |          |     |
| norm.    | 4.38 | 5.79 | 9.60   | 7.13| 7.87| 10.00  | 18.35   | 8.41 | 11.20| 15.45    | 9.82 |
| dirs.    | 14.31| 13.38| 18.32  | 18.37| 16.48| 24.43  | 25.99   | 22.80| 17.14| 26.60    | 19.74|
| ints.    | 0.145| 0.121| 0.150  | 0.232| 0.269| 0.244  | 0.272   | 0.182| 0.139| 0.299    | 0.205|
| Ours (10) |      |      |        |     |     |        |         |      |      |          |     |
| norm.    | 2.49 | 4.90 | 11.58  | 7.47| 10.21| 14.45  | 27.30   | 11.41| 6.57 | 19.85    | 11.62|
| dirs.    | 14.31| 13.38| 18.32  | 18.37| 16.48| 24.43  | 25.99   | 22.80| 17.14| 26.60    | 19.74|
| ints.    | 0.145| 0.121| 0.150  | 0.232| 0.269| 0.244  | 0.272   | 0.182| 0.139| 0.299    | 0.205|
| Ours (16) |      |      |        |     |     |        |         |      |      |          |     |
| norm.    | 2.34 | 4.52 | 10.33  | 7.87| 7.68 | 11.23  | 25.07   | 8.00 | 6.35 | 20.42    | 10.38|
| dirs.    | 2.83 | 4.14 | 5.77   | 8.45| 8.91 | 10.09  | 6.44    | 6.16 | 3.50 | 15.85    | 7.21 |
| ints.    | 0.012| 0.014| 0.033  | 0.038| 0.047| 0.062  | 0.041   | 0.062| 0.026| 0.136    | 0.047|
| Ours (96) |      |      |        |     |     |        |         |      |      |          |     |
| norm.    | 1.17 | 4.49 | 8.73   | 8.73| 6.27 | 9.53   | 18.31   | 7.83 | 5.85 | 12.02    | 7.83 |
| dirs.    | 1.79 | 3.54 | 2.33   | 2.60| 5.81 | 8.45   | 7.40    | 3.73 | 2.10 | 7.91     | 4.57 |
| ints.    | 0.014| 0.011| 0.034  | 0.023| 0.196| 0.045  | 0.032   | 0.071| 0.037| 0.047    | 0.051|

†Paper reported version, using different lights compared with us.
‡Re-test version, using same lights compared with us.
M.2 Qualitative Results on DiLiGENT dataset

We compared our 10 lights and 16 lights models with our 96 lights models and PF14 [44] under the same light setting, as shown in Fig. 22, Fig. 23, Fig. 24. Note that sparse uncalibrated photometric stereo results are from a set of random lights.

Figure 22: Error maps and surface normal of BALL, BEAR, BUDDHA
Figure 23: Error maps and surface normal of CAT, COW, GOBLET and HARVEST
Figure 24: Error maps and surface normal of POT1, POT2, READING
N. Time Analysis

KK21 [33] did not release their code. Therefore, we report the test time (s) for TM18 [55], LL22 [35] and ours in Table 12 under different numbers of input, respectively. Our method has the fastest test time, given 96 images as input and the competitive test time given 16 images as input among the mentioned methods. The tests are implemented on an RTX3090 GPU, and the batch size is full size for lighting and 2048 for spatially sampling. The larger the batch size for spatially sampling, the faster the test, but the GPU memory occupation will also increase. According to our experiments, the parameters we choose are the most cost-efficient, which will occupy around 3.2GB on RTX3090.

The average training time for our method is about 4 hours on RTX 2080 Ti 12GB GPU, while [55] has an average training time of 1 hour, [33] has an average training time of 1 hour, and [35] has an average training time of 6 minutes. The slow training time of our method is mainly because our depth reconstruction algorithm runs on the CPU and joint optimization of LightNet’s CNN Encoder in a per-pixel way.

Table 12: Quantitative comparison in terms of test time (s) on DILIGENT benchmark [53]. Numbers in brackets indicate the number of lights. For example, Ours (96) indicates that our method takes 96 images under different lights as the input.

|                | BALL | BEAR | BUDDHA | CAT | COW | GOBLET | HARVEST | POT1 | POT2 | READING | AVG  |
|----------------|------|------|--------|-----|-----|--------|---------|------|------|---------|------|
| TM18 (96) [55] | 0.21 | 0.70 | 0.79   | 1.02| 0.48| 1.07   | 1.07    | 1.18 | 0.84 | 0.54    | 0.79 |
| LL22 (96) [35] | 1.22 | 0.92 | 1.17   | 1.36| 0.85| 1.26   | 1.23    | 1.08 | 1.14 | 1.06    | 1.13 |
| Ours (96)      | 0.29 | 0.70 | 0.83   | 0.82| 0.50| 0.46   | 1.04    | 1.10 | 0.65 | 0.52    | 0.69 |
| TM18 (16) [55] | 0.04 | 0.11 | 0.12   | 0.17| 0.08| 0.22   | 0.17    | 0.17 | 0.14 | 0.13    | 0.13 |
| LL22 (16) [35] | 0.62 | 0.69 | 0.69   | 0.68| 0.67| 0.71   | 0.66    | 0.69 | 0.77 | 0.68    | 0.68 |
| Ours (16)      | 0.27 | 0.67 | 0.80   | 0.83| 0.49| 0.49   | 1.08    | 1.09 | 0.63 | 0.52    | 0.69 |