Spin-UP: Spin Light for Natural Light Uncalibrated Photometric Stereo

Zongrui Li[1,2,*] Zhan Lu[2,4,*,†] Haojie Yan[3,4] Boxin Shi[5,6] Gang Pan[3,4] Qian Zheng[3,4,‡] Xudong Jiang[1,2]

[1]Rapid-Rich Object Search (ROSE) Lab, Interdisciplinary Graduate Programme, Nanyang Technological University.
[2]School of Electrical and Electronic Engineering, Nanyang Technological University
[3]College of Computer Science and Technology, Zhejiang University
[4]The State Key Lab of Brain-Machine Intelligence, Zhejiang University
[5]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[6]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
{zongrui001,zhan007,EXDJiang}@ntu.edu.sg, {hjyan,qianzheng,gpan}@zju.edu.cn, shiboxin@pku.edu.cn

Abstract

Natural Light Uncalibrated Photometric Stereo (NaUPS) relieves the strict environment and light assumptions in classical Uncalibrated Photometric Stereo (UPS) methods. However, due to the intrinsic ill-posedness and high-dimensional ambiguities, addressing NaUPS is still an open question. Existing works impose strong assumptions on the environment lights and objects' material, restricting the effectiveness in more general scenarios. Alternatively, some methods leverage supervised learning with intricate models while lacking interpretability, resulting in a biased estimation. In this work, we proposed ***Spin*** *Light* ***U****ncalibrated* ***P****hotometric Stereo (****Spin-UP****), an unsupervised method to tackle NaUPS in various environment lights and objects. The proposed method uses a novel setup that captures the object's images on a spinning platform, which mitigates NaUPS's ill-posedness by reducing unknowns and provides reliable priors to alleviate NaUPS's ambiguities. Leveraging neural inverse rendering and the proposed training strategies, Spin-UP recovers surface normals, environment light, and isotropic reflectance under complex natural light. Experiments have shown that Spin-UP outperforms other supervised / unsupervised NaUPS methods and achieves state-of-the-art performance on synthetic and real-world datasets. Codes and data are available at https://github.com/LMozart/CVPR2024-SpinUP.*

1. Introduction

Natural light uncalibrated photometric stereo (NaUPS) [6] is proposed to relieve the dark room and directional light assumption in classical uncalibrated photometric stereo, aiming to reconstruct the surface normal given images of an object captured at arbitrary environment light. The implications of NaUPS are far-reaching: it makes photometric stereo universal. However, solving NaUPS is still an open question because of the intrinsic ill-posedness introduced by the varying light of each image and the high dimensional ambiguities between the light and objects [6].

Previous optimization-based methods use the simple light model to represent the varying environment lights and Lambertian reflectance to represent the material [6, 7, 21]. These models help mitigate the ill-posedness and ambiguities to some extent but become ineffective in handling objects with general reflectance (*e.g.*, non-Lambertian reflectance) under complex environmental lighting, leading to unsatisfactory reconstruction outcomes. Besides, since they solely model the varying lights in each image, the unknowns introduced by the light model may increase with the resolution and numbers of the images, restricting their method to low-resolution and insufficient images.

Considering the difficulties of explicitly mitigating the ill-posedness and ambiguities, recent advances [10, 11] turn to data-driven methods. Specifically, they train a deep learning model on large-scale datasets and implicitly exploit deep light features from images to improve performance. Those methods lack interpretability, making them hard to constrain during training. Consequently, the model may be affected by the data bias and prone to specific types of light sources [10] or reflection variations [11] among images.

Despite a persistent exploration in this research field, a method capable of handling general objects under natural light while free from data bias is still missing. In this paper, we provide a new perspective to solve NaUPS. Specifically, we propose a novel setup that acquires images on a rotatable platform under a static environment light. In such cases, the object is illuminated by rotated environment light.

[*] Co-first author. ‡Corresponding author.

[†] Work completed while interning at the State Key Lab of Brain-Machine Intelligence, Zhejiang University.

The unknowns of light representation are significantly reduced since we model the light as a uniform environment light multiplied by low degree-of-freedom (DOF [5]) rotation matrices. This allows us to implement advanced parametric light models (*e.g.*, spherical Gaussian) and isotropic reflectance models to handle general scenarios. Additionally, based on such a setup, we further derive a reliable light initialization method by analyzing the pixel value at object's occluding boundary. Such light initialization will help the model converge at the beginning, thereby alleviating the ambiguity between light and objects during training.

With the help of the proposed setup and light initialization method, we develop **Spin** Light **U**ncalibrated **P**hotometric Stereo (**Spin-UP**), addressing NaUPS by optimizing inverse rendering framework in an unsupervised manner. To our best knowledge, this is the first unsupervised method that can handle general objects under natural light. Unlike previous methods, Spin-UP can jointly reconstruct arbitrary environment light, isotropic reflectance, and complicated shape with low GPU memory (5GB) and reasonable running time (around 25 minutes). Such low cost is attributed to two proposed training strategies: interval sampling and shrinking range computing. Experiments on synthetic and real-world datasets demonstrate our superior performance over previous methods on general scenarios. In a nutshell, our contributions are summarized as follows:

1. We design a novel setup for NaUPS, which reduces unknowns of light representation and facilitates solving NaUPS in an unsupervised manner.
2. We introduce a light prior, which leverages an object's occluding boundaries to initialize a reliable environment light. Based on the setup and light prior, the unsupervised NaUPS method Spin-UP is proposed.
3. We present two training strategies for robust optimization and fast convergence of Spin-UP.

2. Related Work

In this section, we briefly review recent supervised and unsupervised NaUPS methods. We also summarize other techniques that exploit priors from occluding boundaries. Additionally, we discuss recent advances in 3D vision to distinguish our techniques from other neural inverse rendering approaches. Note that there is a group of works reconstructing 3D surfaces from a single image by deep learning under natural light [15] or shading [4, 12, 32]. Those works suffer from extreme ambiguities and poor reconstruction quality on general objects and fall beyond the scope of this paper.

Natural Light Uncalibrated Photometric Stereo. Unsupervised NaUPS methods jointly recover the light, reflectance properties, and surface normal. These methods explicitly model the environment light by low-order spherical harmonics (SH) [2, 21], spatially varying spherical harmonics (SV-SH) [7, 18] or equivalent directional light [6] to mitigate the ill-posedness, and use integrability constraint [2], shape initialization [7], non-physical lighting regularization [21], or graph-based ambiguity relaxation [6] to alleviate the ambiguity. In contrast, supervised NaUPS methods [10, 11] apply deep learning models like transformers to reconstruct normal maps without explicitly estimating the environment light. The models are trained on a dataset containing images of diverse objects captured under various lighting conditions, including directional, point, and environmental light. Compared to previous work, the proposed Spin-UP distinguishes itself in three key facets: 1) it features a novel setup explicitly designed to model correlations among observed images to mitigate the ill-posedness of NaUPS, 2) leveraging this unique setup, a novel light initialization method is introduced to mitigate ambiguities, and 3) an advanced light and material model is implemented to address a broader range of scenarios.

Priors from the Boundaries. The occluding boundaries of an object are considered to reveal adequate information about the object's shape and the scene's light. Given the fact that the projection of the boundaries' normal to xy-plane is perpendicular to the boundaries in orthographic projection, methods are developed to constraint the surface normal estimation during iterative optimization [16] or recover a rough shape to initialize the geometry in multi-view [8] or photometric stereo [7]. Other methods associate the boundaries normal with the reflectance to estimate a rough position of the directional lights [27]. However, none of them derive the environment light from the boundary reflectance. Given the setup in Spin-UP, we can roughly estimate the environment by analyzing occluding boundaries and the corresponding pixel points. This approach provides a reliable initialization for lighting that alleviates ambiguity in NaUPS.

Inverse rendering in 3D Vision. Neural Radiance Fields (NeRF) [19] implicitly store the scene's shape and reflectance through MLPs optimized by inverse volume rendering. While NeRF can only recover coarse 3D shapes, several subsequent works [3, 25, 26, 29, 31, 33] have been proposed to combine the surface rendering and volume rendering techniques, recovering fine shapes under varying viewpoints but static environment light. In contrast, viewpoints in Spin-UP are relatively static to the objects. While most neural field methods aim to recover the whole 3D geometries, Spin-UP only recovers the object's surface.

3. Proposed Method

In Sec. 3.1, we explain the Spin-UP's setup and how it reduces the unknowns. In Sec. 3.2, we introduce the light prior that alleviates ambiguities in NaUPS, including details of the light initialization method based on that prior. In Sec. 3.3, we describe the implementation details of the proposed Spin-UP framework and losses. In Sec. 3.4, we demonstrate two proposed training strategies.

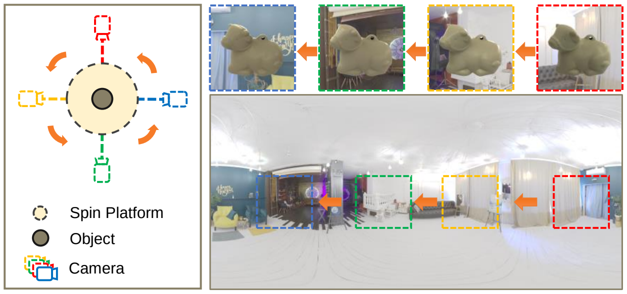

Figure 1. The proposed image capturing setup. Left: an illustration of image-capturing equipment consisting of a rotatable platform, a camera, and the target object. We spin the platform in 360° and capture images of the object. The object and camera rotate together with the platform. Right-top: Four observed images. Right-bottom: The ground truth environment light. Dashed color boxes indicate the camera views at different positions.

3.1. Spin Light Setup

As shown in Fig. 1, we capture a sequence of images $\boldsymbol{I} \triangleq \{I_j | j \in [0, 1, ..., N_I]\}$ for an object by rotating it together with a linear perspective camera in 360° on a rotatable platform[1]. Since the relative positions and orientations between the camera and the object are fixed during rotation, each observed image is aligned with the rotated environment light $\boldsymbol{L}(\boldsymbol{R}_j \cdot \boldsymbol{\omega})$, where $\boldsymbol{\omega} \in \mathbb{R}^3$ indicates the incident light's direction, $\boldsymbol{R}_j = \boldsymbol{R}(\theta_j)$ is the 1-DoF [5] rotation matrix representing rotation about the vertical axis, θ_j is the rotation angle, $\theta_0 = 0$. As we control unidirectional rotation and assume a constant velocity (not strictly required in practice), $\boldsymbol{R}_j$ can be initialized by $\theta_j = 2\pi j(N_I - 1)/N_I$. Given a sequence of images $\boldsymbol{I}$ and the initialized $\boldsymbol{R}$, Spin-UP iteratively optimizes the normal map $\boldsymbol{N}$, the environment light $\boldsymbol{L}$, isotropic BRDF map $\boldsymbol{M}$, and rotation angle $\boldsymbol{R}$ by solving

$$\underset{\boldsymbol{L},\boldsymbol{M},\boldsymbol{N},\boldsymbol{R}}{\arg\min} \sum_{i=1}^{N_P} \sum_{j=1}^{N_I} \mathcal{E}\left(\boldsymbol{m}_{ij}, \hat{\boldsymbol{m}}_{ij}\right), \tag{1}$$

where N_P is the number of sampled points on the surface, $\boldsymbol{m}_{ij}$ and $\hat{\boldsymbol{m}}_{ij}$ indicate the ground truth and estimation of point i's color in image I_j, respectively. $\mathcal{E}(\cdot, \cdot)$ is loss function between $\boldsymbol{m}_{ij}$ and $\hat{\boldsymbol{m}}_{ij}$ (*i.e*, mean absolute error). We adopt the rendering equation to calculate the color $\hat{\boldsymbol{m}}$[2]

$$\begin{aligned}\hat{\boldsymbol{m}} &= \int_{\Omega} s\boldsymbol{L}(\boldsymbol{\omega})\,\rho\,(\boldsymbol{\omega} \cdot \boldsymbol{n})\,\mathrm{d}\boldsymbol{\omega}, \\ &= \int_{\Omega} s\boldsymbol{L}(\boldsymbol{\omega})\,(\rho^s + \rho^d)\,(\boldsymbol{\omega} \cdot \boldsymbol{n})\,\mathrm{d}\boldsymbol{\omega}.\end{aligned} \tag{2}$$

where Ω represents the upper hemisphere centered at the normal vector $\boldsymbol{n}$ and s is the cast shadow. ρ^s and ρ^d indicate the specular and diffuse reflectance, respectively. The ambiguities between the light $\boldsymbol{L}$ and reflectance of the object $\boldsymbol{M}$ are often disregarded [13, 16].

Table 1. A comparison of unknowns' number among Spin-UP and representative unsupervised NaUPS methods.

	Ours	QL15 [21]	HY19 [7]	GM21 [6]
unknowns	434	1.5K	450	1.4M

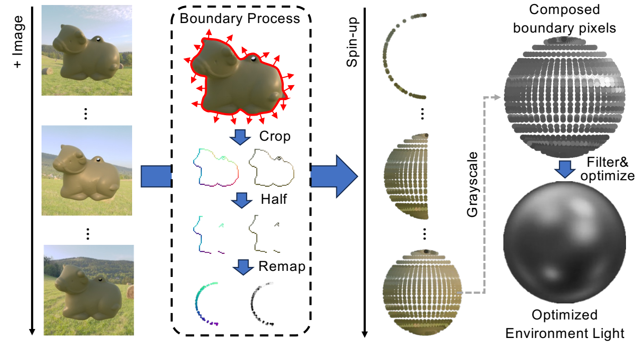

Figure 2. The proposed light initialization method in Spin-UP. We crop the boundary pixels m_b and normal $\boldsymbol{n}_b$ from input images. Here, we use half boundary points as we experimentally find that this improves reconstruction accuracy. Then, we remap them on the sphere and spin them with their corresponding rotations R. Based on a light probe composed of gray-scale boundary pixels, we optimize the SG light model to obtain the environment light.

Unknowns reduction. The proposed spin light setup reduces the unknowns of the light representation $\boldsymbol{L}$ by exploiting correlations between different images. In contrast to previous NaUPS methods that separately model the light for each image, we consider an environment light $\boldsymbol{L}$ represented by the parametric model like spherical Gaussian and 1-DoF rotation angle θ for each image. As such, the unknowns consist of the environment light model's parameters and the number of rotational angles that are quantitatively equal to N_I. The total unknown amount is reduced compared to other methods (Table 1), which helps mitigate the ill-posedness and facilitates solving NaUPS with advanced light and reflectance models in an unsupervised manner.

3.2. Light Prior from Boundaries

Based on the spin light setup, we can exploit priors from the object boundary for light initialization to alleviate the ambiguity. The idea is motivated by the observation that the pixel value m_b at an object's boundary provides insights into the environment light (see Fig. 2). For an object with occluding boundaries, the normal of those boundaries $\boldsymbol{n}_b$ can be pre-computed [13, 16]. By bonding m_b, $\boldsymbol{n}_b$, and $\boldsymbol{R}$, we can roughly derive a light map indicating the light sources' positions and intensities, where m_b directly represent the light intensity $\boldsymbol{L}(\boldsymbol{\omega}_b)$ at $\boldsymbol{\omega}_b = \boldsymbol{R} \cdot \boldsymbol{n}_b$. However, the derived light map for objects with different materials may contain mismatched light source positions and chromatic bias, leading to inaccurate light initialization.

The mismatched light source positions are caused by

[1] We assume a geometrically smooth around boundary (occluding boundary). The rotatable platform can be extended to fit various scenarios.

[2] The subscripts are omitted for simplicity.

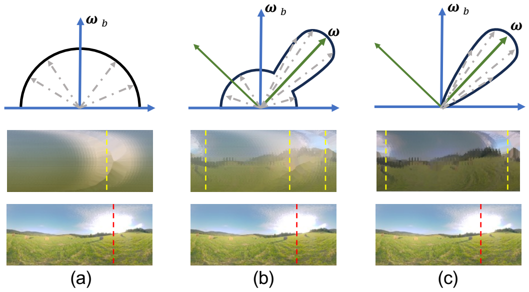

Figure 3. An illustration of mismatched light source positions. Rows from top to bottom: an illustration of reflection on materials with different roughness; the initial environment light before applying any filters given objects' boundary pixels; the ground truth environment light. Objects are (a) the diffuse-dominant sphere, (b) the diffuse and specular mixture sphere, and (c) the specular dominant sphere. The yellow and red lines in rows 2 and 3 indicate a rough position of the light sources.

the specular component m_b^s in m_b. When m_b^s dominates, approximating $\boldsymbol{L}(\omega_b)$ by m_b becomes a biased estimation as m_b^s is a reflection of lights in different directions than ω_b Fig. 3. By contrast, the approximation is more reasonable when the diffuse component (m_b^d) dominates since $m_b^d = \int_\Omega \boldsymbol{L}\left(\boldsymbol{\omega}\right) \rho^d \left(\boldsymbol{\omega} \cdot \boldsymbol{\omega}_b\right) \mathrm{d}\boldsymbol{\omega}$ indicates that $\boldsymbol{L}(\boldsymbol{\omega}_b)$ contributes most to the actual pixel value, making it less biased to use m_b^d to represent $\boldsymbol{L}(\boldsymbol{\omega}_b)$. Therefore, to conduct a less biased estimation of environment light for initialization, a diffuse filter $\mathcal{F}^d(.)$ is necessary on m_b^d to relieve the bias of mismatched light source positions. Similarly, a chromatic filter $\mathcal{F}^c(.)$ is also required on m_b^d to relieve the chromatic bias caused by the spatially varying material at boundaries. The filtered pixel value $\hat{m}_b^d = \mathcal{F}^c(\mathcal{F}^d\left(m_b\right))$ are the basis for our light initialization method.

Light initialization method. The procedure of light initialization method is summarized in Algorithm 1. This method aims to derive an initialized environment light model with parameter Θ. Specifically, we use $N_L = 64$ spherical Gaussian (SG) bases [28] as the light model, where $\boldsymbol{L}(\boldsymbol{\omega}|\boldsymbol{\xi}_t, \lambda_t, \boldsymbol{\mu}_t) = \sum_{t=1}^{N_L} G(\boldsymbol{\omega}; \boldsymbol{\xi}_t, \lambda_t, \boldsymbol{\mu}_t)$. $\boldsymbol{\xi}_t$, λ_t, and $\boldsymbol{\mu}_t$ stands for Gaussian lobes' direction, amplitude, and sharpness, respectively[3]. Inspired by the conclusion in [23] that diffuse reflectance can be approximated by the low-frequency reflectance, we design $\mathcal{F}^d(.)$ as a threshold filter $\mathcal{F}^d_{TH}(.)$ that remove pixels value at a point with intensity higher than 80 percent of the point's intensity profile [23] plus a low-pass filter (*i.e*, 3-order spherical harmonics filters[4]) noted as $\mathcal{F}^d_{SH}$ [22, 23]. $\mathcal{F}^d_{TH}$ further reduces the bias by removing the brightest parts in observed images, usually attributed to the specular reflectance. $\mathcal{F}^d_{SH}$ helps estimate the low-frequency reflectance. Moreover, we design $\mathcal{F}^c$[5] as the converter transferring pixel values into gray-scale, mitigating biases from spatially varying material.

Algorithm 1 Light Initialization Method

```
Input: pre-computed boundary normal $\boldsymbol{n}_b$, relative rotation matrices $\{\boldsymbol{R}_1 ..., \boldsymbol{R}_{N_I}\}$, boundary pixels $\{m_{b1}, ..., m_{bN_I}\}$, diffuse filter $\mathcal{F}^d_{TH}(.)$ and $\mathcal{F}^d_{SH}(.)$, chromatic filter $\mathcal{F}^c(.)$, fitting epochs $N$, environment light with parameters $\boldsymbol{L}(.|\Theta)$, learning rate $\eta$.
Output: An initialized environment lights $\boldsymbol{L}(\boldsymbol{\omega}_b)$.
for $j = 1, ..., N_I$ do
    $\boldsymbol{\omega}_{bj} = \boldsymbol{R}_j \cdot \boldsymbol{n}_b$
$\hat{m}_b^d = \mathcal{F}^c(\mathcal{F}^d_{SH}(\mathcal{F}^d_{TH}(m_b)))$
for $e = 1, ..., N$ do
    $J = \sum(\boldsymbol{L}(\boldsymbol{\omega}_b|\Theta) - \hat{m}_b^d)^2$
    $\Theta \leftarrow \Theta - \eta\partial J/\partial\Theta$
```

3.3. Framework of Spin-UP

With reliable initial SG lights and rotation matrices $\boldsymbol{R}$, we develop Spin-UP based on neural inverse rendering [3, 13, 16, 31] given the rendering equation in Eq. (2).

Shape model. We use the neural depth field to represent the 3D surface. An MLP[6] predicts the depth value given the surface point's image coordinates as the input. The coordinate will first be passed to a positional encoding module [19] with a frequency band equal to 10. then feed to an 8-layer MLP. To compute normal given the surface's depth map, we extend the normal fitting method described in [16] to the perspective projection[7].

Material model. We represent the spatially varying, isotropic reflectance as a simplified Disney Model [31]. The diffuse albedo $\boldsymbol{\rho}^d$ is predicted by another MLP with a similar structure given the query surface point's 2D image coordinate. The frequency band for this MLP's positional encoding module is 6. The spatially varying specular reflectance is calculated as a weighted sum of $N_S = 12$ SG bases, so $\boldsymbol{\rho}^s = \sum_{n=1}^{N_S} c^n \mathcal{D}\left(\boldsymbol{v}, \boldsymbol{\omega}\right) \mathcal{F}\left(\boldsymbol{h}, \boldsymbol{\omega}\right) \mathcal{G}(\boldsymbol{n}, \boldsymbol{\omega}, \boldsymbol{v}, \lambda_n)$, where $\mathcal{D}$, $\mathcal{F}$, and $\mathcal{G}$ accounts for micro-facet's normal distribution, Fresnel effects, and local shadow, respectively, $\boldsymbol{v}$ is the view direction, $\boldsymbol{h}$ is the half-vector, calculated by $\boldsymbol{h} = (\boldsymbol{v} + \boldsymbol{\omega}) / \left\|\boldsymbol{v} + \boldsymbol{\omega}\right\|$, λ_n is the roughness terms initialized as $(0.1 + 0.9(n-1))/(N_S - 1)$ and set as learnable parameters, c^n is the weights predicted by the MLP.

Shadow model. We apply a shadow mask similar in [13] to handle the cast shadow.

Loss functions. Similar to other inverse rendering-based methods [13, 16], we use the inverse rendering loss to train

[3] In practice, we find that initializing Gaussians' parameters by Fibonacci lattice [31] and freezing λ_t gives the best results.

[4] We implement a Gaussian filter and rescale the pixel value's range to $\mathcal{F}^d_{TH}(m_b)$'s range to suppress ringing effect and negative energy in SH.

[5] Despite a gray-scale initialization, we still optimize the light model in RGB channels.

[6] Please refer to the supplementary material for more details about the network structure.

[7] Please refer to the supplementary material for more details about the modified normal fitting method.

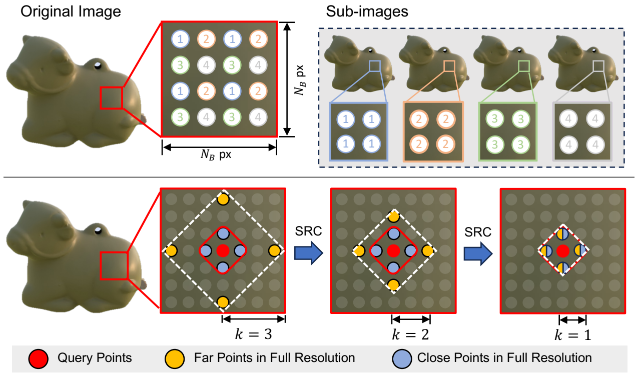

Figure 4. Proposed training strategies. (Top) **Interval sampling (IS)**. The high-resolution images are down-sampled into several low-resolution sub-images by extracting pixels with an interval of N_p, where $N_p = 2$ in this example. (Bottom) **Shrinking range computing (SRC)**. Far points (yellow circles) that are k ($k = 3$ in this example) point away *w.r.t* the query position is selected to interpolate the close points (blue circles)'s depth for normal calculation. During optimization, k will be gradually reduced to 1. Experiments show SRC facilitates convergence at the early stage.

the framework. The three-stage schema [16] is applied, as well as other smoothness terms (total variance regularization [13, 16]) on the normal map, diffuse albedo map, and the Gaussian bases' weights ($\boldsymbol{c}$) for the material. Similar to [30], the normalized color loss is implemented to help Spin-UP learn a better albedo representation. Following [16], we calculate the boundary loss for normal regularization[8].

3.4. Training Strategies

Optimizing Spin-UP in an unsupervised manner requires smoothness terms to facilitate convergence and avoid local optima. However, those terms are often implemented on full-resolution images, which requires extra time and computational cost. To reduce those costs, we propose to use a sampling strategy noted as *interval sampling* (IS). To further improve convergence, we introduce another technique noted as *shrinking range computing* (SRC).

Interval sampling (IS). IS samples ray batches from images to reduce the costs. Unlike random ray sampling [19] or patch-based sampling [20], IS preserves the object's shape. The idea of IS is similar to downsampled techniques in [11, 24], but we don't merge the sub-images to full resolution. We experimentally find this strategy important for training on down-sampled sub-images with smoothness terms to avoid local optima (Sec. 5.2). Specifically, we divide the image in full resolution into non-overlapping blocks, and each block contains $N_B \times N_B$ pixel points. By extracting pixel points from the same position in each block (*e.g.*, the left-top pixel), we obtain maximum $N_B \times N_B$ sub-images with a down-sampled resolution (See Fig. 4 for illustration). When the image resolution is not divisible by N_B, we will pad the image to ensure each sub-image has the same resolution. During training, those sub-images are randomly sampled in each step, and the smoothness terms are calculated based on the sub-images resolution, which ensures the effectiveness of those terms.

Shrinking range computing (SRC). Without merging sub-images to a full-resolution image in IS, there will be an aliasing issue in the inverse rendering process. Such an issue is caused by the fact that the normal calculation in our framework on sub-images requires four adjacent points' depths in sub-images resolution [16], which degrades the precision of normal calculation[9]. Therefore, Shrinking range computing (SRC) is applied for anti-aliasing. It uses points adjacent to the query point (blue circles in Fig. 4) **in the full-resolution image coordinates** to calculate the normal for each pixel in the sub-images. Such a strategy maintains the precision of normal calculation. However, at the early stage of training, calculating the normal based on the blue circles' depth is vulnerable to perturbation in per-pixel training. Therefore, SRC gradually selects points (yellow circles) from far ($k = 3$ points away) to close (blue circles) *w.r.t* query points to interpolate the blue circles' depth, as normal calculation on far points' depth will lead to a more smooth and stable normal map at the early stage, which eventually improves convergence. Experiments have shown this improves the accuracy of the estimated surface normal.

4. Experiments

We validate the effectiveness of the proposed Spin-UP on synthetic and real-world data. We use mean angle error (MAE) to evaluate the normal map's reconstructed quality and PU-PSNR [1], and PU-SSIM [1] to evaluate the reconstructed environment light. Since no existing datasets follow our spin light setup, we collect the dataset using Blender and our device. All experiments are conducted on an RTX A6000 GPU.

4.1. Evaluation on Synthetic Datasets

We collect several objects, environment maps, and materials to render the synthetic dataset in Blender by Cycles. Specifically, five shapes from the DiLiGenT-MV dataset [14] (*i.e*, BUDDHA, BEAR, COW, POT2, and READING) and a generated shape (BALL), five HDR environment maps (*i.e*, LANDSCAPE, QUARRY, URBAN, ATTIC, STUDIO), two PBR materials (*i.e*, RUSTY STEEL, LEATHER), and four synthetic materials (*i.e*, VORONOI DIFF, VORONOI SPEC,

[8] Please refer to the supplementary material for more details about the forms of the loss function and setup of hyperparameters.

[9] According to normal's definition, the smaller the distance between the adjacent points and the query points (red circle in Fig. 4), the more accurately representing the geometry at query points. Therefore, the blue circles' depths are preferred for normal calculation.

Table 2. Qualitative comparison results in terms of MAE on shape group (left-top), light group (left-bottom), reflectance group (right-top), and spatially varying material group (right-bottom). **Bold** numbers indicate the best results in MAE. In light group, {U., A., S., L.} stand for environment map named Urban, Attic, Studio, and Landscape, respectively. In the reflectance group, {D. , S.} stand for material named Green Diff and Green Spec, respectively. In spatially varying material group, {D. , S.} stand for material named Voronoi Diff and Voronoi Spec, respectively.

Method	Shape Group					Reflectance Group				
	Ball	Bear	Buddha	Reading	AVG	Pot2 (D.)	Pot2 (S.)	Reading (D.)	Reading (S.)	AVG
HY19 [7]	41.32	53.88	67.90	54.85	54.49	57.45	37.43	65.48	58.04	54.60
S22 [10]	7.35	14.03	26.37	18.77	16.63	15.56	11.83	18.97	18.38	16.19
S23 [11]	5.56	10.37	18.54	15.10	12.39	13.46	9.75	16.22	12.67	13.03
Spin-UP	**3.54**	**6.33**	**17.30**	**7.71**	**8.72**	**5.83**	**7.11**	**13.09**	**10.30**	**9.08**

Method	Light Group					Spatially Varying Material Group				
	Cow (U.)	Cow (A.)	Cow (S.)	Cow (L.)	AVG	Pot2 (D.)	Pot2 (S.)	Reading (D.)	Reading (S.)	AVG
HY19 [7]	67.63	39.21	40.28	48.47	48.89	40.97	37.46	49.27	48.96	44.17
S22 [10]	17.17	12.74	17.11	11.35	14.59	18.59	17.63	22.80	23.75	20.69
S23 [11]	11.93	7.52	12.38	11.60	10.84	14.22	11.00	14.58	14.31	13.53
Spin-UP	**5.50**	**4.40**	**3.33**	**4.94**	**4.54**	**5.58**	**6.97**	**12.54**	**11.52**	**9.15**

Green Diff, Green Spec)[10] are used for evaluation. We devise four groups of data for evaluation: shape group, light group, reflectance group, and spatially varying material group, each containing four scenes. For each scene, 50 observed images with a resolution of 512×512 are rendered by a perspective camera with the focal of 50mm and a frame size of $36\text{mm} \times 36\text{mm}$. The camera rotation θ for consecutive images follows a non-uniform rotation velocity. We compare Spin-UP with three advanced NaUPS methods, including two supervised NaUPS methods (S22 [10] and S23 [11]) and one unsupervised UPS method (HY19 [7])[11].

Normal estimation comparison. According to results in Table 2, Spin-UP presents a superior performance compared to all other NaUPS methods. Specifically, in **shape group**, the low MAE on Ball, Bear, Buddha, Reading in Rusty Steel rendered under Quarry indicate the practicability of Spin-UP to various shapes. In **light group**, the low-variance of MAE (1.15° for Spin-UP vs. 2.47° for S23 [11]) on Cow in Leather rendered under Quarry, Urban, Attic, and Studio demonstrates robustness toward different environment lights. In **reflectance group**, the results on Pot2 and Reading in Green Diff and Green Spec rendered under Landscape demonstrate the ability to handle non-Lambertian objects. In **spatially varying material group**, results on Pot2 and Reading rendered in Voronoi Diff or Voronoi Spec under Landscape prove adaptability to challenging scenarios. A comparative analysis of the outcomes of the reflectance group and the spatially varying material group in our method reveals that the MAE remains relatively consistent across identical objects with different materials, underscoring the robustness of Spin-UP in handling diverse materials. Also, we find that Spin-UP sometimes performs better on specular objects than on diffuse objects (*i.e*, Reading). We attribute this to the high-frequency details in specular reflectance that may be useful for shape-light reconstruction during training.

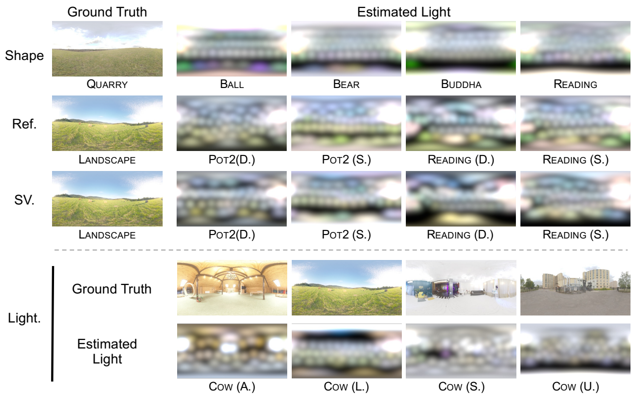

Figure 5. The visual quality comparison of light between the estimated one by Spin-UP (columns 2-4) and the ground truth (column 1) on the four groups, *i.e*, shape (row 1), reflectance (row 2), spatially varying material (row 3), and light group (row 4-5).

Light estimation comparison. Fig. 5 provides a qualitative comparison between the estimated environment light and the ground truth on four groups. We can observe that the learned light map reflects the position of the light source, especially in Cow (S.) given such a challenging setup without any prior information about the object material or shape. It also reconstructs a reasonable light map for objects with diffuse reflectance, such as Pot2 (D.) and Reading (D.), further highlighting the effectiveness of the proposed light initialization. However, it should be noted that the estimated

[10] The generated pattern for Voronoi Diff and Voronoi Spec follow similar setup in CNN-PS [9].

[11] Please refer to the supplementary material for all the qualitative comparison between Spin-UP and other methods on the synthetic and real-world dataset.

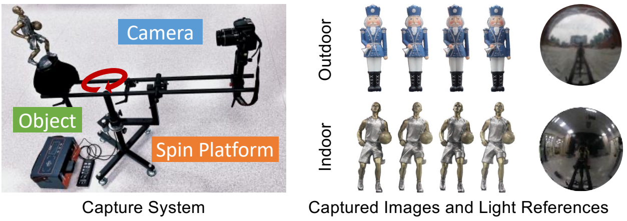

Figure 6. Illustration of the device for real data collection. Please refer to the supplementary material for more details. Left: the capture system contains a camera, a spin platform, and an object. Right: Captured images and paired mirror balls (as light reference) of two objects in the indoor and outdoor scenes, respectively.

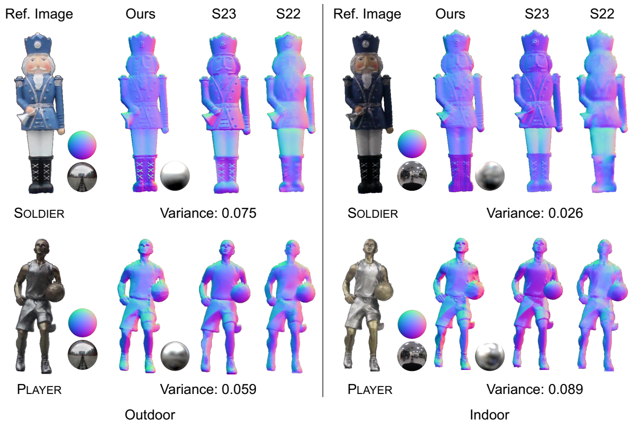

Figure 7. Qualitative comparison for the estimated normal map and environment light on SOLDIER (row 1) and PLAYER (row 2) capturing under outdoor (columns 1-6) and indoor (columns 7-10) between ours (column 3, 8), SS23 [11] (column 4, 9), and S22 [10] (column 5, 10). We also show the reference sphere and reference environment light in columns 2 and 7.

environment light is influenced by the object material, particularly when handling objects with a nearly uniform base color (*e.g.*, reflectance group) or spatially varying material, which may generate unpleasant artifacts (*i.e*, inconsistent color in estimated environment light of READING (D.) in the spatially varying material group.). Those artifacts are hard to eliminate without priors.

4.2. Evaluation on Real-world Datasets

We set up our spin light capture system to collect real data under indoor and outdoor scenes, as shown in Fig. 6. After preprocessing, we end up with 50 images for each object, with a 540×540 resolution. Here, we showcase the normal estimation results on two objects (*i.e*, SOLDIER, and PLAYER) under indoor and outdoor environment lights in Fig. 7, compared with S23 [11], S22 [10]. We do not show HY19 [7] as it failed on the captured data.

Normal estimation comparison. According to the Fig. 7, Spin-UP has competitive performance compared to the state-of-the-art supervised method [11]. In some scenarios, we recover more reasonable results regarding the overall distribution compared to the reference sphere, particularly for SOLDIER indoor and PLAYER indoor. Spin-UP can effectively capture high-frequency details such as wrinkles on clothes in PLAYER and SOLDIER. By contrast, S23 [11] may contain artifacts (*i.e*, an incorrect normal map distribution) even though they have more details than ours, and S22 [10] generates over-smooth results. By comparing indoor and outdoor results, we observe that the performance of S23 [11] degrades significantly in indoor scenarios, which may be attributed to the data bias and low pixel variance (Fig. 7), while our method is not greatly affected.

Table 3. Ablation studies on Spin-UP's alternatives regarding average PU-PSNR [1] and PU-SSIM [1] on synthetic dataset. 'Fib.' and 'Rand.' represent the Fibonacci and random initialization method, respectively. **Bold** number indicates the best results.

	w Rand.	*w* Fib.	Spin-UP
PU-PSNR [1]↑	16.86	18.98	**21.61**
PU-SSIM [1]↑	0.45	0.52	**0.62**

5. Ablation Study

5.1. Light Initialization Validation

To comprehensively validate the effectiveness of our light initialization method, we conduct experiments in two aspects: a comparison of different light initialization methods and the effectiveness of the filters.

Comparison on light initialization methods. We compare our light initialization method with two widely used SG light initialization methods, *i.e*, the random initialization noted as '*w* Rand.' and Fibonacci lattice [31] noted as '*w* Fib.'. A quantitative comparison of reconstructed normal and light maps is shown in Table 3 and Table 4. Compared with '*w* Rand.', the obvious improvement in average (1.81° reduction in MAE on normal estimation and 2.63/0.1 increase in PU-PSNR/PU-SSIM on light estimation, respectively) indicates our light initialization method's adaptability. Compared with '*w* Fib.', we observe an advantage in the shape group (0.61° reduction in MAE), while a smaller advantage in the challenging spatially varying material group (0.19° reduction in MAE). This is because the material and shape will affect the quality of the initial environment light. While the estimated environment light is most accurate on smooth geometry with simple material (*e.g.*, BALL), the quality will degrade on complicated geometry and spatially varying material.

Comparison on the designed filters. We compare Spin-UP with three alternatives (*i.e*, *w/o* $\mathcal{F}^c$, *w/o* $\mathcal{F}^d_{SH}$, and *w/o* $\mathcal{F}^d_{TH}$). The results in Table 4 demonstrate the effectiveness of those filters. Specifically, dropping $\mathcal{F}^d_{TH}$ will lead to mismatching light source position in the initialized environment light introduced by the specular reflectance, which eventually affects the accuracy of the estimated nor-

Table 4. Ablation studies on Spin-UP's alternatives regarding average MAE on four groups (shape, light, reflectance, and spatially varying material group). 'Fib.', 'Rand.', 'Intv.', and 'Shrk.' represent the Fibonacci initialization method, random initialization, interval sampling, and shrinking range computing, respectively. **Bold** number indicates the best results in MAE.

	Shape	Light	Ref.	SV.	AVG
S23 [11]†	12.96	12.29	13.93	16.07	13.81
w Rand.	12.02	6.44	9.79	10.28	9.60
w Fib.	9.33	5.34	10.20	9.34	8.55
w/o Intv.	8.87	5.02	9.83	10.12	8.37
w/o Shrk.	10.10	4.92	10.12	10.43	8.81
w/o $\mathcal{F}^c$	9.38	4.65	**8.77**	9.28	7.95
w/o $\mathcal{F}^d_{SH}$	9.75	8.29	8.93	15.73	10.41
w/o $\mathcal{F}^d_{TH}$	9.30	5.04	9.18	12.49	8.82
Spin-UP	**8.72**	**4.54**	9.08	**9.15**	**7.85**
Spin-UP†	11.62	9.25	11.07	9.07	9.48

† Method with † is tested on dataset with point light + environment light Sec. 5.3.

mal; dropping $\mathcal{F}^d_{SH}$ will harm the performance, especially in the spatially varying material group (6.56° increase in MAE) since $\mathcal{F}^d_{SH}$ is essential in extracting low-frequency reflectance to initialize the environment light; dropping $\mathcal{F}^c$ will increase MAE in average (0.25°), illustrating the necessity of reducing chromatic bias.

5.2. Training Strategies Validation

The interval sampling will facilitate the training of Spin-UP in two ways. First, the training time is two times shorter (25 min per object on average vs. 60 min on average, depending on the image's valid points for the object), and the GPU memory occupation is five times smaller (around 5 GB vs. 25 GB during training) than directly training on original resolution. Second, comparing Spin-UP with '*w/o* Intv.', which applies a random sampling strategy and calculates the smoothness terms on patches (3 × 3 pixels), we find that the performance drops $0.58°$ on average, and most ($0.95°$) on the spatially varying material group. This is because the patch-based smoothness may not work uniformly on different parts of the object, diminishing the effectiveness of smoothness terms, especially on objects in spatially varying material with abrupt texture changes. The shrinking range computing helps avoid local optima when training Spin-UP on down-sampled images while still using full-resolution pixel coordinates. We compared Spin-UP with the alternative '*w/o* Shrk.', which does not implement this strategy. The average MAE on normal estimation for four groups increases $1.02°$, highlighting the importance of this strategy in avoiding local optima.

5.3. Additional Validation on Point Light Source

To ensure a more fair comparison with the state-of-the-art supervised methods (S23 [11]), we add a dominant point light to the environment light in synthetic and real-world dataset[12]. According to Table 4, the proposed Spin-UP has a lower MAE on estimated normal maps than S23 [11] on the synthetic dataset ($9.48°$ ours v.s. $13.81°$ S23 [11]). As shown in Fig. 8, we have a visually comparable result on the real-world dataset given far-field point light (2m) and a better result given near-field point light (0.4m), validating the adaptability of Spin-UP on unseen light sources.

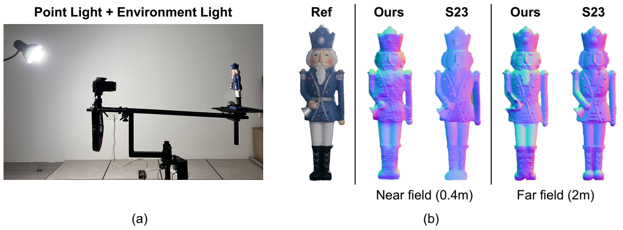

Figure 8. (a) Illustration of new light setup. (b) Qualitative comparison on SOLDIER between our method and S23 [11] with dominant point light.

6. Conclusion

This paper proposes Spin-UP to address NaUPS in an unsupervised manner. Thanks to our setup to mitigate the ill-posedness, the light initialization method to alleviate the ambiguities of NaUPS, and the proposed training strategies to facilitate convergence, Spin-UP can recover surfaces with isotropic reflectance under various lights. Experiments in synthetic and real-world datasets have shown that Spin-UP is robust to various shapes, lights, and reflectances.

Limitations and future work. Although Spin-UP is efficient and robust in solving NaUPS, it has several limitations: 1) Spin-UP assumes infinitely far light sources, which omit the spatially varying lighting; 2) the materials' base color will bias the estimated environment light; 3) Spin-UP assumes objects to have isotropic reflectance, ignoring inter-reflections and anisotropic features, meaning that it cannot perform well on objects with anisotropic reflectance, such as aluminum, or strong inter-reflections, such as a glass bowl; 4) Spin-UP doesn't compute the shadow iteratively, which may result in artifacts on objects with complicated shapes. Overcoming those limitations will be regarded as one of our future works. Also, we find it interesting to improve the setup by relieving the requirement for single-axis $360°$ rotation to free rotations for easier implementation on portable devices.

Acknowledgments. This work is supported by the National Natural Science Foundation of China under Grant No. 62136001, 62088102, Rapid-Rich Object Search (ROSE) Lab, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore, and the State Key Lab of Brain-Machine Intelligence, Zhejiang University, Hangzhou, China.

[12] Point light's setup follows [11] and [17].

References

[1] Maryam Azimi et al. PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR. In *Picture Coding Symposium (PCS)*, 2021. 5, 7

[2] Ronen Basri, David Jacobs, and Ira Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of Computer Vision (IJCV)*, 2007. 2

[3] 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. 2, 4

[4] Jean-Denis Durou, Maurizio Falcone, and Manuela Sagona. Numerical methods for shape-from-shading: A new survey with benchmarks. *Computer Vision and Image Understanding*, 2008. 2

[5] James Gardner, Bernhard Egger, and William Smith. Rotation-Equivariant conditional spherical neural fields for learning a natural illumination prior. In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3

[6] Heng Guo, Zhipeng Mo, Boxin Shi, Feng Lu, Sai-Kit Yeung, Ping Tan, and Yasuyuki Matsushita. Patch-based uncalibrated photometric stereo under natural illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 1, 2, 3

[7] Bjoern Haefner, Zhenzhang Ye, Maolin Gao, Tao Wu, Yvain Quéau, and Daniel Cremers. Variational uncalibrated photometric stereo under general lighting. In *Proc. International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 3, 6, 7

[8] Carlos Hernandez, George Vogiatzis, and Roberto Cipolla. Multiview photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2008. 2

[9] Satoshi Ikehata. CNN-PS: CNN-based photometric stereo for general non-convex surfaces. In *Proc. European Conference on Computer Vision (ECCV)*, 2018. 6

[10] Satoshi Ikehata. Universal photometric stereo network using global lighting contexts. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 6, 7

[11] Satoshi Ikehata. Scalable, detailed and mask-free universal photometric stereo. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 5, 6, 7, 8

[12] Micah K Johnson and Edward H Adelson. Shape estimation in natural illumination. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2011. 2

[13] Junxuan Li and Hongdong Li. Self-calibrating photometric stereo by neural inverse rendering. In *Proc. European Conference on Computer Vision (ECCV)*, 2022. 3, 4, 5

[14] Min Li, Zhenglong Zhou, Zhe Wu, Boxin Shi, Changyu Diao, and Ping Tan. Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials. *IEEE Transactions on Image Processing (TIP)*, 2020. 5

[15] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM TOG*, 2018. 2

[16] Zongrui Li, Qian Zheng, Boxin Shi, Gang Pan, and Xudong Jiang. DANI-Net: Uncalibrated photometric stereo by differentiable shadow handling, anisotropic reflectance modeling, and neural inverse rendering. *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 4, 5

[17] Chao Liu, Srinivasa G Narasimhan, and Artur W Dubrawski. Near-light photometric stereo using circularly placed point light sources. In *Proc. International Conference on Computational Photography (ICCP)*, 2018. 8

[18] Robert Maier, Kihwan Kim, Daniel Cremers, Jan Kautz, and Matthias Nießner. Intrinsic3D: High-quality 3D reconstruction by joint appearance and geometry optimization with spatially-varying lighting. In *Proc. International Conference on Computer Vision (ICCV)*, 2017. 2

[19] 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. 2, 4, 5

[20] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. RegNeRf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[21] Yvain Quéau, François Lauze, and Jean-Denis Durou. A L^1-TV algorithm for robust perspective photometric stereo with spatially-varying lightings. In *International Conference on Scale Space and Variational Methods in Computer Vision*, 2015. 1, 2, 3

[22] Robert T Seeley. Spherical harmonics. *The American Mathematical Monthly*, 1966. 4

[23] Boxin Shi, Ping Tan, Yasuyuki Matsushita, and Katsushi Ikeuchi. Bi-polynomial modeling of low-frequency reflectances. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2013. 4

[24] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[25] 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. 2

[26] 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. 2

[27] George Vogiatzis, Carlos Hernandez, and Roberto Cipolla. Reconstruction in the round using photometric normals and silhouettes. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2006. 2

[28] Jiaping Wang, Peiran Ren, Minmin Gong, John Snyder, and Baining Guo. All-frequency rendering of dynamic, spatially-varying reflectance. In *ACM SIGGRAPH Asia*, 2009. 4

[29] 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. 2

[30] Weicai Ye, Shuo Chen, Chong Bao, Hujun Bao, Marc Pollefeys, Zhaopeng Cui, and Guofeng Zhang. IntrinsicNeRF: Learning intrinsic neural radiance fields for editable novel view synthesis. *Proc. International Conference on Computer Vision (ICCV)*, 2023. 5

[31] 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. 2, 4, 7

[32] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1999. 2

[33] 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. 2

Spin-UP: Spin Light for Natural Light Uncalibrated Photometric Stereo
Supplementary Material

Zongrui Li[1,2,*] Zhan Lu[2,4,*,†] Haojie Yan[3,4] Boxin Shi[5,6] Gang Pan[3,4] Qian Zheng[3,4,‡] Xudong Jiang[1,2]

[1]Rapid-Rich Object Search (ROSE) Lab, Interdisciplinary Graduate Programme, Nanyang Technological University.
[2]School of Electrical and Electronic Engineering, Nanyang Technological University
[3]College of Computer Science and Technology, Zhejiang University
[4]The State Key Lab of Brain-Machine Intelligence, Zhejiang University
[5]National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[6]National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
{zongrui001,zhan007,EXDJiang}@ntu.edu.sg, {hjyan,qianzheng,gpan}@zju.edu.cn, shiboxin@pku.edu.cn

In this supplementary material,

1. we give more implementation details in Sec. 7, including details of framework structure (footnote 6), loss functions, and hyperparameters setup (footnote 8).
2. we introduce more about boundary normal calculation and normal calculation for rendering equation in perspective projection in Sec. 8 (footnote 7);
3. we provide an overview of the synthetic and real-world dataset in Sec. 9. We also explain how we collect and preprocess the real-world dataset;
4. we showcase a qualitative comparison between Spin-UP and other methods on the real-world dataset in Sec. 10 (footnote 11). More results from the real-world dataset are also included in this section (footnote 11);

7. Implementation Details

7.1. Network Structure

We use the similar multi-layer perceptrons (MLPs)' structures in [4, 5], shown in Fig. 9. The input of MLPs is pixels' 2D coordinate ($p = (u, v)$) in an image, which will pass through a positional encoding module similar in [6] calculated as

$$E(p) = \left(\sin\left(2^0\pi p\right), \cos\left(2^0\pi p\right), \cdots, \sin\left(2^{L-1}\pi p\right), \cos\left(2^{L-1}\pi p\right)\right), \tag{3}$$

where L is the positional code's dimension, which we set as 10.

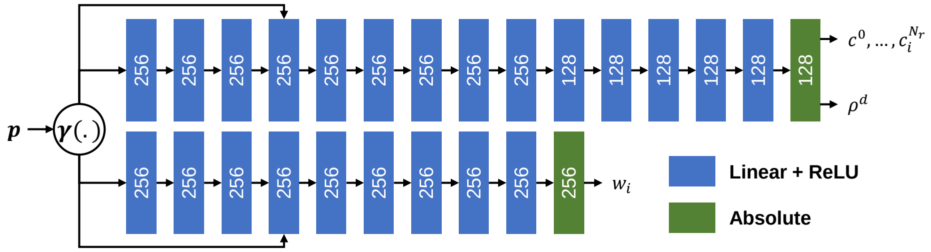

Figure 9. Network structure of MLPs for depth and material estimation in Spin-UP.

*Co-first author. ‡Corresponding author.
†Work completed while interning at the State Key Lab of Brain-Machine Intelligence, Zhejiang University.

7.2. Loss Functions and Hyperparameters Setup

In Spin-UP, we implement:

1. L1 inverse rendering loss L_r calculated as, $\sum_{i=1}^{N_P}\sum_{l=1}^{f}|\boldsymbol{m}_{il}-\hat{\boldsymbol{m}}_{il}|$, where, m_{il} and $\hat{m}_{il}$ is the ground truth and rendered pixel intensity, f is the number of images.
2. Normalized color loss L_{color}, calculated as, $\lambda_c\|\text{Nor}(\boldsymbol{A})-\text{Nor}(\boldsymbol{I})\|$, where $\lambda_c=0.5$, $\boldsymbol{A}$ is the albedo map, $\text{Nor}(.)$ is the vector normalization operator.
3. Boundary loss L_{b}, calculated as the cosine similarity between the pre-computed and estimated boundary normal.
4. Smoothness terms L_{sm} on albedo map $\boldsymbol{A}$, normal map $\boldsymbol{N}$, spatially varying Gaussian bases weights c^n, is calculated as,

$$L_{\text{sm}}=\frac{\lambda}{N_P}\sum_{i=1}^{N_P}\left|\frac{\partial\boldsymbol{A}}{\partial u}+\frac{\partial\boldsymbol{A}}{\partial v}\right|+\frac{\lambda_N}{N_P}\sum_{i=1}^{N_P}\left|\frac{\partial\boldsymbol{N}}{\partial u}+\frac{\partial\boldsymbol{N}}{\partial v}\right|+\frac{\lambda_S}{N_P}\sum_{n=1}^{N_S}\sum_{i=1}^{N_P}\left|\frac{\partial c_i^n}{\partial u}+\frac{\partial c_i^n}{\partial v}\right|, \tag{4}$$

where, $\lambda=0.01$, $\lambda_N=0.02$, $\lambda_S=0.01$.

We train the Spin-UP in three stages similar to [5]. For the first stage, the loss $\mathcal{L}_{\text{stage1}}$ is calculated as below for a faster convergence

$$\mathcal{L}_{\text{stage1}}=\mathcal{L}_r+\mathcal{L}_{\text{b}}+\lambda_c\mathcal{L}_{\text{color}}+\mathcal{L}_{\text{sm}}, \tag{5}$$

For the second stage, we drop the smoothness term on the albedo map and reduce λ_N to 0.05 for details refinement, where $\mathcal{L}_N$ is the smoothness term on normal map

$$\mathcal{L}_{\text{stage2}}=\mathcal{L}_r+\mathcal{L}_{\text{b}}+\lambda_c\mathcal{L}_{\text{color}}+\lambda_N\mathcal{L}_N, \tag{6}$$

For the third stage, we drop the smoothness terms $\mathcal{L}_N$ to further refine the details

$$\mathcal{L}_{\text{stage3}}=\mathcal{L}_r+\mathcal{L}_{\text{b}}+\lambda_c\mathcal{L}_{\text{color}}. \tag{7}$$

The three stages take 500, 1000, and 500 epochs, respectively. During training, we use Adam as the optimizer with a learning rate $\alpha_l=0.001$ and a batch size of 4 images per iteration.

8. Normal Calculation in Perspective View

8.1. Boundary Normal Calculation

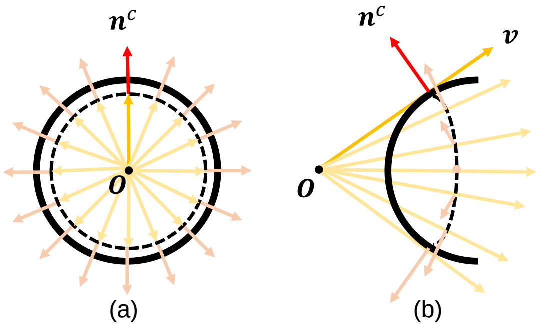

Figure 10. An illustration of occluding boundaries' normal relationship with view directions for (a) front view and (b) side view of a surface. The dotted line in (b) indicates the outermost boundaries of an object in perspective projection.

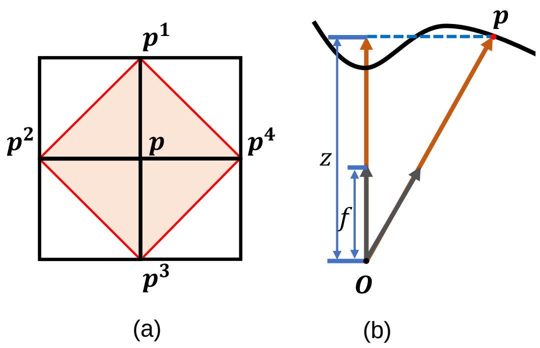

Figure 11. An illustration of (a) adjacent points' positions for normal fitting method [5] in perspective projection, (b) Eq. (11).

In perspective projection, the surface normal is perpendicular to the object's occluding boundaries $B(u,v)$ and view direction $\boldsymbol{v}$, as shown in Fig. 10. Therefore, the boundaries' normal $\boldsymbol{n}^b$ is calculated as

$$\boldsymbol{n}^b\cdot\boldsymbol{v}^b=0,\ \boldsymbol{n}^b\cdot(\frac{\partial B}{u},\frac{\partial B}{v},1)^\top=0, \tag{8}$$

Table 5. Length, width, height, and capturing distance for SOLDIER, PLAYER, DANCER, POLICEMAN and EEVEE.

Properties	SOLDIER	PLAYER	DANCER	POLICEMAN	EEVEE
Length (cm)	9.50	11.50	4.00	4.00	4.00
Width (cm)	7.00	11.00	5.00	4.00	4.00
Height (cm)	3.00	28.00	4.00	9.00	9.00
Distance (m)	0.90	0.90	0.40	0.40	0.30

In practice, the outer boundaries of an object in images may not precisely match its actual boundaries due to limited image resolution. Therefore, we add a small offset ($\beta = 0.1$) to make the pre-computed boundaries normal more accurate:

$$\boldsymbol{n}^b = \mathrm{Nor}(n^{bx}, n^{by}, n^{bz} + \beta). \tag{9}$$

8.2. Normal Calculation For Rendering Equation

The normal fitting method [5] in orthogonal projection is shown below:

$$\begin{aligned} \boldsymbol{n} &= \sum_{k=1}^{4} \gamma^k \boldsymbol{n}^k = \sum_{k=1}^{4} \gamma^k \,\mathrm{Nor}\left[\left(\boldsymbol{p}^{k+1} - \boldsymbol{p}\right) \times \left(\boldsymbol{p}^k - \boldsymbol{p}\right) \right]^\top, \\ \gamma^k &= \frac{\left|d^k\right|^{-1}}{\sum_{k=1}^{4} \left|d^k\right|^{-1}}, \quad d^k = w^k + w^{k+1} - 2w, \end{aligned} \tag{10}$$

where, $\boldsymbol{p}^k = (u^k, v^k, w^k)$ is the adjacent point of the query point $\boldsymbol{p} = (u, v, w)$, $k = 1$ if $k + 1 > 4$, as shown in Fig. 11, (a). To extend the normal fitting method to the perspective projection, we first compute the points' coordinates in the camera coordinate system by

$$\begin{aligned} \boldsymbol{p}^{k\prime} &= (u^k \frac{w^k}{f} s_x, v^k \frac{w^k}{f} s_y, w^k), \\ \boldsymbol{p}' &= (u \frac{w^k}{f} s_x, v \frac{w}{f} s_y, w). \end{aligned} \tag{11}$$

where f is the camera's focal, s_x and s_y are the width and height of the camera's frame. Replace $\boldsymbol{p}^k$ and $\boldsymbol{p}$ in Eq. (10) by $\boldsymbol{p}^{k\prime}$ and $\boldsymbol{p}'$, we get the normal fitting method in perspective projection.

9. Datasets

9.1. Synthetic Dataset

In Fig. 12, we showcase all 5 objects with 6 materials under 5 HDR environment maps rendered by Blender Cycles[1]. This results in 16 scenes[2] of synthetic data that are classified into 4 groups, *i.e*, the shape group, light group, reflectance group, and spatially varying group. We also show sample images with additional dominant point light in Fig. 12.

9.2. Real-world Dataset

The real-world dataset contains 5 objects captured under indoor and outdoor environments with spatially-varying materials. The five real-world objects used in our study are the SOLDIER, PLAYER, POLICEMAN, DANCER, and EEVEE. The objects' sizes are shown in Table 5.

Device introduction. SOLDIER, PLAYER, POLICEMAN, and DANCER's observed images were captured by a customized device shown in Fig. 13 (left), which consists of two stands (one for holding the subject being photographed, the other for supporting the camera) and a rotating mechanism. The distance from the camera to the object is adjustable. In addition to this, we also consider a more portable device shown in Fig. 13 (right), which is made up of a wooden rotatable platform[3] with a diameter of 39mm and the camera. We capture EEVEE's observed images based on this device.

[1] https://www.blender.org

[2] One scene representing an object with one material rendered under HDR environment maps.

[3] https://www.ikea.com/sg/en/p/snudda-lazy-susan-solid-wood-40176460/

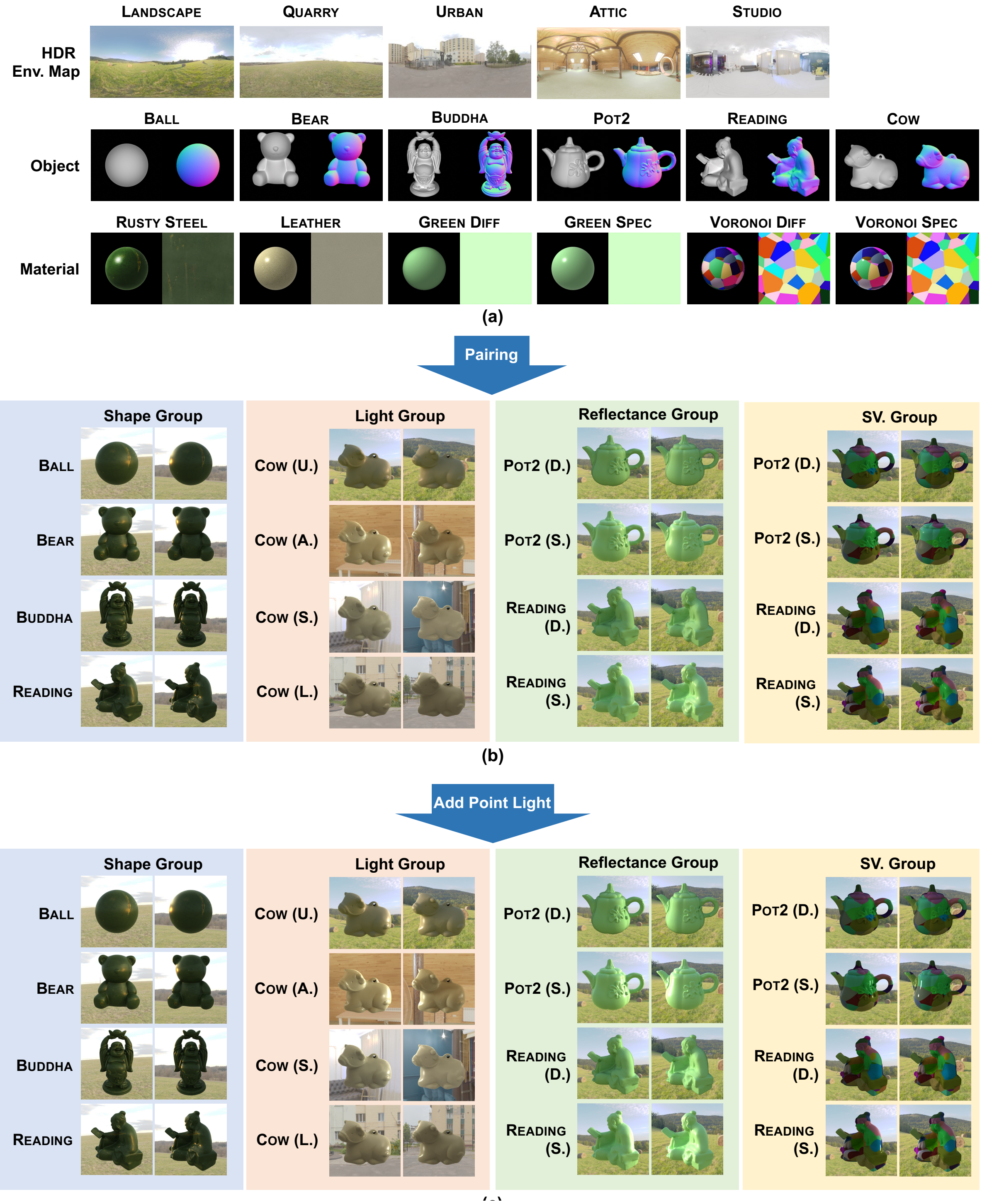

Figure 12. (a) HDR environment maps (row 1), objects (row 2), and materials (row 3) involved in the synthetic dataset. Each figure in row 2 consists of two subfigures for 3D model preview (left) and normal map (right). Each figure in row 3 consists of two subfigures for material rendered on a sphere (left) and albedo (right). (b) Example images from each scene in four groups. (c) Example images from the scene with additional dominant point source.

Photographing requirements. Before photographing, the distance between the camera and the object is determined based on the proportion of the object in the viewfinder, ensuring a balance of the occupied portion between the objects and the

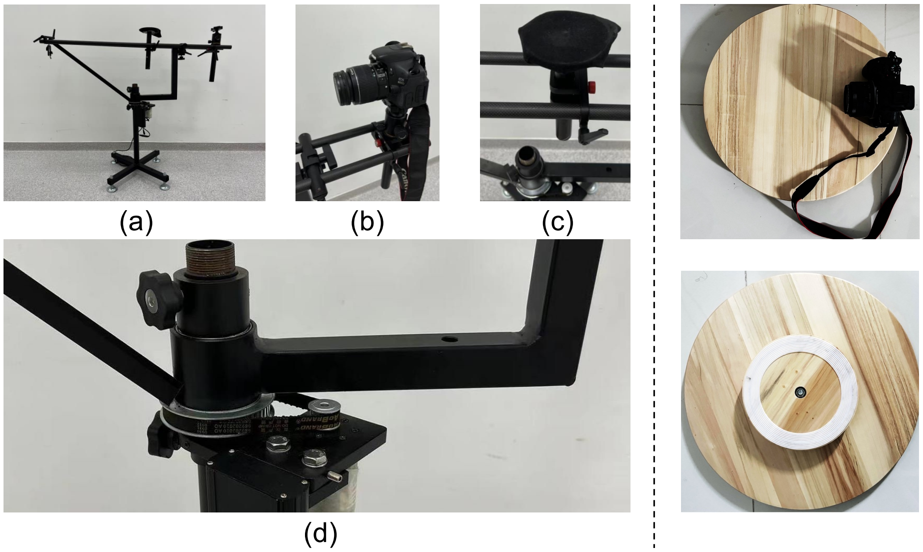

Figure 13. Left: (a) Overview of the device, (b) Stand for the camera, (c) Stand for the object being photographed with dark cloth for interreflection removal, (d) Rotating hinge. Right: A portable version of image capturing device, shown in top and bottom views.

camera. Three typical distances were used: 0.9 meters for large and 0.4 meters (or 0.3 meters) for small objects. During photographing, the thumb rule is to capture a clear image with less noise and keep rotation velocity as uniform as possible. For the camera's parameters, we chose ISO 1600, an aperture size of f/13 for outdoor scenes; and ISO 3200, an aperture size of f/6.3 for indoor scenes, respectively. The focal size is fixed at 31mm for different scenes.

Pre-processing pipeline. In the pre-processing pipeline, we extracted 50 images from the video at equal intervals to use as our data. We then obtain the objects' masks in each scene from the first frame by Photoshop. Those masks help separate objects and backgrounds. In practice, there are translational motions in the horizontal and vertical directions, mostly obvious on objects due to structural instability. Therefore, after calculating the relative rotation angle θ_j, we used a simple algorithm for motion correction, assuming that the only motion of the object relative to the camera was translational in the horizontal and vertical directions. Specifically, we pre-set the range of motion and iterate over the distance vector to find the distance of movement (plus or minus 20 pixels) that minimizes the difference between the front and back frames after applying the mask. Note that although large movement is corrected in this step, minor movements still exist and are hard to eliminate. Fortunately, our method can tolerate those minor movements.

10. Qualitative Comparison

10.1. Qualitative Comparison on Synthetic Dataset

We show all the estimated normal maps, error maps of Spin-UP, S23 [3], S22 [2], and HY19 [1] of shape, light, reflectance, and spatially-varying material groups in Fig. 15-Fig. 17.

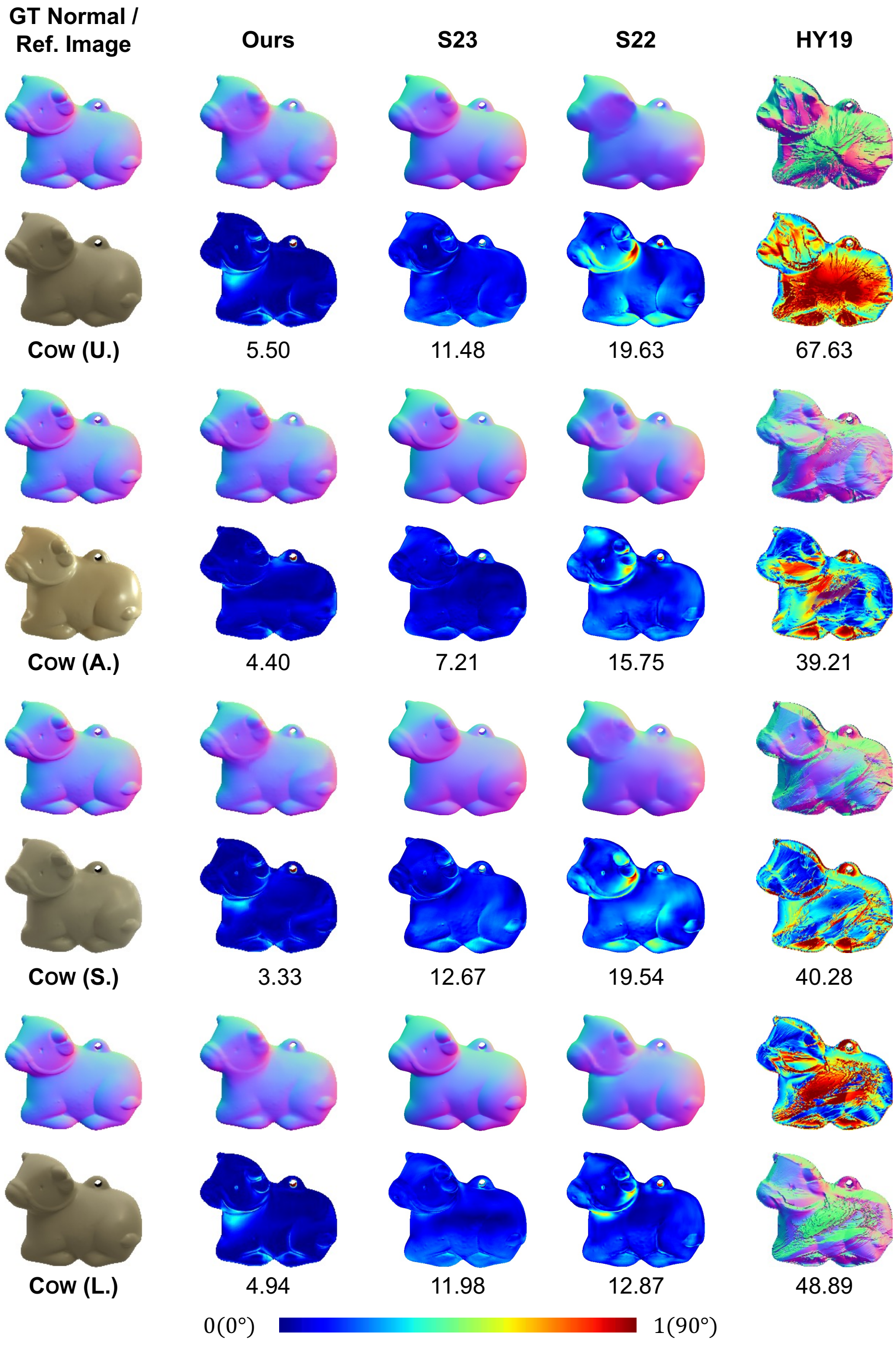

Figure 14. The visual quality comparison among Spin-UP, S23 [3], S22 [2], and HY19 [1] on the light group in terms of normal map (row 1, 3, 5, 7), error map (row 2, 4, 6, 8). Numbers indicate the MAE for surface normal.

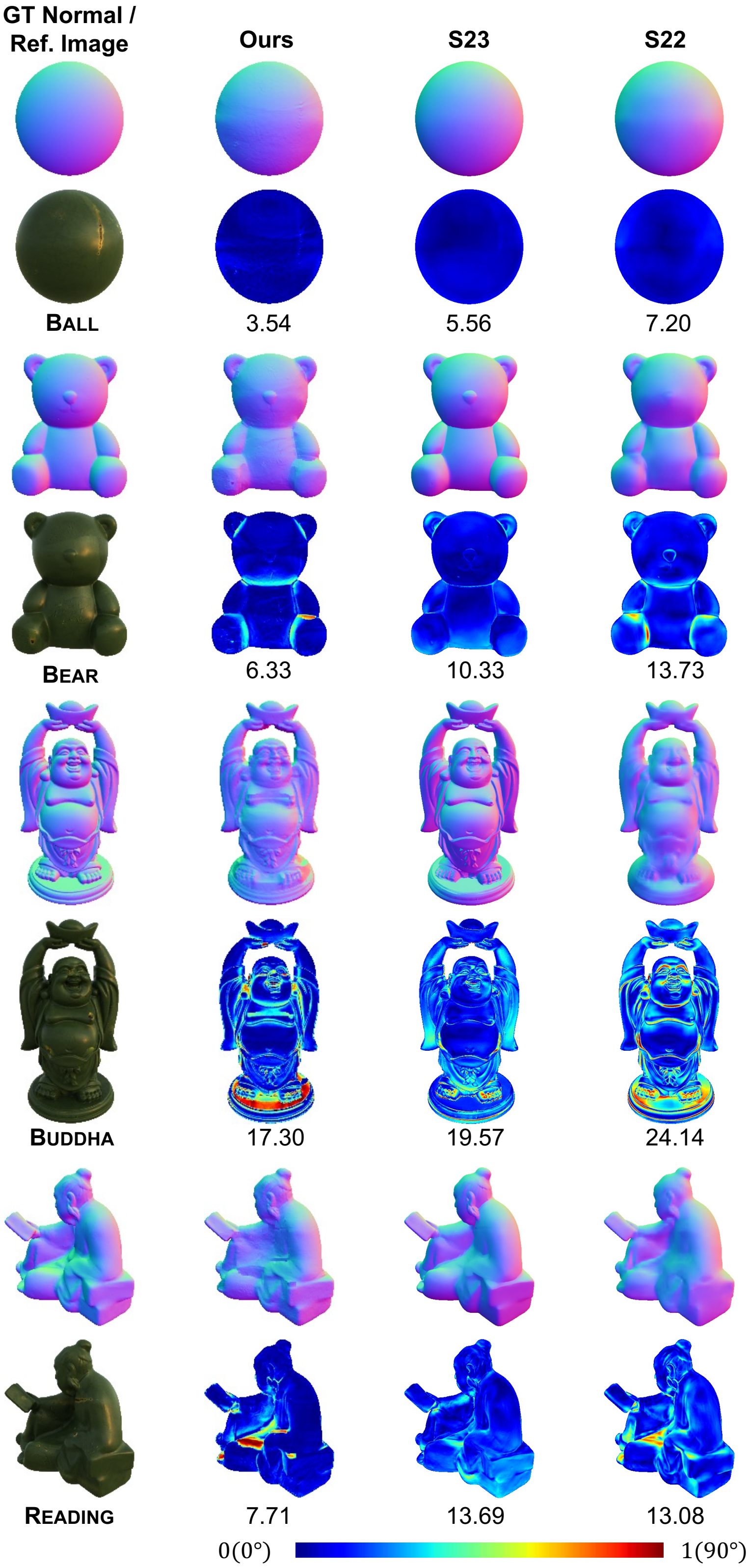

Figure 15. The visual quality comparison among Spin-UP, S23 [3], S22 [2], and HY19 [1] on the shape group in terms of normal map (row 1, 3, 5, 7), error map (row 2, 4, 6, 8). Numbers indicate the MAE for surface normal.

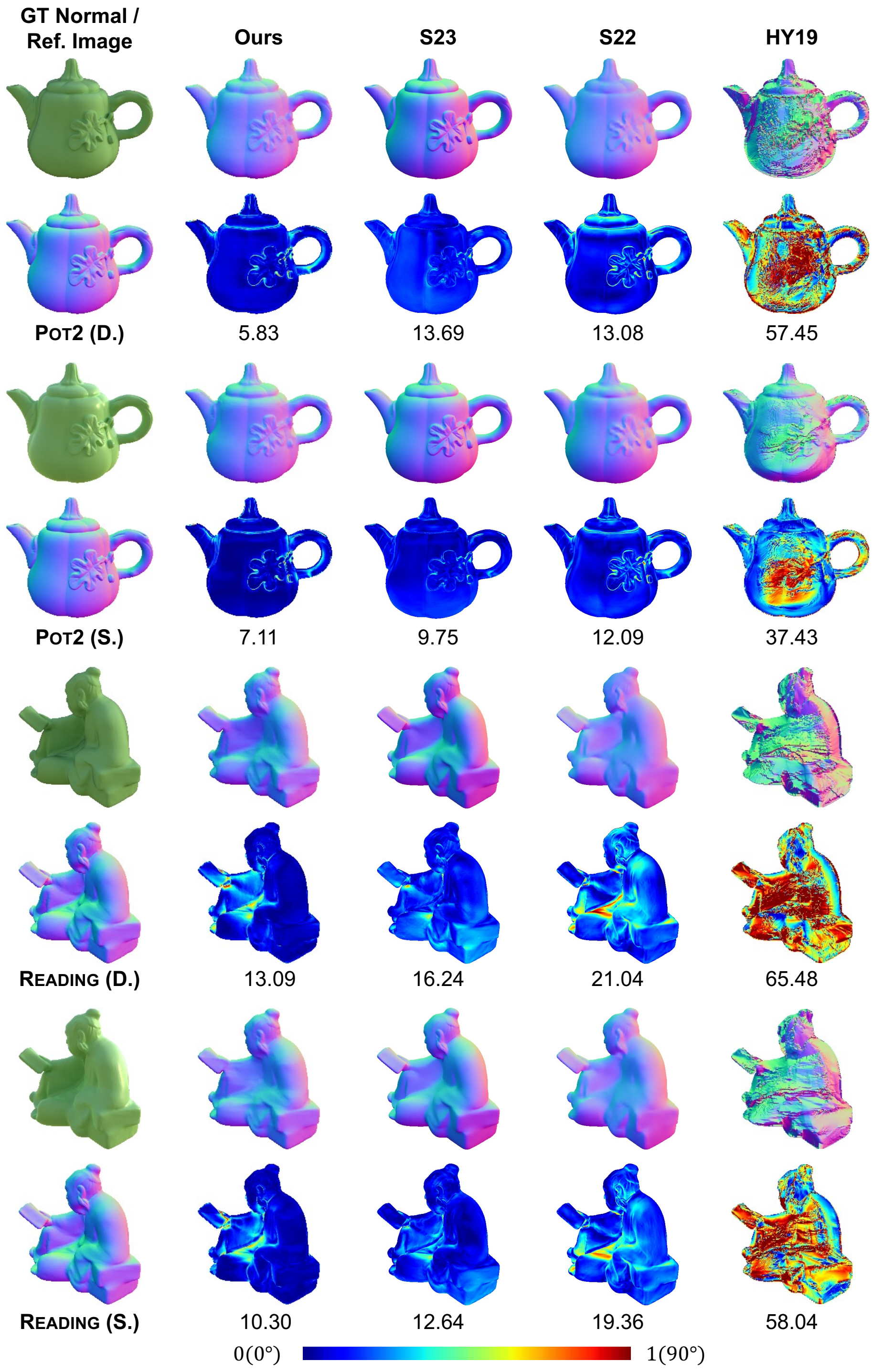

Figure 16. The visual quality comparison among Spin-UP, S23 [3], S22 [2], and HY19 [1] on the reflectance group in terms of normal map (row 1, 3, 5, 7), error map (row 2, 4, 6, 8). Numbers indicate the MAE for surface normal.

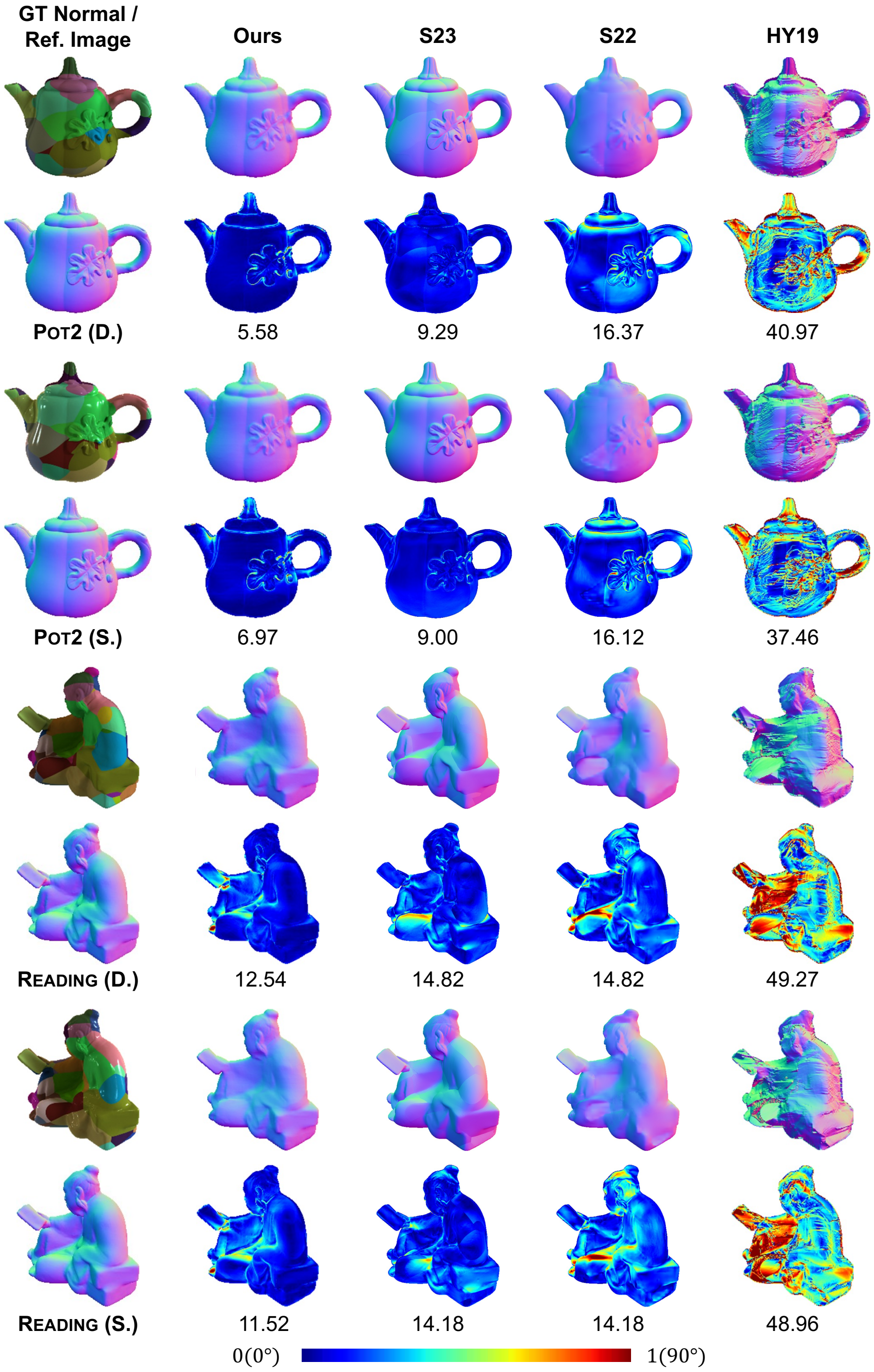

Figure 17. The visual quality comparison among Spin-UP, S23 [3], S22 [2], and HY19 [1] on the spatially varying material group in terms of normal map (rows 1, 3, 5, 7), error map (rows 2, 4, 6, 8). Numbers indicate the MAE for surface normal.

10.2. Qualitative Comparison on Real-world Dataset

We show all the estimated normal maps of Spin-UP, S23 [3], and S22 [2] of real-world dataset in Fig. 18 and Fig. 19.

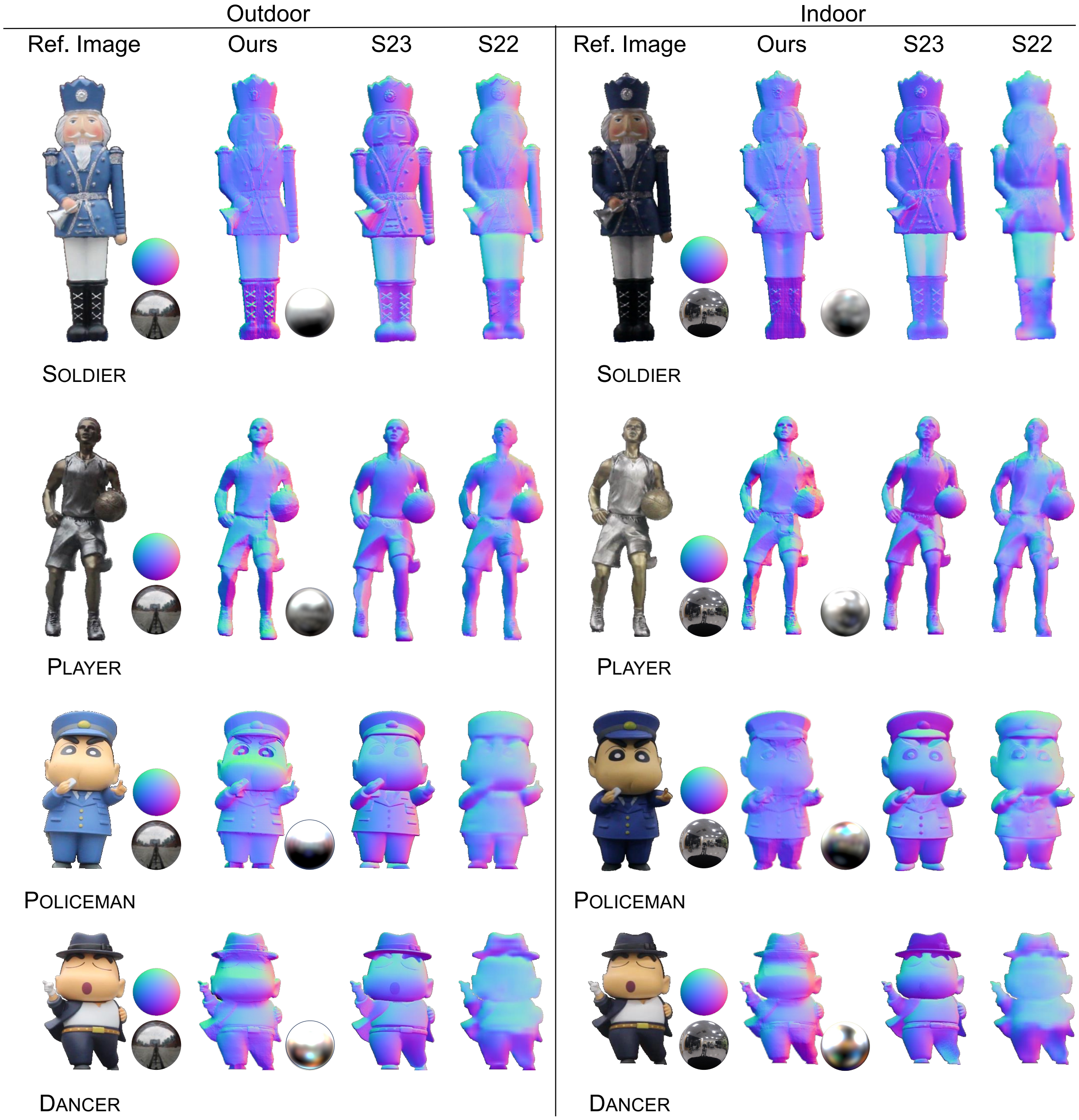

Figure 18. The visual quality comparison among Spin-UP, S23 [3], and S22 [2] on the SOLDIER, PLAYER, POLICEMAN, and DANCER in terms of the normal map. Left (right) side of the solid line: objects captured in CAMPUS (WORKPLACE) environment.

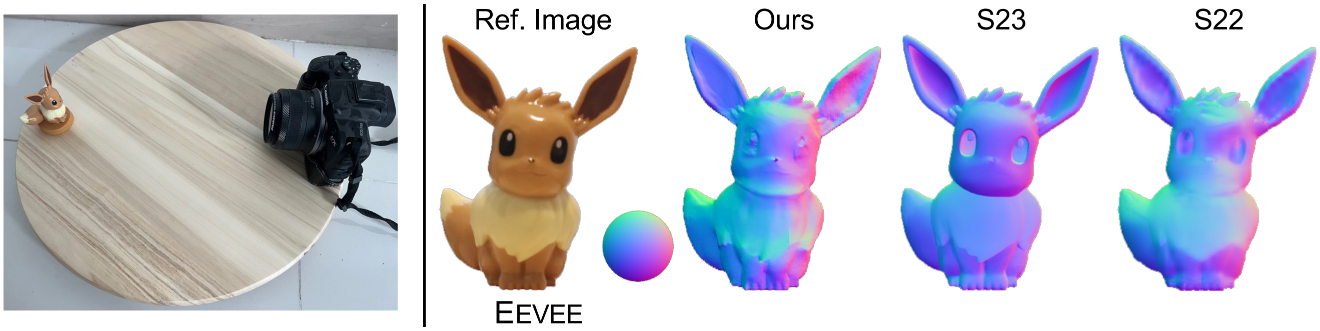

Figure 19. The visual quality comparison among Spin-UP, S23 [3], and S22 [2] on EEVEE captured in a living room in terms of the normal map based on more portable device.

References

[1] Bjoern Haefner, Zhenzhang Ye, Maolin Gao, Tao Wu, Yvain Quéau, and Daniel Cremers. Variational uncalibrated photometric stereo under general lighting. In *Proc. International Conference on Computer Vision (ICCV)*, 2019. 6, 7, 8, 9

[2] Satoshi Ikehata. Universal photometric stereo network using global lighting contexts. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2022. 6, 7, 8, 9, 10, 11

[3] Satoshi Ikehata. Scalable, detailed and mask-free universal photometric stereo. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2023. 6, 7, 8, 9, 10, 11

[4] Junxuan Li and Hongdong Li. Self-calibrating photometric stereo by neural inverse rendering. In *Proc. European Conference on Computer Vision (ECCV)*, 2022. 1

[5] Zongrui Li, Qian Zheng, Boxin Shi, Gang Pan, and Xudong Jiang. DANI-Net: Uncalibrated photometric stereo by differentiable shadow handling, anisotropic reflectance modeling, and neural inverse rendering. *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3

[6] 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. 1