Shape from Polarization with Distant Lighting Estimation

Youwei Lyu, Lingran Zhao, Si Li, and Boxin Shi*, *Senior Member, IEEE*

Abstract—This paper presents a new approach for surface normal recovery from polarization images under an unknown distant light. Polarization provides rich cues of object geometry and material, but it is also influenced by different lighting conditions. Different from previous Shape-from-Polarization (SfP) methods, which rely on handcrafted or data-driven priors, we analytically investigate the benefits of estimating distant lighting for resolving the ambiguity in normal estimation from SfP using the polarimetric Bidirectional Reflectance Distribution Function (pBRDF) based image formation model. We then propose a two-stage learning framework that first effectively exploits polarization and shading cues to estimate the reflectance and lighting information and then optimizes the initial normal as the geometric prior. Leveraging the normal prior with the polarization cues from the input images, our network further generates the surface normal with more details in the second stage. We also present a data generation pipeline derived from the pBRDF model enabling model training and create a real dataset for evaluation of SfP approaches. Extensive ablation studies show the effectiveness of our designed architecture, and our approach outperforms existing methods in quantitative and qualitative experiments on real data.

Index Terms—Shape from polarization, lighting, physics-based vision

1 Introduction

Changes in the polarization status during light propagation provide useful cues for estimating shapes and material information of objects. Shape-from-polarization (SfP) methods aim to recover the surface normal from single-view polarization images by utilizing the angle of polarization (AoP) and degree of polarization (DoP) of the incident light. Shape cues explored from polarization inherently contain pixel-wise geometric information, which could be much higher resolution than consumer-level 3D sensors, such as the Kinect [1]. However, they also introduce ambiguities in normal prediction. The polarizer can distinguish the oscillating orientation of light, but it produces the azimuth angle with π-ambiguity [2]. Moreover, the coexisting diffuse and specular reflections have divergent polarization properties, introducing additional $\pi/2$ deviation of the phase angle [3]. These two primary ambiguities in SfP make the problem under-determined, and extensive research has been conducted to deal with this challenging task.

To relieve the ill-posedness of SfP, researchers tend to assume that the reflectance is dominated by a single component or can be classified as the diffuse component and specular component, and they tackle each one separately. The dichromatic reflectance model [2], [4] and heuristic prior of intensities and DoP [5], [6] are adopted to tell the two types of polarization apart. By investigating only the diffuse dominant case, the difficulty of handling mixed polarization could be partly circumvented [7], [8]. Disambiguation is the next challenge after determining the type of polarization. Based on observations and properties of common objects, the convex prior and boundary constraints of the shape [2], [5], [6], [7], [8], smoothness priors [6], and shading clues [9] are popular choices to disambiguate the azimuth angle of normal. Formulating the polarization constraints as a linear system of equations, the least squares method [6] is presented to solve the ambiguity in a global manner. However, the results of these methods are easily affected by deviation from ideal conditions in real-world scenarios when handcrafted priors are poorly observed.

The lighting condition could significantly influence the polarization properties of the illuminated objects according to the microfacet theory [10]. Recent SfP works attempt to acquire quality appealing normal under specified lighting setups. The special polarization distribution of the sky, such as sunny weather, could provide extra constraints to facilitate the normal recovery [11]. Researchers find polarimetric diffuse reflectance dominates most of the regions under the frontal flash illumination [12] except for the directly lit regions, which simplifies the polarization imaging model. Thus, the frontal flash setup is used in pBRDF (polarimetric BRDF) acquisition and normal recovery [12], [13], [14]. When the frontal flash is altered to illuminate the object from a different direction, the position of specular reflection part changes, but the diffuse reflection is still dominant in the rest of the lit regions. Therefore, it is interesting to explore how a frontal flash setup or a single distant light will benefit the disambiguation of normal azimuth angles.

As deep learning achieves great success in vision tasks, the first deep learning-based SfP solution (DeepSfP) [15] has been introduced to address the limitation of handcrafted priors and robustness issue of optimization methods. Instead of solving the normal ambiguity explicitly, DeepSfP directly predicts the normal map by taking in polarization images

*Corresponding author.

- *Youwei Lyu and Si Li are with School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China. Email: {youweilv, lisi}@bupt.edu.cn.*
- *Lingran Zhao and Boxin Shi are with National Key Laboratory for Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing 100871, China. Email: {calvinzhao, shiboxin}@pku.edu.cn.*

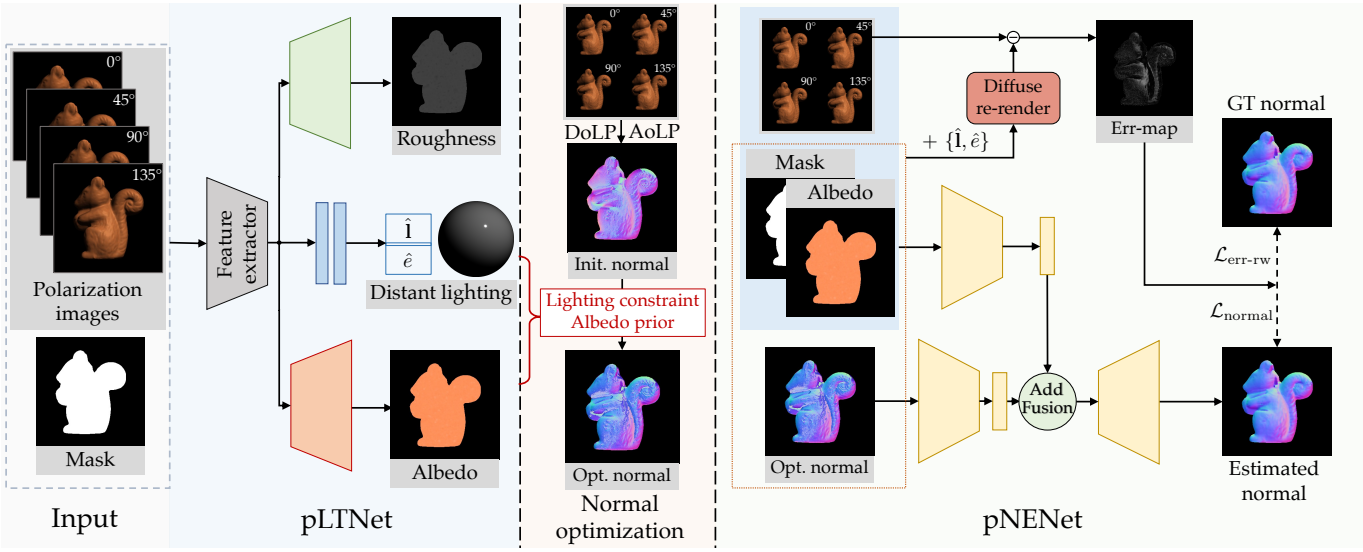


Fig. 1: The overview of our framework, which has a two-stage architecture consisting of pLTNet and pNENet. Taking in four polarization images and the mask as input, pLTNet estimates distant lighting and texture information; with the lighting and albedo prior, the optimized normal could be computed from the initial normal. Then we introduce pNENet as well as the diffuse-rendering error map to further refine the optimized normal and obtain the final results.

and ambiguous normals calculated from polarization. Despite taking priors of ambiguous normals into account, DeepSfP mainly relies on the data-driven prior for disambiguation without considering physics constraints from lighting and shading. A more recent work [12] additionally takes the Stokes maps and normalized color maps as a complement for estimating object normals and textures, which also resorts to the network for normal optimization instead of explicitly exploiting constraints for disambiguation. The remaining challenges in disambiguation of SfP normal estimation inspire us to take mutual benefits of physics and data-driven priors to complementarily narrow down the ambiguous solution space.

In this paper, we exploit polarization images to estimate the lighting information, textures, and surface normal of objects under an unknown single distant light. The adopted lighting setup has merits in two aspects: 1) We could compute the initial normal map from input polarization images with the approximation of dominant diffuse reflection; 2) as a spatially uniform vector regardless of pixel locations, such a lighting model makes it much easier to be estimated compared to the normal map. Analytically investigating the relationship between lighting and normal, we derive the formula that integrates distant lighting and surface albedo for normal disambiguation. With the explicit constraints, we could solve the π-ambiguity of azimuth angles and generate the optimized normal map, which prompts further refinement. Based on the above analysis, we design a two-stage learning framework to estimate the surface normal by combining the polarimetric cues and the shading constraint, which consists of two sub-networks, *i.e.*, polarization Lighting and Texture Network (pLTNet) and polarization Normal Estimation Network (pNENet), as shown in Fig. 1.

In the first stage, we present pLTNet to simultaneously predict the lighting parameters and object appearance (Fig. 1-pLTNet), which play a crucial role in the polarization imaging model. Then we design a non-learning scheme to optimize the ambiguous normal computed from polarization properties with the predicted lighting and albedo (Fig. 1-Normal optimization). In addition, the estimation error of pLTNet could affect the optimized results, so we carefully compare different lighting representations and select the best to boost the performance.

Despite the fact that the normal ambiguities are restricted in the normal optimization, the optimized results may be unsatisfactory for several reasons. First of all, we only take account of the diffuse reflection in the optimization step. However, the reflection can be a diffuse-specular mixture or dominated by the specular component (even if in a small region), which breaks our diffuse dominant assumption. On the other hand, part of normal is hardly constrained by the light direction due to the trivial solution of the optimization formula. To address this issue, we design a diffuse-rendering error to mark the pixels regarding specular reflection or inaccurate normals with large values and reweigh the training loss to enforce the network focusing on the local normal recovery of the specular regions. Thus, we further propose pNENet to overcome the limitation of the physical constraint and refine the normal map by integrating the polarization priors and shading information (Fig. 1-pNENet).

Overall, the main contributions of this paper are summarized as:

- We explicitly analyze distant lighting constraints for normal disambiguation in SfP under a clear observation of diffuse dominant phenomenon for the first time.
- We propose a two-stage deep neural network for joint shape and appearance acquisition from polarization, demonstrating the benefits of incorporating distant lighting estimation.
- We show the proposed method produces more accurate normal estimates, which quantitatively outperforms both optimization and learning-based approaches on the synthetic and real-world data.

The structure of this paper is organized as follows. In Sec. 2, we first introduce the related work of SfP, polarimetric BRDF, and distant lighting estimation. In Sec. 3, we then briefly revisit the basics of SfP and elaborate on our disambiguation method using the shading constraint. In Sec. 4 and Sec. 5, we introduce the proposed learning-based framework and the real dataset for evaluation, respectively. The experimental results, including ablation studies and comparison experiments, are presented in Sec. 6. At last, we discuss the limitations and conclude our paper in Sec. 7 and Sec. 8, respectively.

2 Related Work

2.1 Shape from Polarization

Shape-from-polarization methods tend to resort to the separation of specular and diffuse compositions and prior knowledge of surface shape for determining the surface normals. Handcrafted priors such as the convexity prior and boundary constraints [2], [7], [8] are popular choices to facilitate the disambiguation in earlier works. Researchers attempt to characterize the reflection as linear equations [6] or PDEs [16] and solve the normal via optimization. As deep learning flourishes in these years, learning-based methods are proposed to take in the normal priors from polarization and utilize CNN to produce more robust normal maps [12], [15]. The polarimetric information can be integrated with additional constraints to compensate for the limitation of relying on polarization cues only. Jointly formulating shape from shading and SfP in linear equations enables direct depth estimation of the surface [5], [6], [17], [18], [19]. The polarization properties could change drastically with regard to the surface roughness and lighting conditions, and SfP under specified illumination is explored for more compelling results. The SVBRDF (spatially varying BRDF) and shape acquisition approaches under a projector light [13], a polarized frontal light [20], or a frontal flash light [12], [14] have been proposed. The polarization patterns derived from the sky lighting model are investigated to include useful constraints on normal predictions [11]. Traditional SfP works focus on object shape recovery, and researchers attempt to handle the scene-level SfP by introducing a viewing encoding and a more effective polarization representation [21].

SfP could be solved by integrating additional constraints from various aspects. Multi-spectral measurements [22], introducing the material dispersion constraint, assist in predicting the refractive index and recover the shape simultaneously. In addition, coarse depth maps acquired by the RGBD camera [23], [24] could provide normal priors for SfP disambiguation. Stereo polarization cues have also been used for depth estimation [25], [26] and dense SLAM reconstruction [27]. In multi-view stereo, polarization methods help in enabling transparent surface modeling [28], solving normal vector for accurate correspondence [29], recovery of surface shape in featureless regions [30], [31], and normal estimate from specular reflection [32] and polarimetric cost volume [10]. Incorporated with Helmholtz stereopsis, polarization clues also benefit surface reconstruction with reciprocal image pairs [33]. Most of the SfP methods are proposed based on the orthographic camera model, *e.g.*, [6], [7], [12], while recent work [34] presents to determine plane normal by imposing the perspective projection constraint on the phase angle. Our method explicitly considers the image formation model in normal optimization, and the perspective projection can be applied in our framework given the intrinsic matrix of the camera. Some approaches propose to estimate epipolar geometry by phase information [35] or geometric information available from polarization cameras [36].

The pBRDF model describes the interaction of normal, reflectance, and lighting under the polarization context. The polarization properties of specular reflection are well explored, while the diffuse component is often modeled as unpolarized light to reduce the computational complexity [37], [38], [39].To improve the accuracy of pBRDF, Baek *et al*. [13] propose a diffuse-specular pBRDF by additionally modeling the polarization state changes in the light transmission process. Kondo *et al*. [20] further introduce the depolarization component into the BRDF model for depicting a wider range of real-world materials.

2.2 Distant Lighting Estimation

Distant lighting information could be inferred by inversely analyzing the photometric image formation model. Early methods assume known geometric and reflectance properties in the scene and estimate illumination from shading [40], [41], [42]. Estimation of distant lighting is necessary for solving uncalibrated photometric stereo, *e.g.*, SDPS-Net [43] determines both shape and distant lighting information of an object with unknown arbitrary reflectance using a lighting and normal estimation network, respectively. Thanh *et al*. [44] incorporate photometric stereo and polarization constraints to simultaneously estimate the surface normals and light directions.

Distant lighting estimation could be extended to deal with distant environment maps (usually represented as a panoramic HDR image) rather than a single point of distant light by relying on geometric priors. Given a single image and geometry of an object, Weber *et al*. [45] propose a framework to estimate the environmental map surrounding the object. Yi *et al*. [46] use faces as light probes and estimate the environment map via diffuse and specular components separation. Gardner *et al*. [47] propose to use discrete lights and the ambient term to represent the lighting conditions. More recent lighting estimation approaches directly predicting outdoor (*e.g.*, [48]) or indoor (*e.g.*, [49]) environment maps (without inferring shape) using deep learning are beyond the scope of this paper.

We adopt the setup of a single distant light and predict the lighting information from polarization images. We later use the estimated lighting and shading information to facilitate SfP normal disambiguation.

3 Polarimetric BRDF Model under Distant Lighting for SfP

In this section, we first review the basis of polarized light and SfP and then describe the relationship between the distant lighting and normal estimation based on polarimetric Bidirectional Reflectance Distribution Function (pBRDF),

TABLE 1: The lighting setup, camera model, and assumptions made in our method.

Lighting	A single distant light Unpolarized light source
Object	Dielectric material
Camera model	Orthographic/perspective
Assumption	No inter-reflections

which inspires us to disambiguate normals with the help of the direction of a distant light and object albedo. Our method is based on the specific lighting setup and several assumptions, which are summarized in Table 1.

3.1 Polarization of Light and Stokes Vector

As an electromagnetic wave, the oscillating orientation of the light is perpendicular to the propagation direction, and we conventionally use polarization to describe the oscillating status of a light ray. For acquisitions of the polarization images, a polarizer can be mounted in front of the camera under different polarizer angles ϑ, as shown in Fig. 2. We denote the intensity of polarization images as $I(\vartheta)$.

To describe the polarization formation model in detail, we adopt Stokes vectors to measure the polarization status of received light. A Stokes vector consists of four components: $\mathbf{s} = [\mathrm{s}(0), \mathrm{s}(1), \mathrm{s}(2), \mathrm{s}(3)]^\top$, $\mathrm{s}(0)$ is the total intensity of the light, $\mathrm{s}(1), \mathrm{s}(2)$ represent the intensity of linear polarization at $\vartheta = 0°, 45°$, respectively, and $\mathrm{s}(3)$ denotes the intensity of circular polarization [50]. Concentrating on the linear polarization properties for normal estimation, we use the first three entries of Stokes vectors and omit the circular polarization term $\mathrm{s}(3)$ like previous SfP methods [12], [15], [21]. Then the observed Stokes vector can be computed via four polarization images $I(0°), I(45°), I(90°)$ and $I(135°)$ [50],

$$\mathbf{s}_o = \begin{bmatrix} \left(I(0°) + I(45°) + I(90°) + I(135°)\right)/2 \\ I(0°) - I(90°) \\ I(45°) - I(135°) \end{bmatrix}, \quad (1)$$

where $I(0°), I(45°), I(90°)$ and $I(135°)$ are easily obtained by a DSLR camera attaching a rotating polarizer or by a quad-Bayer polarization camera [51].

Given the observed Stokes vector $\mathbf{s}_o$, we could also compute polarization images taken under different polarizer angles,

$$I(\vartheta) = \frac{1}{2}\begin{bmatrix} 1 & \cos 2\vartheta & \sin 2\vartheta \end{bmatrix}\mathbf{s}_o, \quad (2)$$

which enables the synthetic data creation.

3.2 SfP and Its Ambiguities

To make this paper self-contained, we briefly review the SfP method and its relevant ambiguity issues. Previous SfP methods [2], [6], [7], [9] attempt to predict the zenith angle θ_o and azimuth angle ϕ_o of the surface normal by assuming dominant diffuse or specular polarization. The degree of linear polarization (DoLP) ρ is related to the zenith angle θ_o and can be derived from the Stokes vector:$\rho = \sqrt{\mathrm{s}_o(1)^2 + \mathrm{s}_o(2)^2}/\mathrm{s}_o(0)$. When the diffuse component dominates in the reflection, the closed-form solution of θ_o could be computed by [6]:

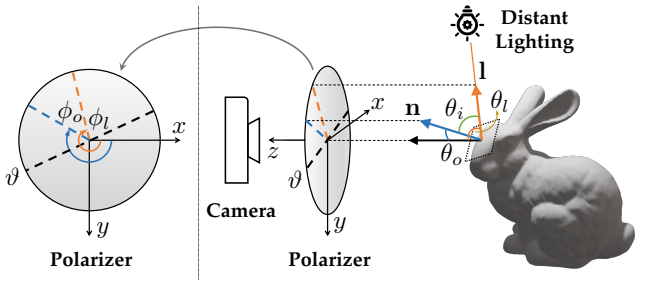


Fig. 2: A diagram for illustration of the symbols defined in our paper. The mesh model of Stanford bunny [52] is used as an example.

$$\cos\theta_o = \sqrt{\frac{\eta^4(1-\rho^2)+2\eta^2(2\rho^2+\rho-1)+\rho^2+2\rho-4\eta^3\rho\sqrt{1-\rho^2}+1}{(\rho+1)^2\left(\eta^4+1\right)+2\eta^2\left(3\rho^2+2\rho-1\right)}}, \quad (3)$$

where η denotes the refractive index. The azimuth angle is given by

$$\phi_o = \frac{1}{2}\arctan_2\frac{\mathrm{s}_o(2)}{\mathrm{s}_o(1)} \text{ or } \phi_o = \frac{1}{2}\arctan_2\frac{\mathrm{s}_o(2)}{\mathrm{s}_o(1)} + \pi. \quad (4)$$

With the calculated zenith and azimuth angle (with ambiguity), the normal vector can be obtained by

$$\mathbf{n} = \begin{bmatrix}\sin\theta_o\cos\phi_o & \sin\theta_o\sin\phi_o & \cos\theta_o\end{bmatrix}^\top. \quad (5)$$

Eq. (3), (4), and (5) are derived based on orthographic projection. Due to the complicated expression, we leave the derivation of the ambiguous normal under perspective projection in Sec. 2 of the supplementary material. To resolve the π-ambiguity in Eq. (4), previous work relies on handcrafted [2], [5], [6], [9] or data-driven priors [15], without explicitly considering the shading or lighting constraints.

3.3 Polarimetric BRDF and Lighting Constraint

The observed Stokes vector of exitant light can be expressed by matrix multiplication of a Mueller matrix $\mathbf{M}$ and an incident Stokes vector $\mathbf{s}_i$, and this formulation can be split into diffuse and specular components:

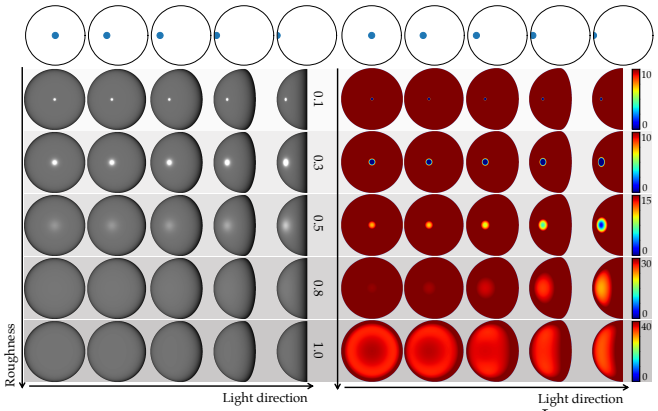
$$\mathbf{s}_o = \left(\mathbf{M}_d(\mathbf{l}) + \mathbf{M}_s(\mathbf{l})\right)\mathbf{s}_i, \quad (6)$$

where $\mathbf{l}$ is the direction of the incident light. $\mathbf{s}_i$ denotes the Stokes vector of the incident light from direction $\mathbf{l}$. The Mueller matrices $\mathbf{M}_{d,s}$ correspond to the diffuse and specular reflections, respectively, which inherently contain information about shape geometry, lighting conditions, and BRDF of the surface.

To reveal the relationship between surface normals and the light direction, we replace the Mueller matrix according to the diffuse-specular pBRDF model [13]:

$$\mathbf{s}_o = \left(\begin{bmatrix} T_o^+ \\ T_o^-\cos 2\psi_o \\ -T_o^-\sin 2\psi_o \end{bmatrix} a T_i^+ + \frac{D(\mathbf{l},\mathbf{n};\sigma)G(\theta_i,\theta_o;\sigma)}{4\cos\theta_o\cos\theta_i}\begin{bmatrix} R^+ \\ R^-\cos 2\varphi_o \\ -R^-\sin 2\varphi_o \end{bmatrix}\right)(\mathbf{n}\cdot\mathbf{l})e, \quad (7)$$

in which $T_{i,o}^\pm$ and $R^\pm$ are Fresnel terms, and $T_{i,o}^\pm$ with the subscripts $\{i,o\}$ are short for $T^\pm(\theta_o), T^\pm(\theta_i)$, respectively; a denotes diffuse albedo; $\mathbf{n}$ is the surface normal vector;



(a) Sphere images (b) $10 \log_e \frac{I_{\text{diff}}}{I_{\text{spec}}}$

Fig. 3: We render the images of spheres with ranging roughness parameters under different lights in (a), based on the adopted pBRDF [13]. The diffuse-specular ratios in (b) show that the diffuse component dominates most of the regions when the object surface is smooth and the intensity of values the diffuse is 10 times greater than those of the specular when the surface is very rough (roughness>0.8).

θ_i denotes the angle between the surface normal and the incident light direction; ψ_o is the rotation angle regarding azimuth of surface normal, and φ_o is the rotation angle about azimuth of the incident light; D describes the normal distribution function (NDF) of microgeometry [53] and G is the geometry function [54]; σ denotes the specular roughness and e is the light intensity[1]. The definitions of the symbols are illustrated in Fig. 2.

Precisely modeling real-world environment illumination requires an HDR panoramic image [55], which is far too complicated for inverse problems. On the other hand, the polarization properties of the same object could change dramatically under varying illumination conditions [10], [12], which is one of the key factors that makes SfP problems intractable. For shape recovery or spatially-varying BRDF acquisition, researchers tend to restrict the lighting setup, such as sunny/cloudy sky [11] or a frontal projector/flash light [12], [13], [14]. We further study SfP by extending the frontal light to a distant light from an unknown and arbitrary direction and propose to estimate the lighting information for facilitating normal map prediction. The distant lighting setup also has been widely adopted in photometric stereo [56] and data capture in SfP (*e.g.*, SONY dataset [44]).

Under the distant lighting setup, diffuse polarization will dominate a majority of the regions, and the specular reflection is distributed on a small patch in the directly illuminated regions [12]. As a visualization, we plot the log ratio of the diffuse to specular components on a sphere in Fig. 3 with regard to the distant light directions varying along the horizontal axis and the roughness parameter in Eq. (7) ranging from 0.1 to 1 in the vertical axis.

Under the diffuse dominant assumption, the observed Stokes vector in Eq. (7) can be approximated by omitting the specular term, which is given by

$$\mathbf{s}_{\text{diff}} = a \begin{bmatrix} T_o^+ \\ T_o^- \cos 2\psi_o \\ -T_o^- \sin 2\psi_o \end{bmatrix} \cos\theta_i T_i^+ e. \tag{8}$$

Then the intensity of received light is approximated by

$$\mathbf{s}_{\text{diff}}(0) = a T_o^+ \cos\theta_i T_i^+ e. \tag{9}$$

In Eq. (9), the ambiguous azimuth angle of surface normal ϕ_o is inherently contained in $\cos\theta_i$:

$$\cos\theta_i(\phi_o) = \cos\theta_l \cos\theta_o + \sin\theta_l \sin\theta_o \cos(\phi_o - \phi_l), \tag{10}$$

where θ_l and ϕ_l represent the zenith angle and the azimuth angle of a distant light, respectively. To relieve the π-ambiguity of the normal azimuth angle, we introduce the shading related term s denoted as

$$s = \cos\theta_i T_i^+, \tag{11}$$

where T_i^+ is a function regarding $\cos\theta_i$, and the magnitude curves of T_i^+ and s are shown in the supplement.

Investigating Eq. (9), the zenith angle of the normal θ_o can be computed by Eq. (3), and the azimuth angle ϕ_o is estimated by Eq. (4) with π-ambiguity. We find the lighting terms contain much fewer unknowns compared to spatially varying normal vectors, and albedo is close to the observed image intensities, which are expected to be much easier to estimate. Assuming the lighting conditions and object albedo could be predicted, the shading term s provides reliable constraints on removing the azimuth ambiguity of the surface normal by optimizing

$$\hat{\phi}_o = \underset{\phi_o}{\operatorname{argmin}} \left| s\left(\cos\theta_i(\phi_o)\right) - \frac{\mathbf{s}_o(0)}{a e T_o^+} \right|. \tag{12}$$

This objective function regarding normal azimuth is easy to solve since Eq. (4) has restricted the solution space to two feasible points. The disambiguation process is illustrated in Fig. 4. Note that the disambiguation method may fail in the trivial solutions according to Eq. (10): When $\sin\theta_o = 0$, the zenith angle of normal equals zero, and there is no ambiguity; when $\cos(\phi_o - \phi_l) = 0$, the direction of the azimuth angle of normal is orthogonal to that of the distant light, and lighting provides no azimuth cues of normal (Fig. 4(c)); $\sin\theta_l = 0$ is the special case of $\cos(\phi_o - \phi_l) = 0$, where the view direction and the light direction are collinear (this may happen in real capture). Except for these degenerated cases, Eq. (12) facilitates the normal disambiguation for the object pixels.

Therefore, we propose to predict the lighting conditions and albedo map from the polarimetric images first, which will assist in distinguishing the ambiguity in azimuth angles of normal from SfP. By applying diffuse dominant assumption, the specular component is omitted for deriving the objective function for normal estimation. However, specular reflections still exist in real data, which affects the accuracy of the optimized normal map; then, we design a neural network to generate ambiguity-free normal maps by incorporating the lighting constraint and polarization information. Note that Eq. (12) is obtained based on the orthographic camera projection. We also derive the objective function under perspective projection (a slight variant of Eq. (12)) and further analyze the influence of different camera models on the performance of our method in Sec. 2 of the supplement.

1. Please refer to the supplementary material for details about the pBRDF model and formulation of the Fresnel terms.

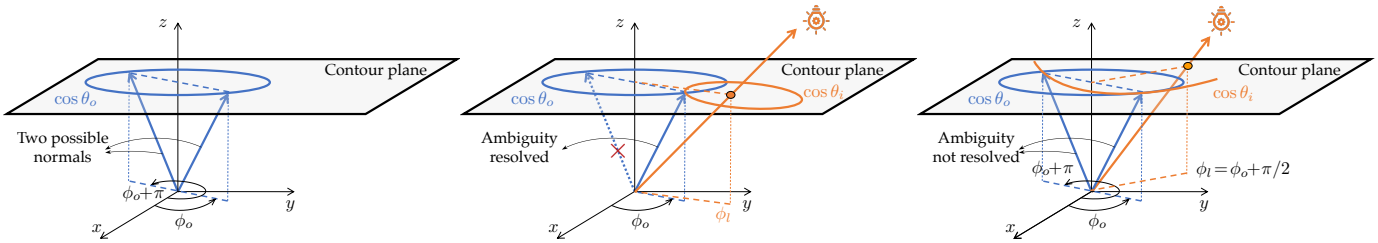


Fig. 4: Illustration of the disambiguation process. (a) Two possible normals computed from DoLP and AoLP. (b) Normal ambiguity resolved under the constraint of the light direction. (c) The light direction fails to constrain the optimization when $\cos(\phi_o - \phi_l) = 0$.

4 Learning to Predict Lighting, Texture and Normal from Polarization

In this section, we introduce our two-stage framework for distant lighting, albedo and roughness estimation, and surface normal recovery, which takes advantage of SfP under a single distant light.

4.1 Network Architecture

As shown in Fig. 1, our network takes a two-stage architecture for surface normal reconstruction, consisting of polarization Lighting and Texture Network (pLTNet) and polarization Normal Estimation Network (pNENet).

4.1.1 Polarization Lighting and Texture Network (pLTNet)

pLTNet is targeted at predicting the lighting (L), the albedo map (A) and the roughness map (R) by taking four polarization images (I_{pol}) and the corresponding mask (M) as input. The network structure is composed of independent decoder branches for the three sub-tasks, and the three branches share the same encoder since the lighting information and texture are coupled in object appearance. The architecture of pLTNet is illustrated in Fig. 1-pLTNet.

The lighting branch is adopted to estimate the distant light direction and intensities in RGB. The light direction can be either represented as a unit vector in the Cartesian coordinate or denoted by a pair of zenith and azimuth angles in the spherical coordinates. However, predicting the zenith and azimuth angle separately may be inefficient: The azimuth angle contributes less to the accuracy of the light direction as the zenith angle approaches zero. We find it more effective to regress the light direction vector in the Cartesian coordinate rather than estimating the direction angles in the spherical space, which is attributed to the representation continuity of solution space [57]. The light direction is assumed within the hemisphere of the view direction, so the light direction vector satisfies: $|\mathbf{l}| = 1, \mathbf{l} = [l_x, l_y, l_z]^\top, l_z > 0$. For the light intensity, we empirically cast the estimation as a classification problem with C_{int} categories.

The architecture of the shared feature extractor consists of five residual blocks [58], which down-sample the input tensor in the height-width dimension and dilate the planes of encoded features to 1024. In the lighting estimate branch, the encoder is followed by two fully connected layers to produce the light direction vector and softmax probabilities of the light intensity. We then convert the output to a unit vector of the light direction and RGB color of the light by taking the middle value of the category with the highest probability.

Learning the object albedo enables extracting the shading term for normal disambiguation, and estimating specular roughness helps with the acquisition of complete SVBRDF parameters. In line with the encoder architecture, we utilize five convolutional blocks with bilinear interpolation for upsampling for each branch. Also, skip connections between the encoder and decoder are adopted to preserve more details from the extracted features. We denote the pLTNet as

$$\hat{\mathbf{l}}, \hat{e}, \hat{a}, \hat{\sigma} = \mathcal{F}_{\text{pLTNet}}(I_{\text{pol}}, M). \tag{13}$$

Loss function The cosine loss is widely adopted in the direction estimation,

$$\mathcal{L}_{\cos}(\mathbf{z}, \hat{\mathbf{z}}) = 1 - \mathbf{z}^\top \hat{\mathbf{z}}, \tag{14}$$

where $\mathbf{z}$ and $\hat{\mathbf{z}}$ denote the reference vector and the predicted vector, respectively. We employ this function to supervise distant light direction estimates. For light intensity supervision, we use the multi-class cross-entropy loss denoted as $\mathcal{L}_{\text{CE}}$. We supervise the albedo and roughness branches by measuring $\mathcal{L}_1$ and $\mathcal{L}_2$ distance between the predicted map with the corresponding ground truth. In summary, the total loss function of pLTNet is given by

$$\begin{aligned}\mathcal{L}_{\text{pLTNet}} = \lambda_{\text{dir}}\mathcal{L}_{\cos}(\mathbf{l}, \hat{\mathbf{l}}) + \lambda_{\text{int}}\mathcal{L}_{\text{CE}}(e, \hat{e}) + \lambda_{a1}\mathcal{L}_1(a, \hat{a}) \\ + \lambda_{a2}\mathcal{L}_2(a, \hat{a}) + \lambda_{\sigma1}\mathcal{L}_1(\sigma, \hat{\sigma}) + \lambda_{\sigma2}\mathcal{L}_2(\sigma, \hat{\sigma}),\end{aligned} \tag{15}$$

in which $\{\lambda_{\text{dir}}, \lambda_{\text{int}}, \lambda_{a1}, \lambda_{a2}, \lambda_{\sigma1}, \lambda_{\sigma2}\}$ are weighting parameters to balance each loss.

4.1.2 Normal Optimization

The initial normal with ambiguity ($\mathbf{n}_{\text{init}}$) can be calculated via Eq. (3) and (4). We could derive the shading-related term s from Eq. (11). Given the light intensities and object albedo estimated from the preceding stage, the ambiguous azimuth angle is constrained by the distant light direction (Eq. (10)). It is easy to solve azimuth angles by evaluating only two alternative values in Eq. (12) for each pixel, and we could then derive the optimized normal ($\mathbf{n}_{\text{w/lt}}$) to facilitate the final estimate.

4.1.3 Polarization Normal Estimate Network (pNENet)

To tolerate errors in lighting estimation and mitigate trivial solutions in Eq. (12), we build pNENet for integrating the physical constraint from polarization and shading clues from captured images, as shown in Fig. 1-pNENet. The

optimized normal also suffers from artifacts in the directly illuminated regions since the observed light is approximated as the diffuse reflection in the normal optimization step. We propose to compute the rendering error map for loss function reweighing at the training stage, which further benefits the normal detail recovery.

We employ a widely adopted encoder-decoder architecture to refine the normal map, which consists of ResNet [58] blocks with the SPADE [59] normalization layers. pNENet concatenates the input polarization images, the object albedo, and the corresponding mask as one input stream and feeds it into the encoding layers for extracting high-level shape features from appearance information. The optimized normal map is simultaneously taken as another stream to utilize the polarimetric constraint. Each encoder consists of five convolutional layers for extracting high-level features of polarization and normal information. For integrating features from two different sources, we fuse the latent features from two streams through an addition operation. The following decoder, fed with the fused features, has five upsampling layers with skip connections from the same hierarchical level encoding blocks. The final output is normalized to unit vectors. We denote the pNENet as

$$\hat{\mathbf{n}} = \mathcal{F}_{\text{pNENet}}(\mathbf{n}_{\text{w/lt}}, I_{\text{pol}}, M, \hat{a}). \tag{16}$$

There still remains a global convex/concave ambiguity in lighting estimate and normal recovery [6]. We adopt the convexity prior to objects in this paper, and one could retrain the model to apply it to concave objects.

Loss function pNENet exploits the shape prior knowledge to generate final results by taking in the optimized normal. However, the optimized normal is computed under the dominant diffuse reflection, which in the specular/specular-diffuse-mixed regions can be inaccurate. We further propose the diffuse-rendering error, which measures the discrepancy between the observed image intensities and the re-rendered diffuse ones:

$$\mathbf{E} = \left| \mathbf{s}_o(0) - \hat{a} T_o^+ \cos\theta_i T_i^+ \hat{e} \right|, \tag{17}$$

of which large values correspond to the specular regions (green bounding box) and inter-reflections in concave regions (red bounding box), as shown in Fig. 6. We re-render the diffuse components, *i.e.*, $\hat{a} T_o^+ \cos\theta_i T_i^+ \hat{e}$, instead of the total intensity, because the optimized normal only correlates with the diffuse reflection. At the training stage, we reweigh the cosine loss with the error map penalizing more on the normal without the reliable prior, which is expressed as

$$\mathcal{L}_{\text{err-rw}} = \frac{1}{\sum_k^N E_k} \sum_k^N E_k \mathcal{L}_{\cos}(\mathbf{n}_k, \hat{\mathbf{n}}_k), \tag{18}$$

where N is the number of valid pixels in the image, $\mathbf{n}_k$ and $\hat{\mathbf{n}}_k$ denote the reference normal vector and the estimated normal vector at pixel k, respectively, and E_k is the value of the error map at pixel k. Besides, the regular cosine loss is also adopted to guarantee the overall quality of the normal map:

$$\mathcal{L}_{\text{normal}} = \frac{1}{N} \sum_k^N \mathcal{L}_{\cos}(\mathbf{n}_k, \hat{\mathbf{n}}_k), \tag{19}$$

and the total loss function of pNENet is given by

$$\mathcal{L}_{\text{pNENet}} = \mathcal{L}_{\text{err-rw}} + \mathcal{L}_{\text{normal}}. \tag{20}$$

4.2 Implementation Details

We implemented our model with the PyTorch framework [60], used Adam solver [61] with default parameters, and set the weighting parameters of pLTNet loss $\{\lambda_{\text{dir}}, \lambda_{\text{int}}, \lambda_{a1}, \lambda_{a2}, \lambda_{\sigma 1}, \lambda_{\sigma 2}\} = \{1, 0.01, 1, 1, 1, 1\}$. We experimentally set $C_{\text{int}} = 40$ for categories of light intensity of each RGB channel. In the normal optimization stage, the refractive index was fixed to 1.5. We discuss the selection of the refractive index in Sec. 6.1. We first trained the pLTNet with the initial learning rate at 8×10^{-4} for 65 epochs and multiplied the learning rate by 0.5 every 20 epochs, and then successively trained pNENet with the lighting and albedo estimated by pLTNet, allowing pNENet to learn to better tolerate error in the first stage. The base learning rate was set to 1.6×10^{-4} and halved at the 20th epoch, and the model was trained for 30 epochs. To test our method on the grayscale dataset like SONY [44], we expand the channel of grayscale images to 3 and take the expanded tensor as input in our network.

5 SfP Dataset under Distant Lighting

5.1 Real Data Acquisition

Though DeepSfP [15] provides polarization images of various objects and their corresponding ground truth normal maps, the data are collected under indoor/natural outdoor environment, which is different from ours. Deschaintre *et al*. [12] create a test set consisting of RGB polarization images of 12 objects without ground truth normal, and these data only can be used for qualitative comparison. SONY dataset [44] is proposed for polarization-photometric shape recovery, which contains 8 sets of polarization images under 8 different distant lights corresponding to each object. In addition, the ground truth of the light directions and object normal is provided in the SONY set [44], and we could adopt it for quantitative evaluation of our lighting and normal estimation model. However, polarization images in SONY dataset [44] is captured in grayscale. To complement the existing real-world SfP dataset, we collect a set of RGB polarization images under the distant lighting setup and provide the ground truth normal maps. The data acquisition setup is shown in Fig. 5. We adopt a Lucid Vision Triton polarization camera (with Sony IMX250MYR CMOS and a 16mm lens) to take four polarization images under polarizer angles of 0°, 45°, 90°, and 135° at a single shot. For acquisition of ground truth normal, we first scan the objects with the Shining 3D EinScan–SP scanner[2] to obtain the point cloud. Based on structured light technology, this scanner produces 3D point cloud with a point spacing about 0.2mm and automatically generates the complete mesh. After getting the mesh model of the objects, we follow the method proposed in [62] to conduct the shape-to-image alignment and then render the "ground truth" normal with Mitsuba2 [63]. We employ an LED flashlight with a lens in the front to radiate a near-parallel beam of light towards the

2. https://www.einscan.com/einscan-sp/einscan-sp-specs/

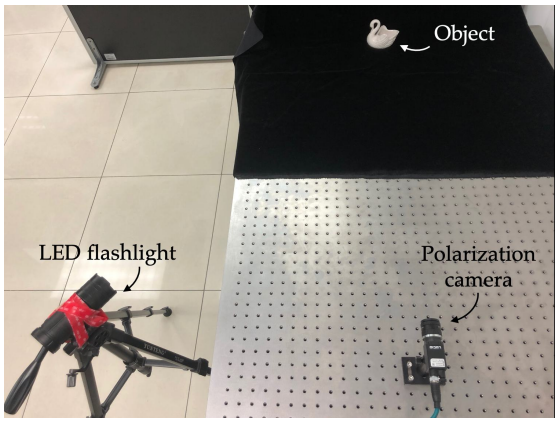


Fig. 5: Illustration of our data acquisition setup. For better visualization, the setup image is taken with ceiling lights on, while real data are captured only with an LED flashlight on.

object, which was placed 120cm away to mimic a distant light like [56]. This dataset consists of 40 RGB polarization images of 4 objects with a resolution of 1024×1224, and we take images of each object under 10 different light directions. It supports quantitative evaluation of existing SfP methods under distant lighting.

5.2 Synthetic Data Generation

Due to the fact that there is no large-scale SfP dataset collected under a distant light, we resort to the pBRDF model [13] for training and validation data synthesis. We use 43 object shapes provided by [12], [15], [44] and generate the normal maps by random rotation. To simulate different distant lights, we randomly sample light directions within the hemisphere of the view direction. We collect publicly available SVBRDFs from [64], which contains 1064 high-quality SVBRDFs covering a large range of materials. We further augment the BRDF maps by randomly changing RGB in the HSV color space, rotating, scaling, and finally cropping to the desired size. We randomly choose the value of the refractive index from $\mathcal{U}(1.4, 1.6)$ to mimic typical dielectric materials. With the parameters above prepared, we acquire the output Stokes vector via Eq. (6) and generate a rendering for polarization images by Eq. (2). We further augment the synthetic images by adding Gaussian noise to simulate the capturing process in a realistic manner. We simulate polarization images with 512×512 resolution for training and validation.

In the training stage, we render a total of 43,520 sets of images by 25 object shapes from the training set of [15] and 862 materials from [64] and randomly crop them into 256×256 patches for data augmentation. To avoid obtaining blank patches, we pre-compute the indices of feasible patches, which consist of over 25% valid pixels for each set of data. During training, we randomly select a feasible index to crop the original images for each set of data. Our test set consists of 7,800 sets of data simulated with the remaining 18 shapes (8 from the test set of [15], 8 from [44], and 2 from [12]) and 202 materials from [64]. Examples of rendered polarization images and corresponding DoLP and angle of linear polarization (AoLP) maps are provided in the supplementary material.

TABLE 2: Ablation studies of the lighting branch of pLTNet. We conduct the experiment on SONY [44] and our synthetic dataset. The estimated light directions (Dir.) and intensities (Int.) are evaluated by MAngE and MSE, respectively.

Model	Dir. on SONY ↓	Dir. on Sync. ↓	Int. Sync. ↓
pLTNet-Light	**6.226**	**6.559**	**.09105**
pLTNet-Light$_{\text{angle-cls}}$	7.429	6.881	.10261
pLTNet-Light$_{\text{angle-reg}}$	13.70	14.70	.09983
pLTNet-Light$_{\text{int-reg}}$	7.017	6.760	.09544
pLTNet-Light$_{\text{w/ unpol}}$	16.09	15.85	.09340

TABLE 3: Ablation studies of pNENet. Quantitative evaluation is conducted on SONY [44], the synthetic dataset, and our real-world dataset.

		Angular error ↓			Accuracy ↑		
		Mean	Median	ARMSE	11.25°	22.5°	30.0°
SONY [44]	$\mathbf{n}_{\text{init}}$	41.98	29.66	57.02	.3266	.4500	.5052
	$\mathbf{n}_{\text{w/lt}}$ (ours, η=1.5)	25.56	12.63	38.42	.4979	.6600	.7139
	$\mathbf{n}_{\text{w/lt}}$ (η=1.4)	29.37	15.17	42.73	.4368	.6306	.6852
	$\mathbf{n}_{\text{w/lt}}$ (η=1.6)	25.06	12.83	37.52	.4947	.6672	.7222
	pNENet	**8.868**	**5.682**	**13.32**	**.7853**	**.9181**	**.9505**
	pNENet$_{\text{w/o } \mathbf{n}_{\text{w/lt}}}$	10.46	6.904	15.20	.7398	.8922	.9330
	pNENet$_{\text{w/ } \mathbf{n}_{\text{init}}}$	10.28	6.375	15.44	.7505	.8921	.9305
	pNENet$_{\text{sgl branch}}$	9.661	6.123	14.50	.7626	.9018	.9404
	pNENet$_{\text{concat fuse}}$	9.294	5.832	13.97	.7697	.9086	.9450
	pNENet$_{\text{w/o err-rw loss}}$	9.638	6.294	14.21	.7664	.9071	.9433
	pNENet$_{\text{w/o albedo}}$	9.550	6.014	14.39	.7670	.9040	.9407
	pNENet$_{\text{w/o mask}}$	9.218	5.947	13.72	.7724	.9104	.9459
Sync.	$\mathbf{n}_{\text{init}}$	47.58	37.03	61.25	.2433	.3694	.4370
	$\mathbf{n}_{\text{w/lt}}$ (ours, η=1.5)	31.76	21.85	44.30	.3729	.5383	.6162
	$\mathbf{n}_{\text{w/lt}}$ (η=1.4)	34.24	24.26	46.49	.3272	.5093	.5887
	$\mathbf{n}_{\text{w/lt}}$ (η=1.6)	31.15	20.91	43.63	.3780	.5533	.6317
	pNENet	**4.924**	**3.981**	**6.573**	**.9294**	**.9873**	**.9940**
	pNENet$_{\text{w/o } \mathbf{n}_{\text{w/lt}}}$	5.946	4.896	7.740	.8950	.9814	.9915
	pNENet$_{\text{w/ } \mathbf{n}_{\text{init}}}$	5.812	4.599	7.896	.8979	.9801	.9902
	pNENet$_{\text{sgl branch}}$	5.396	4.395	7.146	.9142	.9843	.9927
	pNENet$_{\text{concat fuse}}$	5.017	4.055	6.680	.9267	.9871	.9938
	pNENet$_{\text{w/o err-rw loss}}$	5.275	4.302	6.931	.9175	.9858	.9934
	pNENet$_{\text{w/o albedo}}$	5.336	4.327	7.043	.9150	.9854	.9933
	pNENet$_{\text{w/o mask}}$	5.147	4.187	6.823	.9219	.9864	.9936
Our Real	$\mathbf{n}_{\text{init}}$	53.28	44.56	66.19	.1331	.3003	.3806
	$\mathbf{n}_{\text{w/lt}}$ (ours, η=1.5)	41.81	31.58	53.15	.1466	.3773	.4969
	$\mathbf{n}_{\text{w/lt}}$ (η=1.4)	44.03	33.51	55.45	.1200	.3373	.4638
	$\mathbf{n}_{\text{w/lt}}$ (η=1.6)	40.74	30.54	51.96	.1552	.3995	.5147
	pNENet	**15.58**	**12.65**	**19.64**	**.4387**	**.8275**	**.9111**
	pNENet$_{\text{w/o } \mathbf{n}_{\text{w/lt}}}$	16.68	13.38	21.19	.4199	.8037	.8900
	pNENet$_{\text{w/ } \mathbf{n}_{\text{init}}}$	16.57	13.29	20.91	.4035	.7968	.8899
	pNENet$_{\text{sgl branch}}$	16.22	13.16	20.75	.4099	.8252	.9078
	pNENet$_{\text{concat fuse}}$	15.81	12.90	19.76	.4264	.8217	.9106
	pNENet$_{\text{w/o err-rw loss}}$	16.17	13.19	20.21	.4063	.8149	.9044
	pNENet$_{\text{w/o albedo}}$	16.20	13.21	20.37	.4087	.8119	.9029
	pNENet$_{\text{w/o mask}}$	16.08	13.11	20.23	.4137	.8193	.9033

6 Experimental Results

First, we conduct the ablation study on the simulated/real dataset to show the effectiveness of the proposed framework. Then we quantitatively and qualitatively evaluate our approach using both synthetic and real-world data by comparing it with state-of-the-art methods. We adopt mean/median angular error, angular root-mean-square error (ARMSE) (↓ denotes lower is better), and angular accuracy percentage (↑ denotes higher is better) to quantitatively evaluate the quality of estimated surface normal maps. For distant light directions, mean angular error (MAngE) is utilized for evaluation. Mean squared error (MSE) and mean absolute error (MAE) are adopted to measure the predicted albedo and roughness.

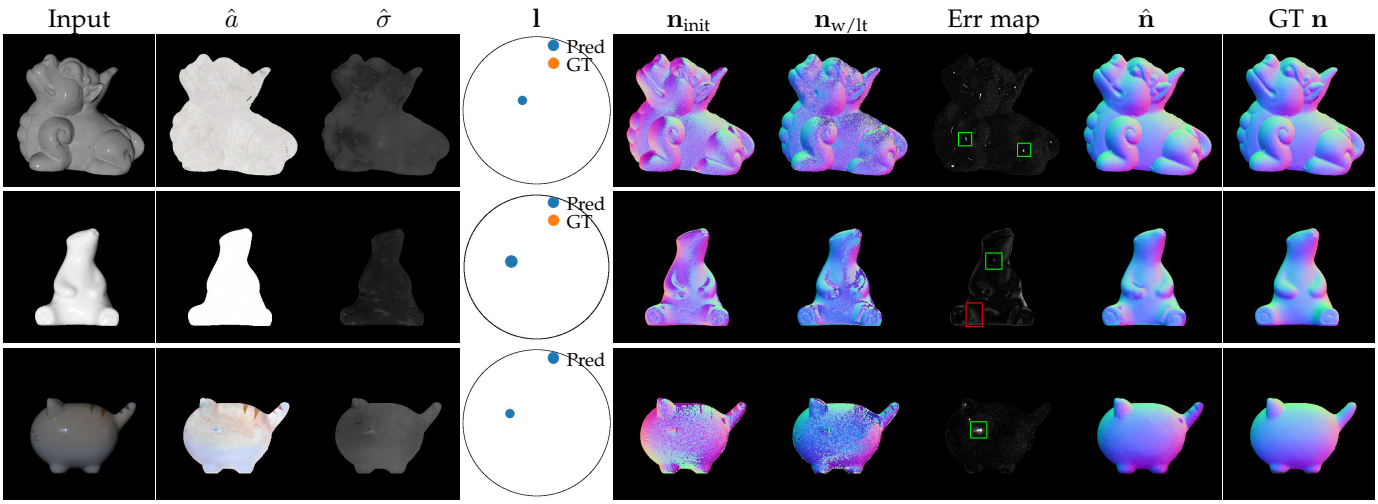


Fig. 6: Qualitative evaluation of our method on synthetic data, SONY [44], and our real data. We show the texture maps ($\hat{a}, \hat{\sigma}$) and light directions predicted by pLTNet, the initial normal ($\mathbf{n}_{\text{init}}$) obtained from Eq. (3) and Eq. (4), the normal optimized ($\mathbf{n}_{\text{w/lt}}$) under the distant lighting constraint, the diffuse-rendering error map (rescaled for better visualization), and the normal predicted by pNENet ($\hat{\mathbf{n}}$). In the error map, part of the specular and inter-reflection regions is marked with the green and red bounding boxes, respectively.

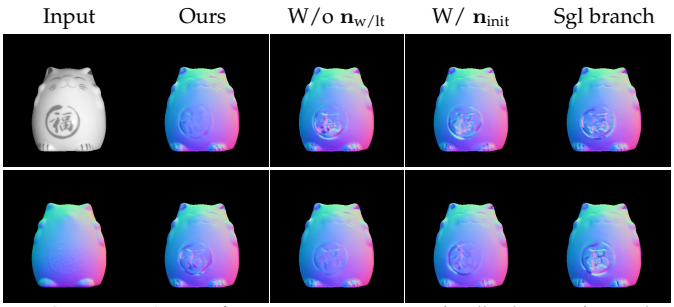


Fig. 7: Qualitative evaluation results in ablation studies of pNENet.

6.1 Ablation Studies and Analyses

To evaluate how manual settings in each model influence the final performance, we analyze our framework by changing or disabling each component respectively. The quantitative ablation study of the lighting module is included in Table 2. We further ablate pNENet on the synthetic dataset, SONY [44], and our real-world dataset. The quantitative evaluation is listed in Table 3, and visual results are shown in Fig. 6 and Fig. 7.

Effectiveness of the lighting branch We analyze the effectiveness of light direction estimation in pLTNet by comparing the vector and angle representations, as shown in Table 2. Modeling a light direction as a pair of the zenith and azimuth angles, we discretize the space of the two angles into 60 and 90 categories, respectively, and separately estimate the two parameters ($\text{Light}_{\text{angle-cls}}$). Besides, we directly predict the zenith and azimuth angle values ($\text{Light}_{\text{angle-reg}}$). As shown in Table 2, the lighting branches perform better with the vector representation of light directions, in accordance with the contiguous representation of 2D rotations [57]. We compare the classification method against the regression method for light intensity estimate. We train the classification model, which has the same architecture except for the number of neurons in output linear layers, until convergence. The quantitative evaluation in "Light" and "$\text{Light}_{\text{int-reg}}$" of Table 2 shows the classification method produces more accurate intensities. For intensity estimation, carefully converting regression to the multi-category classification problem could narrow down the solution space of intensities and balance the quantization error. Further, we train our network to take in a single unpolarized image instead of four polarization images, and the result of "Light" compared to "$\text{Light}_{\text{w/ unpol}}$" indicates polarization cues benefit light direction estimate.

Effectiveness of the normal optimization Comparing $\mathbf{n}_{\text{init}}$ and $\mathbf{n}_{\text{w/lt}}$ in Table 3 and Fig. 6, we find that the azimuth ambiguities in initial normals are greatly relieved with

TABLE 4: Quantitative evaluation on the synthetic data, SONY [44], and our real-world data, compared to the state-of-the-art methods. DeepSfP [15] and SfP-wild [21] are re-trained on our synthetic data for a fair comparison.

	Methods	Angular error ↓			Accuracy ↑		
		Mean	Median	ARMSE	11.25°	22.5°	30.0°
Sync.	Miyazaki [2]	43.79	39.70	50.86	.0857	.2432	.3631
	Mahmoud [9]	44.81	39.36	52.22	.0653	.2338	.3620
	Smith [6]	54.59	53.65	59.15	.0663	.1814	.2673
	Li [65]	42.62	35.37	51.70	.1158	.3187	.4440
	DeepSfP [15]	12.09	9.178	17.31	.6543	.8903	.9324
	Deschaintre [12]	15.07	10.29	23.21	.6243	.7866	.8409
	SfP-wild [21]	5.690	**3.927**	10.59	**.9337**	.9744	.9793
	Ours	**4.924**	3.981	**6.573**	.9294	**.9873**	**.9940**
SONY [44]	Miyazaki [2]	39.21	35.57	45.74	.1058	.2919	.4191
	Mahmoud [9]	43.22	37.70	51.47	.1006	.2811	.3973
	Smith [6]	40.52	38.50	46.06	.0714	.2311	.3618
	Li [65]	22.34	17.12	32.68	.2766	.6910	.8426
	DeepSfP [15]	11.26	8.328	15.65	.6887	.8868	.9334
	Deschaintre [12]	12.45	8.288	17.79	.6653	.8576	.9056
	SfP-wild [21]	10.49	8.126	14.29	.7427	.9153	.9488
	Ours	**8.868**	**5.682**	**13.32**	**.7853**	**.9181**	**.9505**
Our Real	Miyazaki [2]	45.80	41.63	52.76	.0610	.2054	.3233
	Mahmoud [9]	45.30	38.46	53.36	.0490	.2214	.3631
	Smith [6]	61.28	59.94	65.01	.0377	.1270	.2050
	Li [65]	28.19	21.62	37.96	.1786	.5280	.7180
	DeepSfP [15]	16.88	13.16	22.07	.4082	.7985	.8846
	Deschaintre [12]	21.24	17.54	26.64	.2695	.6587	.7993
	SfP-wild [21]	17.62	14.47	21.47	.3144	.7934	.8900
	Ours	**15.58**	**12.65**	**19.64**	**.4387**	**.8275**	**.9111**

TABLE 5: Quantitative comparison between our model and the state-of-the-art SVBRDF methods, evaluated on synthetic data.

Methods	Albedo		Roughness		Rendering	
	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓
Li [65]	.0954	.2053	.1047	.2477	.0834	.3452
Deschaintre [12]	.0727	.1963	.1162	.2783	.1228	.4398
Ours	**.0174**	**.0870**	**.0347**	**.1195**	**.0289**	**.1907**

predicted lighting and albedo information. For evaluation of the influence of refractive indices in the normalization optimization, we compute the optimized normal on the SONY dataset [44] and our synthetic data with η ranging from 1.4 to 1.6. As shown in Table 3, the angular errors of optimized normal are close between η=1.5 and η=1.6, and we fix η of 1.5 in the optimization process of our framework.

Effectiveness of pNENet Quantitative and qualitative results are shown in Table 3 and Fig. 7. We first verify the contribution of the normal optimization module by 1) removing $\mathbf{n}_{\text{w/lt}}$ from the input of the network; 2) using $\mathbf{n}_{\text{init}}$ as a substitute for $\mathbf{n}_{\text{w/lt}}$. Angular error and accuracy metrics averaged over 7,800 samples are listed in "pNENet$_{\text{w/o } \mathbf{n}_{\text{w/lt}}}$" and "pNENet$_{\text{w/}\mathbf{n}_{\text{init}}}$", respectively. pNENet produces more reliable results under the guidance of optimized normal maps, and the initial normal also contributes to the final results but does not work well as the optimized one. We compare our two-branch architecture to a network with a single stream, *i.e.*, we remove the branch fed with $\mathbf{n}_{\text{w/lt}}$ and stack $\mathbf{n}_{\text{w/lt}}$ with the albedo map and polarization images as input. The results in the "pNENet$_{\text{sgl branch}}$" demonstrate the effectiveness of the two-branch design of pNENet. We also train pNENet with the concatenation fusion method instead of the addition operation. Our network performs slightly better with less number of parameters (20.4M) than "pNENet$_{\text{concat fuse}}$" (22.7M).

Moreover, the adopted diffuse-rendering error loss is evaluated in "pNENet$_{\text{w/o err-rw loss}}$" row of Table 3. By giving additional weight to the loss function on pixels of inaccurate normal prior, the diffuse-rendering term helps the network to concentrate more on the local normal recovery in the specular regions. We then investigate the benefit of the predicted albedo to final normal estimations. pNENet is trained without the object albedo as input, and the results are in "pNENet$_{\text{w/o albedo}}$", showing that more prior knowledge encoded in the network further improves the model performance. Additionally, we find that concatenating the mask into input tensors also enables better estimation of surface normals, as listed in "pNENet$_{\text{w/o mask}}$".

Overall, it is the additional input $\mathbf{n}_{\text{w/lt}}$, the designed architecture and the loss functions that contribute most to the final normal. By introducing $\mathbf{n}_{\text{w/lt}}$ as input, our pNENet can generate results slightly better than SfP-wild [21]. Then the two-branch network structure is designed to effectively utilize the optimized normal, and the loss function helps further improve the normal accuracy.

6.2 Evaluation on Synthetic Data

We use 7,800 rendered polarization images and their corresponding ground-truth normal maps and texture maps as the test set to quantitatively evaluate our method against the SfP and SVBRDF approaches. Miyazaki *et al*. [2], Mahmoud *et al*. [9], and Smith *et al*. [6] are non-learning methods, which predict the surface normal based on polarization cues and other priors. Li *et al*. [65] take a single image for SVBRDF recovering and Deschaintre *et al*.'s method [12] is based on a deep network utilizing the polarization images. DeepSfP [15] is proposed to exploit a network to solve the ambiguity in traditional SfP. For a fair comparison, we retrain the model of DeepSfP [15] using our dataset with the same strategy as stated in their paper. Recently presented SfP-wild [21] aims at estimating the normal of the scene. To make their method applicable on the object level, we also retrained the model on our dataset until it gets converged.

Qualitative evaluation of our method on the synthetic data is shown in the first row of Fig. 6. Quantitative comparisons of normal predictions among these methods are listed in Table 4. DeepSfP [15], Deschaintre *et al*. [12], SfP-wild [21], and our method performs significantly better than other methods. Compared to Li *et al*. [65], which only relies on a single image, SfP methods produce more reliable normal maps with less angular error. With the assistance of the lighting information, our method outperforms DeepSfP [15], Deschaintre *et al*. [12], and SfP-wild [21]. We quantitatively compare our method to Deschaintre *et al*. [12] and Li *et al*. [65] regarding albedo, roughness, and re-rendered images. The results are shown in Table 5. Our framework also generates compelling results as the state-of-the-art approaches.

6.3 Evaluation on Real Dataset

We take SONY dataset [44] and our collected real dataset as the benchmark to evaluate SfP methods. Table 2 contains the quantitative evaluation of pLTNet, and Fig. 6 shows some visual results of our method. More intermediate results (*e.g.*, albedo) on our real dataset are provided in the supplement. pNENet is quantitatively and qualitatively compared with previous state-of-the-art methods on SONY dataset [44], as shown in Fig. 8 and Table 4. Since the non-learning methods [2], [6], [9] show large errors on the real-world images, only the visual results of [9] are displayed in Fig. 8. All the listed methods fail to accurately recover the normals near the edge of DOLL's handbag. pBRDF changes with the albedo near the handbag contour, which makes the polarimetric information unstable and also affects the predictions of SfP methods. We conduct further analysis on this phenomenon in Sec. 4.2 of the supplement. SfP-wild [21] also produces comparable normal maps as our method. However, the "sharper" results of SfP-wild [21] suffer from more artifacts near the edges, such as the SPARROW tail and the SWAN feather. Our model generalizes well on real-world images and produces reliable normal maps on a majority of these objects. More qualitative comparisons on the rest of the objects are provided in the supplementary material.

We also conduct qualitative evaluation on the real data released by Deschaintre *et al*. [12], as shown in Fig. 9. Our method generates quality reliable normal maps with fewer artifacts, such as normals at the edge of the raspberryPi. The images re-rendered by our method are also closer to the original input than the rendered results of Deschaintre *et al*. [12] and Li *et al*. [65]. The rest of the results are provided in the supplementary material.

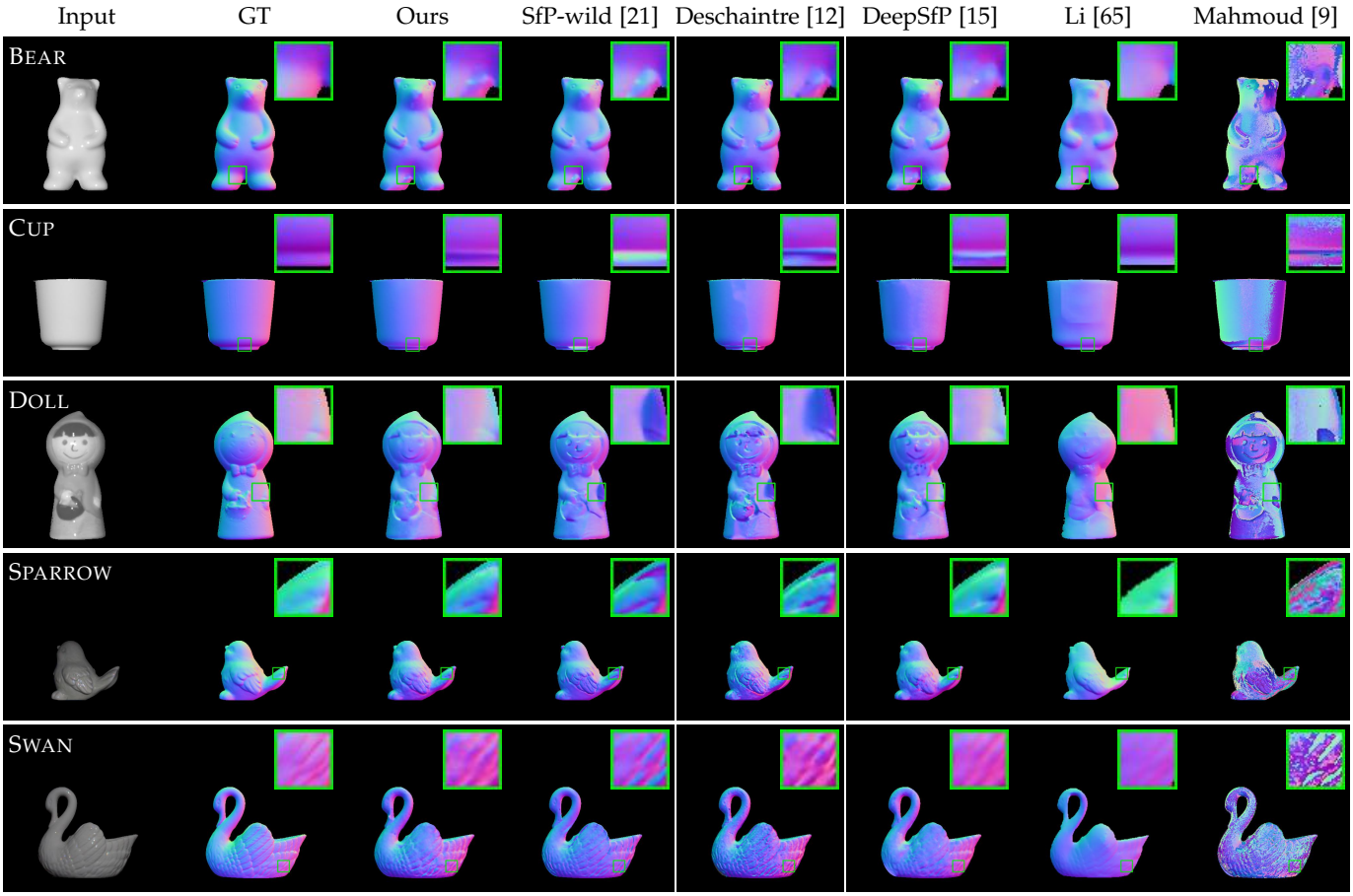


Fig. 8: Qualitative comparisons among the state-of-the-art approaches, SfP-wild [21], Deschaintre *et al*. [12], DeepSfP [15], Li *et al*. [65], and Mahmoud *et al*. [9], evaluated on SONY dataset [44] and our captured real data.

TABLE 6: Model parameters, FLOPs, and inference time comparison between our method and learning-based approaches. N, A, R, S, D, and L are abbreviations for normal, albedo, roughness, specular reflection, depth, and lighting, respectively.

	Ours	Deschaintre	Li	DeepSfP	SfP-wild
Output	N, A, R, L	N, A, R, S, D	N, A, R, L, D	N	N
# Parameters	35.7M	188.3M	74.6M	10.8M	42.5M
FLOPs	113.5G	176.3G	225.4G	77.78G	195.4G
Time (sec/iter)	0.733	0.882	2.106	0.720	0.806

6.4 Model Complexity and Inference Time

We compare the computational costs of Deschaintre *et al*. [12] (a single-stage model), Li *et al*. [65] (a multi-stage model, iterative refinement), DeepSfP [15] (a single-stage model), SfP-wild [21] (a single-stage model for scene-level SfP), and our framework (a two-stage model) on Manjaro Linux with an AMD Ryzen9 5950X CPU and an NVIDIA GeForce RTX 3090. The FLOPs of each method are measured by processing a single test sample with a size of 512×512. For testing running time, we set the batch size to 32 and run the five models on 8,000 samples, and the results are listed in Table 6. Our method has advantages over [12], [21], [65] in inference time, and also achieves comparable performance compared to [15]. Thanks to the integration of shading and polarimetric information, our framework predicts normal map as well as texture information and consists of moderate model parameters like SfP approaches.

7 Limitations and Discussions

Since our model is designed under the setup of a single distant light, it will degenerate when dealing with images captured under different lighting conditions, such as outdoor lighting (polarization of the sky should be considered) and indoor scenes with multiple light sources (specular reflection cannot be ignored, which makes the diffuse dominant basis invalid). To evaluate the robustness of the method in more complex environments, we take polarization images under a distant light along with different ceiling lights (mimic ambient lighting), as shown in Fig. 10. Our method still performs well under distant and ambient lighting (Fig. 10(a)(b)), but we find that the estimated normal quality downgrades as the intensity of ambient lighting increases (Fig. 10(b)). Without the distant lighting, the normal maps shown in Fig. 10(c) suffer from artifacts due to the influence of specular reflection. This experiment qualitatively demonstrates that our method works reliably in the scenarios where distant light dominates and also echoes our analysis that distant lighting causes stronger diffuse reflectance. Moreover, our framework does not model global illumination such as inter-reflection, which causes the artifacts near the leg of BEAR in Fig. 8. To further analyze this issue, we test our method on a statue with a concave shape under a single distant light, as shown in Fig. 10(d). The cast-shadow area marked by the green box and the concave surface marked by the red box are mainly illuminated by the reflected light from other surface

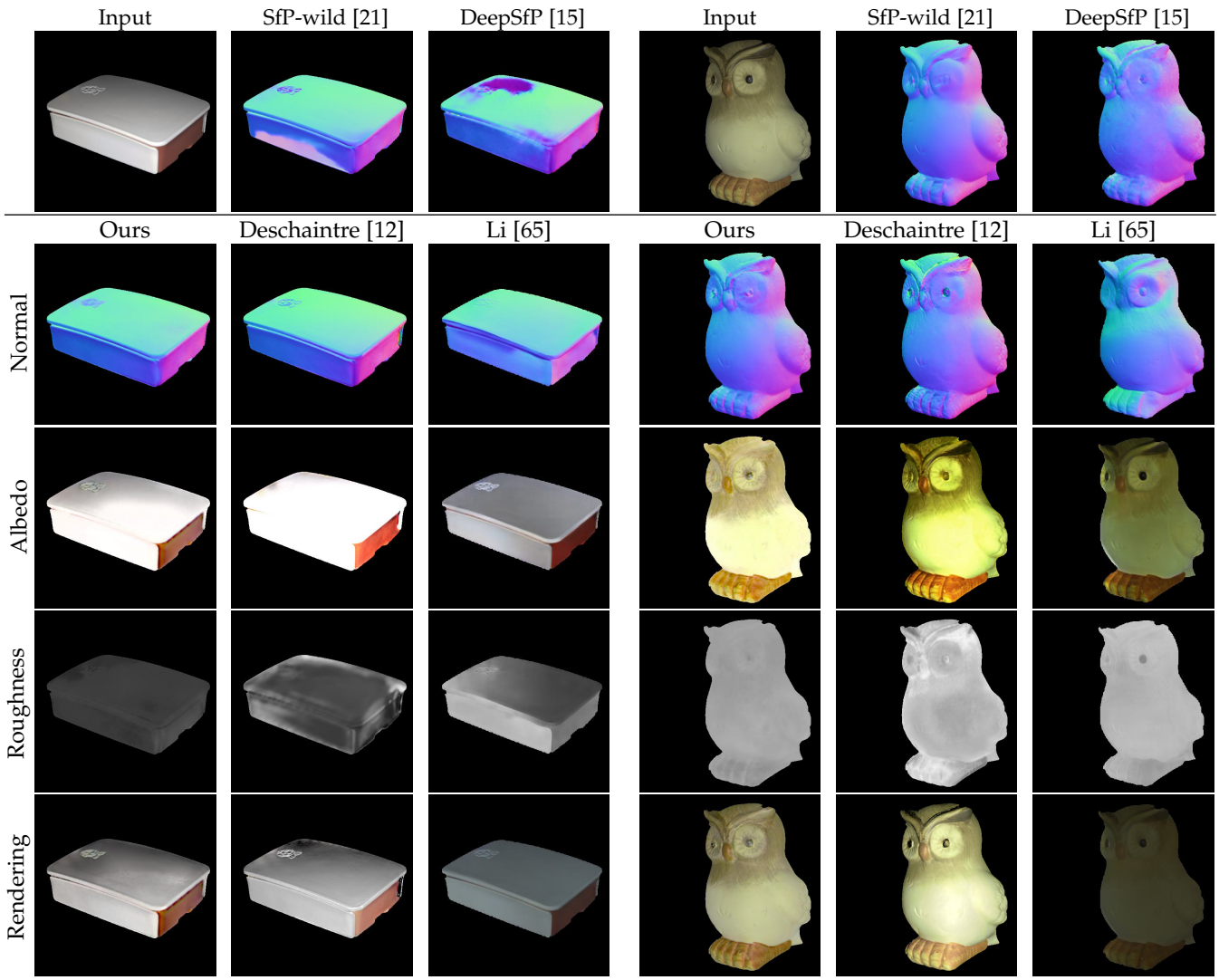


Fig. 9: Qualitative comparisons between learning-based SfP/SVBRDF approaches and our method, evaluated on the real data released by [12].

points. The intricate inter-reflection violates the assumption of unpolarized incident light and consequently degrades the quality of the estimated normal map.

Like previous SfP methods, our model is designed for dielectric objects but may not work on very rough surfaces or conductors. We test SfP methods on three challenging materials, *i.e.*, fabric, rough clay, and brass, as shown in Fig. 11[3]. The inter-reflection and scattering in microfacets of the rough surface of fabric and clay depolarize the exitant light making polarization information (like AoLP) invalid, so SfP methods hardly recover the normal azimuth angle of MONKEY and BUDDHA1. The refractive index of metallic material like brass is a complex number and the pBRDF is also different from that of dielectric [66], so AoLP of BUDDHA2 seems to be affected by noise. SfP of these challenging materials is still an open problem and remains to be solved.

8 Conclusion

In this paper, we propose a learning-based model consisting of a polarimetric lighting and texture estimate module and a normal recovery model for surface normal reconstruction. Since the polarization properties of objects vary with the lighting conditions, we investigate the SfP problem under distant lighting. We derive the shading constraints from polarimetric BRDF and exploit the lighting and albedo cues to enable disambiguation of the azimuth angle. Overall, we derive the two-stage network to achieve shape recovery from polarization: in the first stage, pLTNet takes polarization images to predict lighting conditions and object texture; in the second stage, pNENet makes full use of polarization cues, physical priors calculated by SfP, and lighting information from pLTNet to generate compelling normal maps. By introducing the lighting constraint in SfP, the network uses shading information to assist in resolving normal estimation ambiguity. Experimental results show our approach has superior performance over previous work.

3. Qualitative results of other methods and quantitative evaluation are provided in Sec. 4.3 of the supplement.

Acknowledgments

This work is supported by the National Key R&D Program of China (2021ZD0109800), National Natural Science Foundation of China under Grand No. 62136001, 62088102, program for Youth Innovative Research Team of BUPT

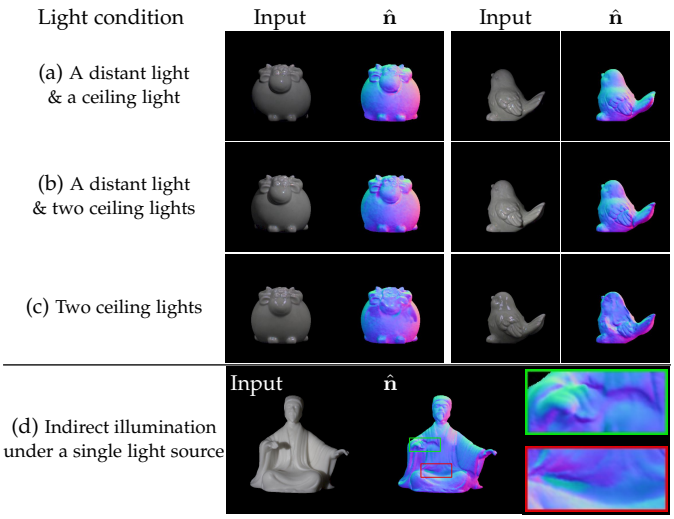


Fig. 10: Qualitative evaluation of our method on the data collected under distant and environment lighting (a)(b)(c), and indirect illumination with a single light source (d).

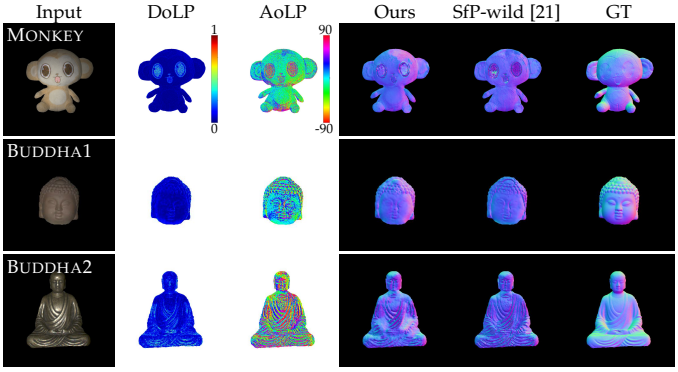


Fig. 11: Qualitative results and polarization properties of the objects made from fabric, rough clay, and brass.

No. 2023QNTD02, and BUPT Excellent Ph.D. Students Foundation No. CX2022233.

REFERENCES

[1] P. Fankhauser, M. Bloesch, D. Rodriguez, R. Kaestner, M. Hutter, and R. Siegwart, "Kinect V2 for mobile robot navigation: Evaluation and modeling," in *Proc. International Conference on Advanced Robotics*, 2015.

[2] D. Miyazaki, M. Kagesawa, and K. Ikeuchi, "Polarization-based transparent surface modeling from two views," in *Proc. International Conference on Computer Vision*, 2003.

[3] S. Rahmann and N. Canterakis, "Reconstruction of specular surfaces using polarization imaging," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2001.

[4] K. E. Torrance and E. M. Sparrow, "Theory for off-specular reflection from roughened surfaces," *Journal of the Optical Society of America*, 1967.

[5] W. A. Smith, R. Ramamoorthi, and S. Tozza, "Linear depth estimation from an uncalibrated, monocular polarisation image," in *Proc. European Conference on Computer Vision*, 2016.

[6] ——, "Height-from-polarisation with unknown lighting or albedo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[7] G. A. Atkinson and E. R. Hancock, "Recovery of surface orientation from diffuse polarization," *IEEE Transactions on Image Processing*, 2006.

[8] C. P. Huynh, A. Robles-Kelly, and E. Hancock, "Shape and refractive index recovery from single-view polarisation images," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2010.

[9] A. H. Mahmoud, M. T. El-Melegy, and A. A. Farag, "Direct method for shape recovery from polarization and shading," in *Proc. International Conference on Image Processing*, 2012.

[10] Y. Fukao, R. Kawahara, S. Nobuhara, and K. Nishino, "Polarimetric normal stereo," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2021.

[11] T. Ichikawa, M. Purri, R. Kawahara, S. Nobuhara, K. Dana, and K. Nishino, "Shape from sky: Polarimetric normal recovery under the sky," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2021.

[12] V. Deschaintre, Y. Lin, and A. Ghosh, "Deep polarization imaging for 3D shape and SVBRDF acquisition," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2021.

[13] S.-H. Baek, D. S. Jeon, X. Tong, and M. H. Kim, "Simultaneous acquisition of polarimetric SVBRDF and normals." *ACM Transactions on Graphics*, 2018.

[14] I. Hwang, D. S. Jeon, A. Muñoz, D. Gutierrez, X. Tong, and M. H. Kim, "Sparse ellipsometry: portable acquisition of polarimetric SVBRDF and shape with unstructured flash photography," *ACM Transactions on Graphics*, 2022.

[15] Y. Ba, A. Gilbert, F. Wang, J. Yang, R. Chen, Y. Wang, L. Yan, B. Shi, and A. Kadambi, "Deep shape from polarization," in *Proc. European Conference on Computer Vision*, 2019.

[16] F. Logothetis, R. Mecca, F. Sgallari, and R. Cipolla, "A differential approach to shape from polarisation: A level-set characterisation," *Springer International Journal of Computer Vision*, 2019.

[17] S. Tozza, W. A. Smith, D. Zhu, R. Ramamoorthi, and E. R. Hancock, "Linear differential constraints for photo-polarimetric height estimation," in *Proc. International Conference on Computer Vision*, 2017.

[18] Y. Yu, D. Zhu, and W. A. Smith, "Shape-from-polarisation: a nonlinear least squares approach," in *Proc. International Conference on Computer Vision Workshops*, 2017.

[19] S. Tozza, D. Zhu, W. A. P. Smith, R. Ramamoorthi, and E. R. Hancock, "Uncalibrated, two source photo-polarimetric stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 5747–5760, 2022.

[20] Y. Kondo, T. Ono, L. Sun, Y. Hirasawa, and J. Murayama, "Accurate polarimetric BRDF for real polarization scene rendering," in *Proc. European Conference on Computer Vision*, 2020.

[21] C. Lei, C. Qi, J. Xie, N. Fan, V. Koltun, and Q. Chen, "Shape from polarization for complex scenes in the wild," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2022.

[22] C. P. Huynh, A. Robles-Kelly, and E. R. Hancock, "Shape and refractive index from single-view spectro-polarimetric images," *Springer International Journal of Computer Vision*, 2013.

[23] A. Kadambi, V. Taamazyan, B. Shi, and R. Raskar, "Polarized 3D: High-quality depth sensing with polarization cues," in *Proc. International Conference on Computer Vision*, 2015.

[24] ——, "Depth sensing using geometrically constrained polarization normals," *Springer International Journal of Computer Vision*, 2017.

[25] K. Berger, R. Voorhies, and L. H. Matthies, "Depth from stereo polarization in specular scenes for urban robotics," in *Proc. International Conference on Robotics and Automation*, 2017.

[26] D. Zhu and W. A. Smith, "Depth from a polarisation+ RGB stereo pair," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2019.

[27] L. Yang, F. Tan, A. Li, Z. Cui, Y. Furukawa, and P. Tan, "Polarimetric dense monocular slam," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2018.

[28] D. Miyazaki, M. Kagesawa, and K. Ikeuchi, "Transparent surface modeling from a pair of polarization images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004.

[29] G. A. Atkinson and E. R. Hancock, "Multi-view surface reconstruction using polarization," in *Proc. International Conference on Computer Vision*, 2005.

[30] Z. Cui, J. Gu, B. Shi, P. Tan, and J. Kautz, "Polarimetric multi-view stereo," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2017.

[31] D. Miyazaki, T. Shigetomi, M. Baba, R. Furukawa, S. Hiura, and N. Asada, "Surface normal estimation of black specular objects from multiview polarization images," *Optical Engineering*, 2016.

[32] D. Miyazaki, R. Furuhashi, and S. Hiura, "Shape estimation of concave specular object from multiview polarization," *Journal of Electronic Imaging*, 2020.

[33] Y. Ding, Y. Ji, M. Zhou, S. B. Kang, and J. Ye, "Polarimetric helmholtz stereopsis," in *Proc. International Conference on Computer Vision*, 2021.

[34] G. Chen, L. He, Y. Guan, and H. Zhang, "Perspective phase angle model for polarimetric 3d reconstruction," in *Proc. European Conference on Computer Vision*, 2022.
[35] L. Chen, Y. Zheng, A. Subpa-asa, and I. Sato, "Polarimetric three-view geometry," in *Proc. European Conference on Computer Vision*, 2018.
[36] Z. Cui, V. Larsson, and M. Pollefeys, "Polarimetric relative pose estimation," in *Proc. International Conference on Computer Vision*, 2019.
[37] M. W. H. IV, J. D. Schmidt, and M. J. Havrilla, "A geometrical optics polarimetric bidirectional reflectance distribution function for dielectric and metallic surfaces," *Opt. Express*, 2009.
[38] Y. Zhang, Y. Zhang, H. Zhao, and Z. Wang, "Improved atmospheric effects elimination method for pBRDF models of painted surfaces," *Opt. Express*, 2017.
[39] I. G. E. Renhorn, T. Hallberg, D. Bergström, and G. D. Boreman, "Four-parameter model for polarization-resolved rough-surface BRDF," *Opt. Express*, 2011.
[40] Y. Li, H. Lu, H.-Y. Shum *et al.*, "Multiple-cue illumination estimation in textured scenes," in *Proc. International Conference on Computer Vision*, 2003.
[41] Y. Wang and D. Samaras, "Estimation of multiple illuminants from a single image of arbitrary known geometry," in *Proc. European Conference on Computer Vision*, 2002.
[42] D. Zoran, D. Krishnan, J. Bento, and B. Freeman, "Shape and illumination from shading using the generic viewpoint assumption," in *Proc. Advances in Neural Information Processing Systems*, 2014.
[43] G. Chen, K. Han, B. Shi, Y. Matsushita, and K.-Y. K. Wong, "Self-calibrating deep photometric stereo networks," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2019.
[44] T. T. Ngo, H. Nagahara, and R. ichiro Taniguchi, "Surface normals and light directions from shading and polarization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
[45] H. Weber, D. Prévost, and J. Lalonde, "Learning to estimate indoor lighting from 3D objects," in *Proc. International Conference on 3D Vision*, 2018.
[46] R. Yi, C. Zhu, P. Tan, and S. Lin, "Faces as lighting probes via unsupervised deep highlight extraction," in *Proc. European Conference on Computer Vision*, 2018.
[47] M.-A. Gardner, Y. Hold-Geoffroy, K. Sunkavalli, C. Gagne, and J.-F. Lalonde, "Deep parametric indoor lighting estimation," in *Proc. International Conference on Computer Vision*, 2019.
[48] J. Zhang, K. Sunkavalli, Y. Hold-Geoffroy, S. Hadap, J. Eisenmann, and J.-F. Lalonde, "All-weather deep outdoor lighting estimation," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2019.
[49] M. Garon, K. Sunkavalli, S. Hadap, N. Carr, and J.-F. Lalonde, "Fast spatially-varying indoor lighting estimation," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2019.
[50] E. Collett, *Field Guide to Polarization*, ser. Field Guides. Society of Photo Optical, 2005.
[51] M. Garcia, C. Edmiston, R. Marinov, A. Vail, and V. Gruev, "Bio-inspired color-polarization imager for real-time in situ imaging," *Optica*, 2017.
[52] G. Turk and M. Levoy, "Zippered polygon meshes from range images," in *Proc. ACM SIGGRAPH*, 1994.
[53] B. Walter, S. R. Marschner, H. Li, and K. E. Torrance, "Microfacet models for refraction through rough surfaces," in *Proc. Eurographics Conference on Rendering Techniques*, 2007.
[54] E. Heitz, "Understanding the masking-shadowing function in microfacet-based BRDFs," *Journal of Computer Graphics Techniques (JCGT)*, 2014.
[55] P. Debevec, "Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography," in *Proc. ACM SIGGRAPH*, 1998.
[56] J. Ren, F. Wang, J. Zhang, Q. Zheng, M. Ren, and B. Shi, "DiLiGenT10^2: A photometric stereo benchmark dataset with controlled shape and material variation," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2022.
[57] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5745–5753.
[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2016.
[59] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2019.
[60] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Advances in Neural Information Processing Systems*, 2019.
[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[62] B. Shi, Z. Mo, Z. Wu, D. Duan, S. Yeung, and P. Tan, "A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
[63] M. Nimier-David, D. Vicini, T. Zeltner, and W. Jakob, "Mitsuba 2: A retargetable forward and inverse renderer," *ACM Transactions on Graphics*, vol. 38, no. 6, 2019.
[64] L. Demes, "ambientcg - free public domain pbr materials," https://ambientcg.com/, accessed: 2021-11-08.
[65] Z. Li, Z. Xu, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker, "Learning to reconstruct shape and spatially-varying reflectance from a single image," *ACM Transactions on Graphics*, 2018.
[66] A. Ghosh, T. Chen, P. Peers, C. A. Wilson, and P. Debevec, "Circularly polarized spherical illumination reflectometry," *ACM Transactions on Graphics*, 2010.

Youwei Lyu received his B.S. degree from Beijing University of Posts and Telecommunications in 2019. He is currently studying at Beijing University of Posts and Telecommunications. His research interests are centered around computational photography and physics-based vision.

Lingran Zhao received his B.S. degree from Peking University in 2023. He is currently working toward an M.S. degree at Carnegie Mellon University. His research interests include computer vision, neural network compression, and multi-modality learning.

Si Li received the Ph.D. degree from the Beijing University of Posts and Telecommunications in 2012. She is currently an Associate Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. Her research interests include multimodal artificial intelligence and machine learning.

Boxin Shi (Senior Member, IEEE) received the BE degree from the Beijing University of Posts and Telecommunications, the ME degree from Peking University, and the PhD degree from the University of Tokyo, in 2007, 2010, and 2013. He is currently a Boya Young Fellow Assistant Professor and Research Professor at Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, National Institute of Advanced Industrial Science and Technology, from 2013 to 2017. His papers were awarded as Best Paper Runner-Up at ICCP 2015 and selected as Best Papers from ICCV 2015 for IJCV Special Issue. He is an associate editor of TPAMI/IJCV and an area chair of CVPR/ICCV.

Shape from Polarization with Distant Lighting Estimation Supplementary Material

Youwei Lyu, Lingran Zhao, Si Li, and Boxin Shi*, *Senior Member, IEEE*

1 Polarimetric BRDF and Fresnel Equations

In this section, we provide more details about the adopted polarimetric Bidirectional Reflectance Distribution Function (pBRDF) and related notation definitions of the symbols in this paper. We then derive that the Fresnel term T_i^+ is a function regarding $\cos\theta_i$, as mentioned in Sec. 3.2 of the main paper.

1.1 Stokes Vector and Rotation Matrix

The Stokes vector is applied to describe the polarization status of the light for better illustration of the pBRDF model, which should be defined in a specific coordinate frame. In this paper, we assume the z-axis of the local frame is aligned with the propagation direction of the light, and the y-axis of the local frame is always in the plane of incidence (PoI) or in the plane of exitant (PoE). The PoI is the plane consisting of the surface normal vector and the incident lighting direction vector, and the PoE is the plane containing the surface normal vector and the exitant lighting direction vector.

Then we employ the Mueller matrix to describe changes in polarization states of light, which operates on the Stokes vectors. Before applying the transformation on the Stokes vector, different coordinates of the incident and exitant vectors should be aligned to the same one. Thus, the rotation matrix is introduced to convert the two coordinates frames,

$$\mathbf{C}(\phi) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos 2\phi & \sin 2\phi \\ 0 & -\sin 2\phi & \cos 2\phi \end{bmatrix}, \tag{1}$$

where ϕ denotes the counter-clockwise rotation angle.

*Corresponding author.

- *Youwei Lyu and Si Li are with School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China. Email: {youweilv, lisi}@bupt.edu.cn.*
- *Lingran Zhao and Boxin Shi are with National Key Laboratory for Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing 100871, China. Email: {calvinzhao, shiboxin}@pku.edu.cn.*

1.2 Fresnel Matrix

Fresnel Matrices [1] are widely used to formulate the transmission and reflection process of polarized light, which are denoted as

$$\mathbf{F} = \begin{bmatrix} F^+ & F^- & 0 \\ F^- & F^+ & 0 \\ 0 & 0 & \sqrt{F^\perp F^\parallel}\cos\delta \end{bmatrix}, \begin{matrix} \mathbf{F} \in \{\mathbf{R}, \mathbf{T}\}, \\ F \in \{R, T\}, \end{matrix} \tag{2}$$

in which δ represents the retardation phase shift between the perpendicular and parallel waves. In dielectric materials, $\delta = \pi$ for any incident angle less than the Brewster angle, and $\delta = 0$ otherwise. $\mathbf{R}$, $\mathbf{T}$ denotes Fresnel Matrices corresponding to the reflected light and transmitted light, and R, T are relative strength of reflected light and transmitted light, respectively, regarding the incident angle θ_1 and exitant angle θ_2 in the transmission process. We let $F^+ = \frac{F^\perp + F^\parallel}{2}, F^- = \frac{F^\perp - F^\parallel}{2}, F \in \{R, T\}$ for conciseness. The superscripts $\perp, \parallel$ correspond to the perpendicular and parallel components of the light. Specifically,

$$T^\perp = \frac{\sin 2\theta_1 \sin 2\theta_2}{\sin^2(\theta_1+\theta_2)}, \; T^\parallel = \frac{\sin 2\theta_1 \sin 2\theta_2}{\sin^2(\theta_1+\theta_2)\cos^2(\theta_1-\theta_2)}, \tag{3}$$

$$R^\perp = 1 - T^\perp, \;\; R^\parallel = 1 - T^\parallel, \tag{4}$$

and we could calculate the exitant angle with the known incident angle and the refractive indices of the two types of medium by Snell's law [1],

$$\cos\theta_2 = \sqrt{1 - \left(\frac{\eta_1}{\eta_2}\sin\theta_1\right)^2} = \frac{1}{\eta_2}\sqrt{\eta_2^2 - \eta_1^2 + \eta_1^2\cos^2\theta_1}. \tag{5}$$

Also, the incident angle can be computed with the known exitant angle and the refractive indices $\eta_{1,2}$.

1.3 Diffuse and Specular Reflection

The observed Stokes vector $\mathbf{s}_o$ of exitant light can be expressed by matrix multiplication of a Mueller matrix $\mathbf{M}$ and the incident Stokes vector $\mathbf{s}_i$ [2]:

$$\mathbf{s}_o = \mathbf{M}\mathbf{s}_i, \tag{6}$$

where the Mueller matrix $\mathbf{M}$ is concerning the surface normal, the light direction, and the BRDF of the surface.

We adopt the diffuse-specular pBRDF [3] for analysis of distant lighting constraints and simulation of training data. According to the polarimetric model [3], the observed polarization light is the sum of the diffuse component and specular component, *i.e.*, we could split the Mueller matrix into two parts: $\mathbf{M} = \mathbf{M}_d + \mathbf{M}_s$, and Eq. (6) could also be rewritten as:

$$\mathbf{s}_o = \mathbf{s}_{o,d} + \mathbf{s}_{o,s} = (\mathbf{M}_d(\mathbf{l}) + \mathbf{M}_s(\mathbf{l}))\mathbf{s}_i, \tag{7}$$

in which we denote the Muller matrix regarding the light direction $\mathbf{l}$ for brevity. We then elaborate the two components as follows.

Diffuse reflection The incident light first penetrates into the object's surface, gets depolarized inside the material, transmits back to the air, and becomes polarized. The Mueller matrix describing the diffuse reflection process is given by

$$\mathbf{M}_d = \mathbf{C}_{n\to o}(\psi_o)\mathbf{T}(\theta_o;\eta)\mathbf{D}(a)\mathbf{T}(\theta_i;\eta)\mathbf{C}_{i\to n}(-\psi_i)(\mathbf{n}\cdot\mathbf{l}), \tag{8}$$

in which $\mathbf{n}$ is the unit vector of the surface normal. θ_i denotes the angle between $\mathbf{n}$ and $\mathbf{l}$, and θ_o is the zenith angle of surface normal. $\mathbf{C}_{n\to o}(\psi_o)$ denotes the rotation transformation with ψ_o from the PoI coordinate to the imaging coordinate, and $\mathbf{C}_{i\to n}(-\psi_i)$ denotes the rotation matrix of the angle $-\psi_i$ from the incident coordinate to the PoI coordinate; the rotation angles ψ_o can be calculated with known surface normals,

$$\phi_o = \frac{\pi}{2} - \arctan\frac{n_y}{n_x}, \tag{9}$$

where (n_x, n_y) is the projection of $\mathbf{n}$ on the imaging coordinate. In Eq. (8), $\mathbf{D}(a)$ is a depolarization matrix with diffuse albedo a:

$$\mathbf{D}(a) = \begin{bmatrix} a & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{10}$$

Specular reflection For specular reflection, the incident light hits the object's surface and then is directly reflected back to the air, which can be described by the half vector $\mathbf{h}$ between the direction of incident light $\mathbf{l}$ and the view direction $\mathbf{v}$. Under orthographic projection, we set $\mathbf{v} = [0, 0, 1]^\top$. The Mueller matrix of specular reflection is given by

$$\mathbf{M}_s = \frac{D(\theta_h;\sigma)G(\theta_i,\theta_o;\sigma)}{4\cos\theta_i\cos\theta_o}\mathbf{C}_{h\to o}(\varphi_o)\mathbf{R}(\theta_d;\eta)\mathbf{C}_{i\to h}(-\varphi_i)(\mathbf{n}\cdot\mathbf{l}), \tag{11}$$

where $D(\theta_h;\sigma)$ is the normal distribution function (NDF) of microsurface [4] with regard to θ_h, the angle between the halfway vector and surface normal, and σ, the roughness of the surface. θ_h can be computed from $\mathbf{n}$ and $\mathbf{l}$, and we could write this term as $D(\mathbf{l},\mathbf{n};\sigma)$, as mentioned in the main paper. $G(\theta_i,\theta_o;\sigma)$ is a function corresponding to D for geometric consistency. We adopt the Trowbridge-Reitz NDF (GGX distribution [4]) and its corresponding geometry function in the paper. $\mathbf{C}_{h\to o}(\varphi_o)$ is the rotation matrix of the angle φ_o from the PoI coordinate to the imaging coordinate, while $\mathbf{C}_{i\to h}(-\varphi_i)$ denotes the rotation matrix of angle $-\varphi_i$ from the incident coordinate to the PoI coordinate; the rotating angles φ_o can be calculated from the halfway vector $\mathbf{h}$,

$$\varphi_o = \frac{\pi}{2} - \arctan\frac{h_y}{h_x}, \tag{12}$$

where (h_x, h_y) is the projection of $\mathbf{h}$ on the imaging coordinate.

Moreover, $\mathbf{R}(\theta_d;\eta)$ is the reflection Fresnel term in relation to θ_d, the angle between the halfway vector $\mathbf{h}$ and direction of incident light $\mathbf{l}$. The incident light is assumed to be unpolarized, *i.e.*, $\mathbf{s}_i = [e, 0, 0]^\top$, in which e is the light intensity. Replacing the diffuse and specular Mueller matrix terms in Eq. (7) by Eq. (8) and Eq. (11), we finally derive the observed Stokes vector:

$$\begin{aligned}\mathbf{s}_o = &\begin{bmatrix} T^+(\theta_o) \\ T^-(\theta_o)\cos 2\psi_o \\ -T^-(\theta_o)\sin 2\psi_o \end{bmatrix} aT^+(\theta_i)(\mathbf{n}\cdot\mathbf{l})e + \\ &\frac{D(\theta_h;\sigma)G(\theta_i,\theta_o;\sigma)}{4\cos\theta_o\cos\theta_i}\begin{bmatrix} R^+(\theta_d) \\ R^-(\theta_d)\cos 2\varphi_o \\ -R^-(\theta_d)\sin 2\varphi_o \end{bmatrix}(\mathbf{n}\cdot\mathbf{l})e.\end{aligned} \tag{13}$$

For brevity, we replace $T^\pm(\theta_o)$ with $T_o^\pm$, substitute $D(\mathbf{l},\mathbf{n};\sigma)$ for $D(\theta_d,\theta_h;\sigma)$, substitute $R^\pm$ for $R^\pm(\theta_d)$, and then obtain Eq. (7) in the main paper.

1.4 Proof of Fresnel Functions

We state in the main paper that T_i^+ is a function regarding $\cos\theta_i$, the proof is given as follows. According to Sec. 1.2,

$$T_i^+ = \frac{T^\perp(\theta_i) + T^\parallel(\theta_i)}{2}. \tag{14}$$

For the incident process, the light ray penetrates the object medium with the refractive index $\eta_2 = \eta$ from the air with $\eta_1 = 1$, and we write Eq. (3) as

$$T^\perp(\theta_i) = \frac{4\eta\cos\theta_i\cos\theta_i'}{(\cos\theta_i + \eta\cos\theta_i')^2},\ T^\parallel(\theta_i) = \frac{4\eta\cos\theta_i\cos\theta_i'}{(\cos\theta_i' + \eta\cos\theta_i)^2}, \tag{15}$$

in which $\cos\theta_i'$ can be calculated via Eq. (5):

$$\cos\theta_i' = \frac{1}{\eta}\sqrt{\eta^2 - 1 + \cos^2\theta_i}. \tag{16}$$

Thus, $T^\perp, T^\parallel$, and T^+ are functions about $\cos\theta_i$, since $\cos\theta_i'$ in Eq. (15) could be replaced by Eq. (16). In addition, we plot the magnitude curves of the shading term s and T_i^+ regarding $\cos\theta_i$, as shown in Fig. 1.

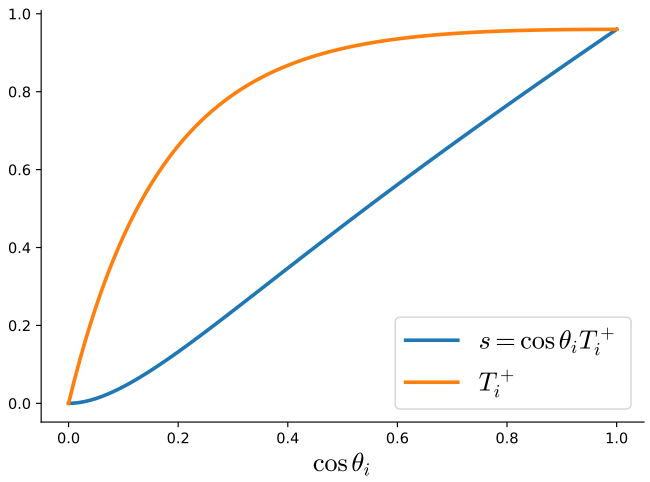


Fig. 1: The magnitude curves of the shading term s and the Fresnel term T_i^+.

2 Applying Our Method under Perspective Projection

In the main paper, we calculate initial normal $\mathbf{n}_{\text{init}}$ and optimize $\mathbf{n}_{\text{w/lt}}$ under the orthographic assumption. Our method can be applied under the perspective projection model without retraining the network. As suggested, we verify the accuracy of our method under different camera models.

Under the perspective model, the orientation of the angle of linear polarization (AoLP) is parallel to the line where xOy and the plane of exitance (PoE) intersect, as shown in Fig. 2. We denote the AoLP as ϕ_p, which is given by

$$[\cos\phi_p, \sin\phi_p, 0]^\top = \frac{\mathbf{z}\times(\mathbf{n}\times\mathbf{v})}{||\mathbf{z}\times(\mathbf{n}\times\mathbf{v})||}, \tag{17}$$

where $\mathbf{z}$ is the direction of the camera optical axis, and we set $\mathbf{z} = [0,0,1]^\top$. The view direction $\mathbf{v}$ is varying over the image plane, which can be obtained by

$$\mathbf{v} = \frac{\mathbf{K}^{-1}\mathbf{x}}{||\mathbf{K}^{-1}\mathbf{x}||}, \tag{18}$$

in which $\mathbf{K}$ denotes the intrinsic matrix of the camera, and $\mathbf{x} = [u, v, 1]^\top$ represents the pixel position on the image plane. To acquire the initial normal $\mathbf{n}_{\text{init}}$, we solve the equations:

$$[\cos\phi_p, \sin\phi_p, 0]^\top = \frac{\mathbf{z}\times(\mathbf{n}\times\mathbf{v})}{||\mathbf{z}\times(\mathbf{n}\times\mathbf{v})||}, \tag{19}$$

$$\mathbf{n}\cdot\mathbf{v} = \sqrt{\frac{\eta^4(1-\rho^2)+2\eta^2(2\rho^2+\rho-1)+\rho^2+2\rho-4\eta^3\rho\sqrt{1-\rho^2}+1}{(\rho+1)^2(\eta^4+1)+2\eta^2(3\rho^2+2\rho-1)}}, \tag{20}$$

$$||\mathbf{n}||_2^2 = 1. \tag{21}$$

Eq. (20) is derived from Eq. (3) of the main paper. There are two possible solutions for the three equations regarding the normal vector. We derive the closed-form solutions of the three equations by Wolfram Mathematica, which are denoted as $\mathbf{n}_{\text{p-init1}}$ and $\mathbf{n}_{\text{p-init2}}$. To resolve the normal ambiguity and obtain $\mathbf{n}_{\text{w/lt}}$, we could resort to minimizing

$$\hat{\theta}_o, \hat{\phi}_o = \underset{\theta_o,\phi_o}{\operatorname{argmin}} \left| s(\theta_o, \phi_o) - \frac{\mathbf{s}_o(0)}{aeT_o^+} \right|, \tag{22}$$

which is a slight variant of Eq. (12) in the main paper. Under the perspective projection, there is an ambiguity in the normal zenith angle as well as in the azimuth angle, which is the only difference between Eq. (22) and Eq. (12) in the main paper. It is easy to optimize Eq. (22) by respectively introducing $\mathbf{n}_{\text{p-init1}}$ and $\mathbf{n}_{\text{p-init2}}$ into the objective function and comparing the values.

Then we evaluate how the camera model contributes to the accuracy of the shape normal estimation. First, we conduct camera calibration by Camera Calibrator in MATLAB to get camera focal length, then calculate the intrinsic matrix $\mathbf{K}$, and compute normal maps estimated under the perspective projection model. We additionally take polarization images of 3 objects under 6 different light directions and generate their corresponding ground truth normal for quantitative evaluation.[1] The results of quantitative evaluation are listed in Table 1 and qualitative results are displayed in Fig. 3. As shown in Table 1, we find that our method still performs well under the perspective camera model without retraining the network. For $\mathbf{n}_{\text{w/lt}}$, the perspective model greatly improves the prediction accuracy but lags in MAngE and ARMSE. Also, using the perspective model marginally improves the final normal maps. However, applying the perspective model on SfP requires additional camera calibration to get $\mathbf{K}$ and object position in the image, which may restrict its application scope. Utilizing the orthographic projection model also generates comparable results, which indicates the validation of the orthographic assumption in our camera setup. In brief, these two camera models excel in different application scenarios, and our proposed framework can be easily adjusted to both models.

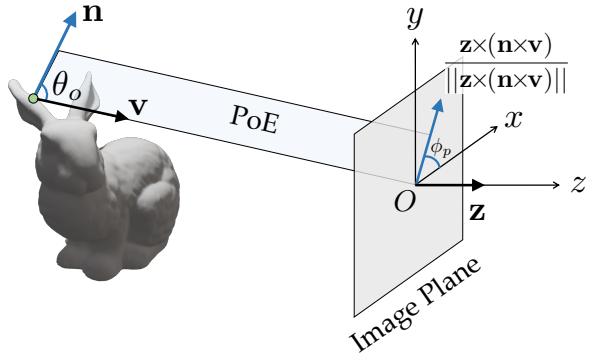


Fig. 2: Illustration of the perspective projection model.

TABLE 1: Quantitative evaluation of our method using orthographic projection and perspective projection, conducted on the newly collected real-world data.

Methods	Angular error ↓			Accuracy ↑		
	Mean	Median	ARMSE	11.25°	22.5°	30.0°
$\mathbf{n}_{\text{init}}$-orthographic	51.77	42.50	64.29	.1251	.3134	.3972
$\mathbf{n}_{\text{init}}$-perspective	51.71	43.02	65.32	.1728	.3352	.4057
$\mathbf{n}_{\text{w/lt}}$-orthographic	**37.99**	27.29	**48.38**	.1361	.4245	.5515
$\mathbf{n}_{\text{w/lt}}$-perspective	38.91	**26.12**	51.53	**.2118**	**.4564**	**.5548**
Ours-orthographic	17.76	14.51	22.47	.3702	.7668	.8781
Ours-perspective	**16.75**	**13.36**	**21.77**	**.4201**	**.7841**	**.8836**

3 Effectiveness of Normal Optimization Module

To further validate the contribution and generalization ability of the normal optimization module, we retrain two learning-based SfP methods (DeepSfP [7] and SfP-wild [8]) with $\mathbf{n}_{\text{w/lt}}$ as the additional input on our synthetic data. Note that the same training strategies are applied in the training process as proposed in their original papers. We test the two retrained models on the synthetic test dataset, SONY dataset [5], and our real dataset, and the quantitative results are shown in Table 2. The quantitative results demonstrate that simply taking the optimized normal as input also

1. SONY dataset [5] and data released by Deschaintre *et al.* [6] cannot be used for the evaluation, since the camera calibration information is not provided.

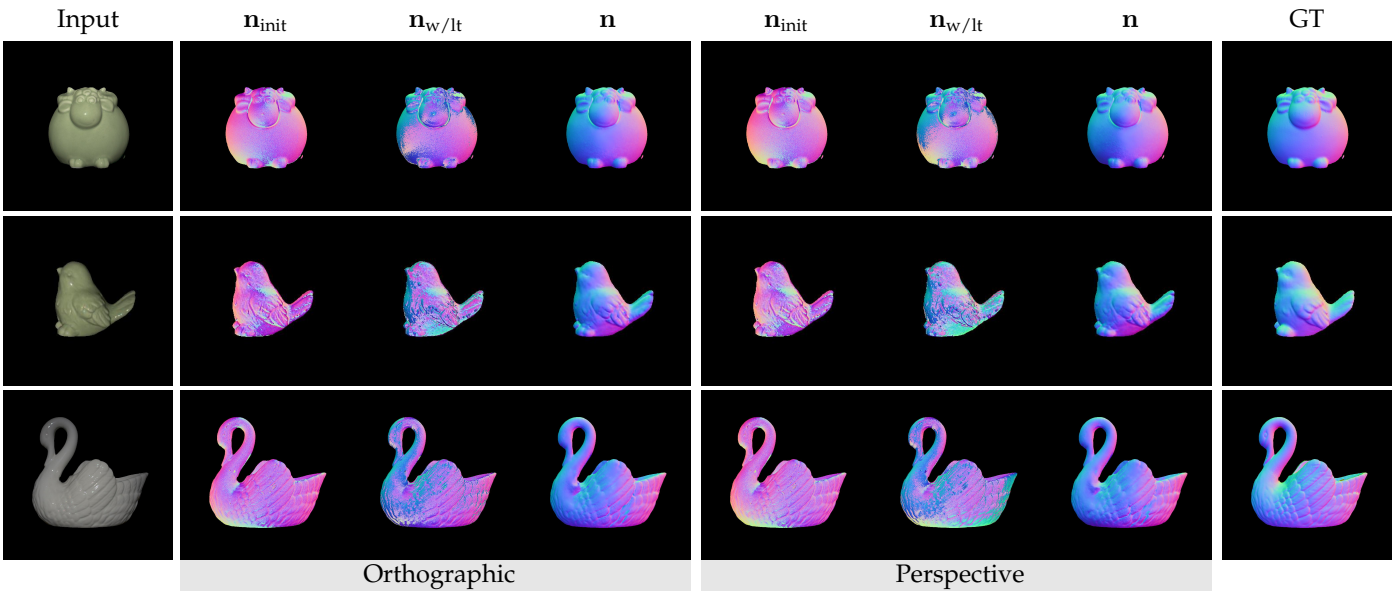


Fig. 3: Qualitative results of our method using orthographic projection and perspective projection, evaluated on the newly collected real-world data.

boosts network performance, which also indicates the effectiveness of light estimation and our normal optimization algorithm.

TABLE 2: Quantitative evaluation on the synthetic data, SONY [5], and our real-world data, compared to the state-of-the-art SfP methods retrained with $\mathbf{n}_{\text{w/lt}}$ as additional input.

	Methods	Angular error ↓			Accuracy ↑		
		Mean	Median	ARMSE	11.25°	22.5°	30.0°
Sync.	DeepSfP [7]	12.09	9.178	17.31	.6543	.8903	.9324
	DeepSfP w/ $\mathbf{n}_{\text{w/lt}}$	8.872	7.545	11.75	.7728	.9517	.9752
	SfP-wild [8]	5.690	3.927	10.59	.9337	.9744	.9793
	SfP-wild w/ $\mathbf{n}_{\text{w/lt}}$	5.436	**3.820**	10.24	**.9361**	.9762	.9809
	Ours	**4.924**	3.981	**6.573**	.9294	**.9873**	**.9940**
SONY [5]	DeepSfP [7]	11.26	8.328	15.65	.6887	.8868	.9334
	DeepSfP w/ $\mathbf{n}_{\text{w/lt}}$	10.62	7.441	15.62	.7321	.8980	.9373
	SfP-wild [8]	10.49	8.126	14.29	.7427	.9153	.9488
	SfP-wild w/ $\mathbf{n}_{\text{w/lt}}$	9.480	6.423	13.89	.7737	.9163	.9490
	Ours	**8.868**	**5.682**	**13.32**	**.7853**	**.9181**	**.9505**
Our Real	DeepSfP [7]	16.88	13.16	22.07	.4082	.7985	.8846
	DeepSfP w/ $\mathbf{n}_{\text{w/lt}}$	16.21	12.69	21.22	.4374	.8104	.8926
	SfP-wild [8]	17.62	14.47	21.47	.3144	.7934	.8900
	SfP-wild w/ $\mathbf{n}_{\text{w/lt}}$	17.28	13.71	22.01	.3907	.7870	.8854
	Ours	**15.58**	**12.65**	**19.64**	**.4387**	**.8275**	**.9111**

4 Further Analysis on Our Model and SfP Methods

4.1 Influence of Image Format

To validate how RGB/grayscale affects the performance, we generate 7,800 sets of test data in grayscale for quantitative evaluation. The results evaluated on the synthetic data are listed in Table 3. The estimated normal on the RGB data are slightly better than (no more than 0.2°) the results on the grayscale data, which may result from the fact that our model is trained on RGB data and color information probably benefit normal recovery. To further verify performance on the real data, we set up a two-camera rig, placing a grayscale polarization camera (Lucid Phoenix 5.0 MP[2] with Sony IMX250MZR CMOS) and an RGB polarization camera (Lucid Triton TRI050S-QC[3] with Sony IMX250MYR CMOS) side by side that reduces the baseline to the minimum. The two cameras are equipped with lens of 16mm focal length, and we simultaneously take images of the same object with the same exposure time and gain. We capture in total 32 sets of grayscale/RGB polarization images of two objects, and each object is illuminated under eight different distant lights. For a fair comparison, we save grayscale/RGB images in 8-bit format. Also, we acquire the ground truth normal for quantitative evaluation. Quantitative evaluation results are listed in Table 3 and visual results are shown in Fig. 4. The advantage of using grayscale or RGB data does not reach a consensus for this test. Our method still performs better on DUCK (Fig. 4 right) by taking RGB images, while the results on SPARROW (Fig. 4 left) estimated from grayscale data are better. We cannot conclude for now whether using RGB data could consistently benefit normal estimation, because the performance differences here are related to two issues that favor either grayscale or RGB images: 1) The noise of Bayer color filter and information loss in the demosaicing may influence the results (grayscale images have higher SNR, which might improve the accuracy of SfP). 2) RGB images do encode richer information in spectrum than the grayscale ones [9], but the spectral polarization properties are not explicitly considered in our current model design (taking such an advantage is beyond the scope of our paper). It is also an intriguing research direction to exploit spectral information in SfP.

4.2 Analysis on the Degenerated Cases

Our model, as well as other SfP methods, fails to produce good normal on CAT, CUP1, and DOLL of SONY [10] dataset. Thus, we show more immediate normal results and polarization properties in Fig. 5 to illustrate this issue. DoLP

2. https://thinklucid.com/product/phoenix-5-0-mp-polarized-model/ 3. https://thinklucid.com/product/triton-5-mp-polarization-camera/

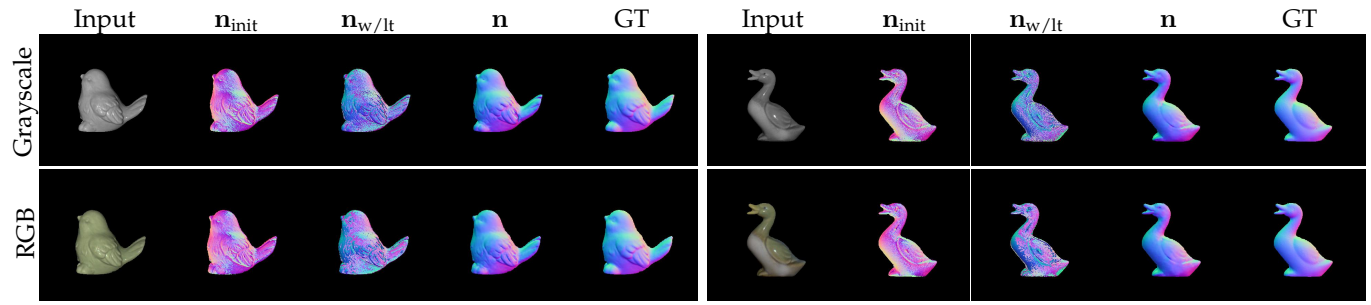


Fig. 4: Qualitative evaluation of our method on the RGB/grayscale polarization images.

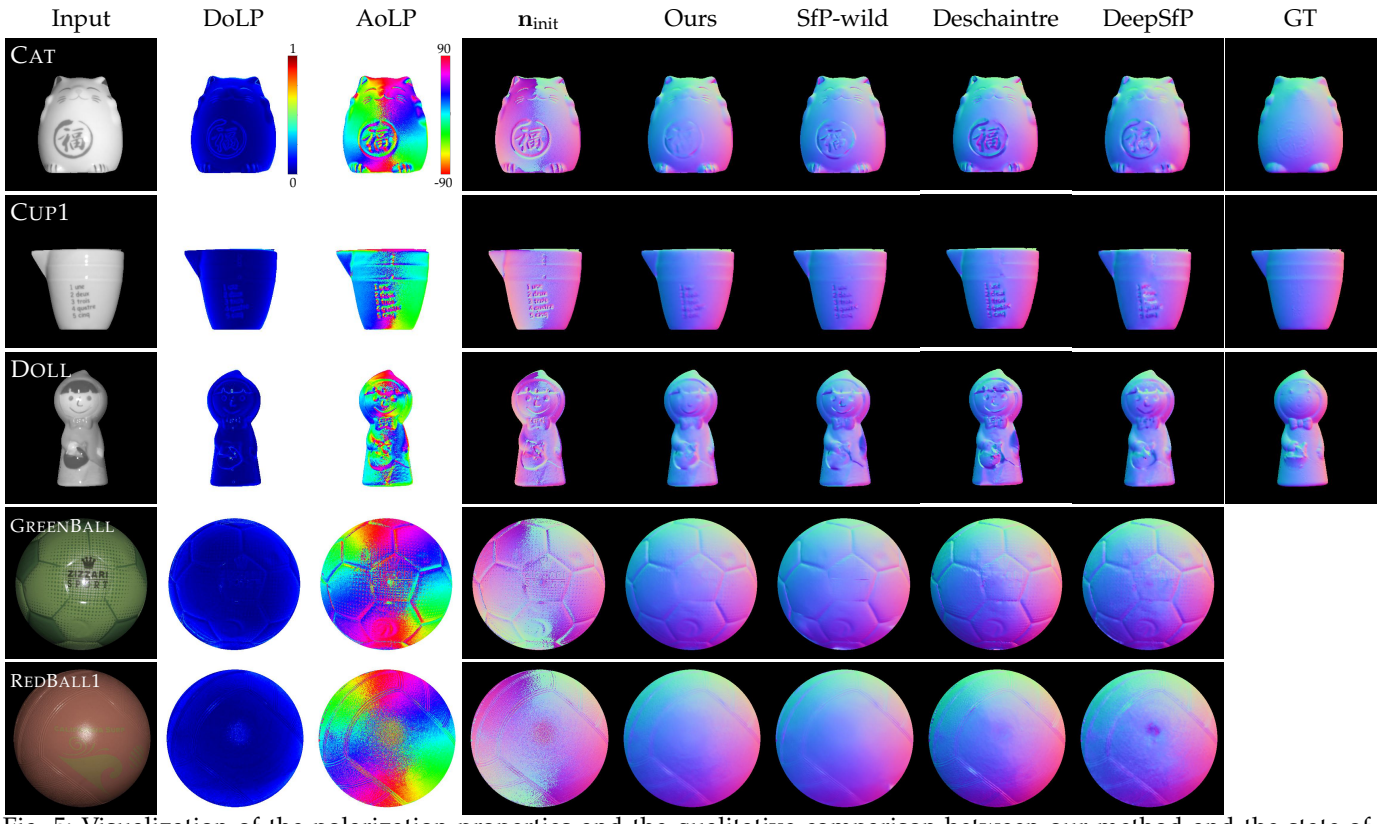


Fig. 5: Visualization of the polarization properties and the qualitative comparison between our method and the state-of-the-art approaches, SfP-wild [8], Deschaintre *et al.* [6], and DeepSfP [7].

TABLE 3: Quantitative evaluation of our method conducted on the RGB/grayscale polarization images.

Data	Format	Angular error ↓			Accuracy ↑		
		Mean	Median	ARMSE	11.25°	22.5°	30.0°
Sync	Grayscale	5.090	4.116	6.760	.9229	.9862	.9935
	RGB	**4.924**	**3.981**	**6.573**	**.9294**	**.9873**	**.9940**
SPARROW	Grayscale	**14.00**	**10.65**	**18.02**	**.5300**	.8306	.9153
	RGB	14.92	12.59	18.34	.4267	**.8481**	**.9276**
DUCK	Grayscale	18.39	17.58	20.33	.1921	.7351	.9143
	RGB	**16.45**	**15.23**	**18.57**	**.2902**	**.7983**	**.9311**

and AoLP maps generated from the polarization images should provide cues about the surface normal and are independent from the albedo according to the polarimetric BRDF (Eq. (7) in the main paper). However, we find DoLP and AoLP on the CAT belly, the text of CUP1, and the handbag of DOLL simultaneously change with the albedo. We think the reason for the abrupt change of polarization properties lies in the different BRDFs in regions with black/white albedo. The varying BRDF may be caused by the material of the black paints on the object. Since the basic polarization cues are affected, the computed $\mathbf{n}_{\text{init}}$ also suffers from artifacts in these regions. That is why all of the polarization-based methods [6], [7], [8] hardly produce satisfactory normal on part of regions of CAT, CUP1, and DOLL. Though our synthetic dataset consists of spatially-varying albedo, the data generation pipeline cannot simulate this kind of BRDF change. It is also difficult for our method to output accurate normal on these samples. The polarization properties of REDBALL1 and GREENBALL are independent of the varying albedo, so our model could handle such albedo, as shown in Fig. 13 and Fig. 15.

4.3 Test SfP Methods on Challenging Materials

We conduct quantitative evaluation of the SfP methods on three challenging materials, *i.e.*, fabric, rough clay, and

brass. The quantitative results are listed in Table 4 and visual comparison is displayed in Fig. 6. From the DoLP and AoLP maps, we could see that fabric (MONKEY) and the rough clay (BUDDHA1) depolarize most of the exitant light and make the azimuth angle cues lost in the AoLP map; polarization state of the light reflected from BUDDHA2 seems to be noisy and unstable since the refractive index of brass is a complex number and the pBRDF of the conductor is completely different from that of the dielectric. The contaminated polarization information makes all of the SfP methods fail to produce reliable normal maps on these objects.

5 Network Details

The detailed network architectures of pLTNet and pNENet are illustrated in Fig. 7 and Fig. 8, respectively. pLTNet is designed for predicting lighting conditions and object texture information, which takes in four polarization images and a mask and predicts albedo, roughness, and the direction and intensity of the distant light. The feature extractor consists of five down-sampling ResNet blocks [11] and a feature output block. In each block, the down-sampling layer has a convolutional filter with a 3×3 kernel size and a stride of 2, and the vanilla convolutional layer has a filter with a kernel size of 3 and a stride of 1. We use the Leaky ReLU (α=0.1) as the activation function and instance normalization in the encoder blocks. The features then are processed by the feature output block, which consists of two standard 3×3 convolutional layers with a stride of 1. To predict the light direction and intensities, we conduct Adapted Max Pooling on the extracted feature map to produce a 512-dimensional latent vector. Then we use the two-layer classifier to regress the light direction vector and output categories of intensity values. For estimating the albedo and roughness, we adopt five up-sampling blocks in the decoder, each of which has a skip connection to the encoder at the same hierarchical level. Fed with the decoded features, the image output layer produces the texture maps of the object.

pNENet is fed with the optimized normal, four polarization images, albedo, and mask, which produces refined normal maps as the final results. We employ a widely adopted encoder-decoder architecture to build pNENet. The encoder consists of five ResNet blocks [11] and a feature output block, which takes a down-sampling convolutional filter with a kernel size of 3 and a stride of 2, and a convolutional layer with a 3×3 kernel size and a stride of 1. Also, the Leaky ReLU activation (α=0.1) and instance normalization are used between the two layers. Corresponding to the encoder blocks, the decoder has five ResNet blocks as well, and each block consists of a bilinear interpolation module and two convolutional layers with a kernel size of 3 and a stride of 1. We use the SPADE [7], [12] blocks between two layers and adopt the bilinear interpolation to interpolate the feature maps instead of the transposed convolution. To preserve the high-frequency information, we connect the encoder and the decoder blocks at the same hierarchical level with skip connections. Finally, the output layer takes the preceding features and produces the estimated normal maps.

6 Additional Qualitative Comparisons on Real Data

In this section, we first show the additional results of SfP methods on SONY dataset [5] (additionally including BEAR1, CUP1, EGG, CAT, and SQUIRREL compared to Fig. 8 of the main paper) and our real-world dataset in Fig. 9. More intermediate results (including albedo, roughness, and light directions) of our method compared to Deschaintre *et al*. [6] on our real dataset are displayed in Fig. 10. Table 5 contains quantitative results on each object of SONY [5]. Also, we qualitatively evaluate our method against the state-of-the-art approaches on the rest real data released by Deschaintre *et al*. [17], as shown in Fig. 11, Fig. 12, Fig. 13, Fig. 14, and Fig. 15.

7 Synthetic Data Generation

The data generation pipeline is elaborated in Sec. 5.2 of the main paper. In this section, we show an example of our simulated polarization images, the corresponding polarization properties, *i.e.*, DoLP and AoLP, and the computed initial normal $\mathbf{n}_{\text{init}}$ as shown in Fig. 16. The visual comparisons between the real captured image EGG from SONY dataset [5] and the simulated data demonstrate the validity of our data simulation method.

References

[1] E. Hecht, *Optics*. Pearson education, 2002.
[2] E. Collett, *Field Guide to Polarization*, ser. Field Guides. Society of Photo Optical, 2005.
[3] S.-H. Baek, D. S. Jeon, X. Tong, and M. H. Kim, "Simultaneous acquisition of polarimetric SVBRDF and normals." *ACM Transactions on Graphics*, 2018.
[4] B. Walter, S. R. Marschner, H. Li, and K. E. Torrance, "Microfacet models for refraction through rough surfaces," in *Proc. Eurographics Conference on Rendering Techniques*, 2007.
[5] T. T. Ngo, H. Nagahara, and R. ichiro Taniguchi, "Surface normals and light directions from shading and polarization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
[6] V. Deschaintre, Y. Lin, and A. Ghosh, "Deep polarization imaging for 3D shape and SVBRDF acquisition," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2021.
[7] Y. Ba, A. Gilbert, F. Wang, J. Yang, R. Chen, Y. Wang, L. Yan, B. Shi, and A. Kadambi, "Deep shape from polarization," in *Proc. European Conference on Computer Vision*, 2019.
[8] C. Lei, C. Qi, J. Xie, N. Fan, V. Koltun, and Q. Chen, "Shape from polarization for complex scenes in the wild," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2022.
[9] C. P. Huynh, A. Robles-Kelly, and E. R. Hancock, "Shape and refractive index from single-view spectro-polarimetric images," *Springer International Journal of Computer Vision*, 2013.
[10] T. Ngo Thanh, H. Nagahara, and R.-i. Taniguchi, "Shape and light directions from shading and polarization," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2015.
[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2016.
[12] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. Conference on Computer Vision and Pattern Recognition*, 2019.
[13] Z. Li, Z. Xu, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker, "Learning to reconstruct shape and spatially-varying reflectance from a single image," *ACM Transactions on Graphics*, 2018.
[14] A. H. Mahmoud, M. T. El-Melegy, and A. A. Farag, "Direct method for shape recovery from polarization and shading," in *Proc. International Conference on Image Processing*, 2012.
[15] D. Miyazaki, M. Kagesawa, and K. Ikeuchi, "Polarization-based transparent surface modeling from two views," in *Proc. International Conference on Computer Vision*, 2003.

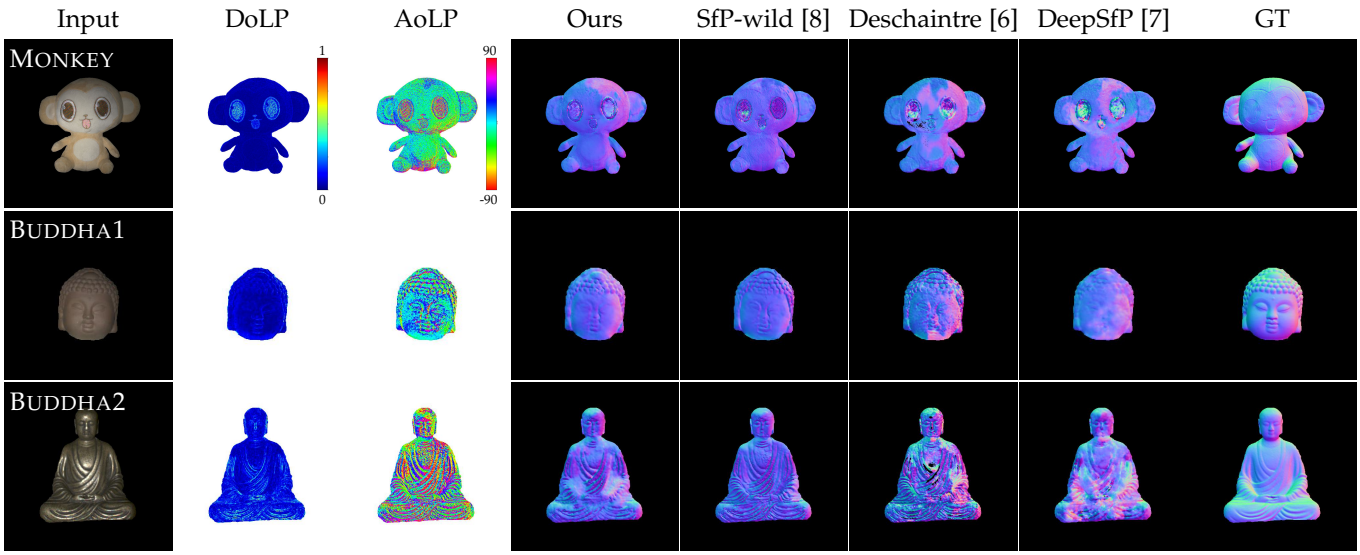


Fig. 6: Qualitative results and polarization properties of the objects made from fabric, rough clay, and brass.

TABLE 4: Quantitative evaluation on the objects made from fabric, rough clay, and brass, compared with the state-of-the-art methods. MAngE is adopted to measure the estimated normal maps.

Method	MONKEY-FABRIC	BUDDHA1-CLAY	BUDDHA2-BRASS	Avg.
DeepSfP [7]	36.01	30.64	42.81	36.49
Deschaintre [6]	35.77	38.00	52.00	41.92
SfP-wild [8]	35.13	41.96	50.37	42.49
Ours	34.28	34.80	42.05	37.04

TABLE 5: Quantitative evaluation on SONY dataset [5] compared with the state-of-the-art methods. MAngE is adopted to measure the estimated normal maps.

Method	BEAR1	BEAR2	CAT	CUP1	CUP2	DOLL	EGG	SQUIRREL	Avg.
Miyazaki [15]	36.71	42.58	46.886	36.51	37.08	44.99	32.46	36.43	39.21
Mahmoud [14]	37.41	42.07	43.48	45.62	46.79	49.32	43.24	37.88	43.22
Smith [16]	36.43	42.98	45.44	38.12	36.39	44.35	47.35	33.09	40.52
Li [13]	30.17	27.09	22.06	17.73	16.72	22.02	18.93	24.01	22.34
DeepSfP [7]	14.31	10.72	10.58	9.543	8.653	16.21	7.086	13	11.26
Deschaintre [6]	18.26	11.44	12.34	8.63	7.287	20.32	6.267	15.02	12.45
SfP-wild [8]	13.52	10.05	9.612	7.855	7.66	15.32	7.026	12.9	10.49
Ours	11.63	9.059	7.873	6.43	5.586	13.6	5.617	11.15	**8.868**

[16] W. A. Smith, R. Ramamoorthi, and S. Tozza, "Height-from-polarisation with unknown lighting or albedo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.
[17] V. Deschaintre, M. Aittala, F. Durand, G. Drettakis, and A. Bousseau, "Single-image SVBRDF capture with a rendering-aware deep network," *ACM Transactions on Graphics*, 2018.

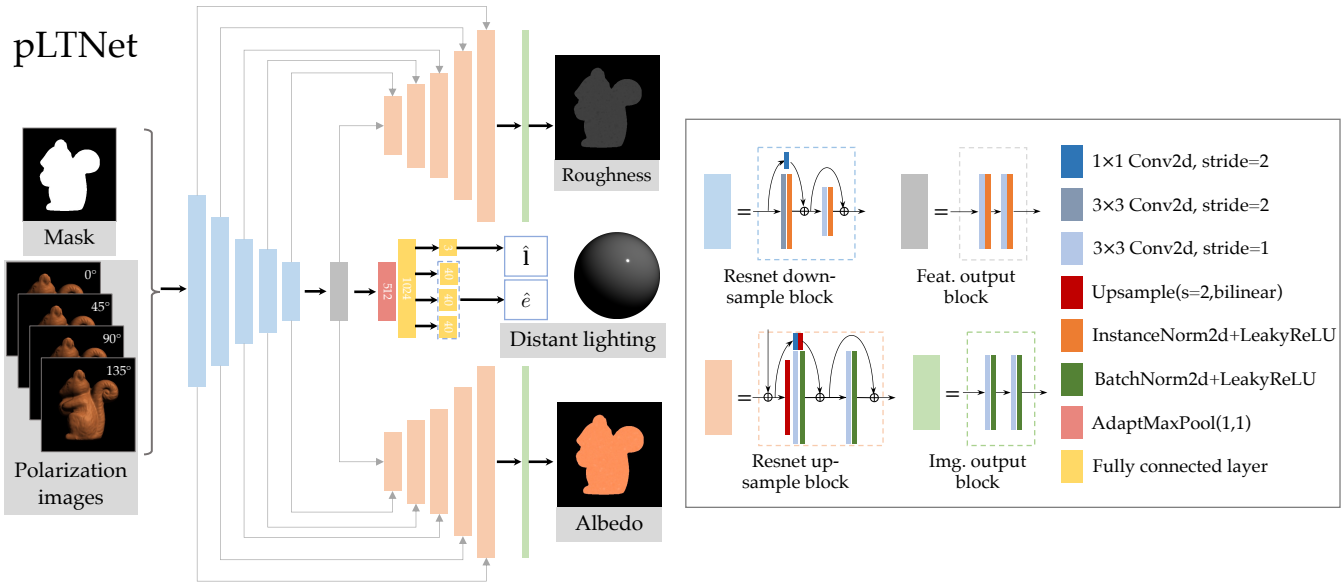


Fig. 7: Network architecture of pLTNet.

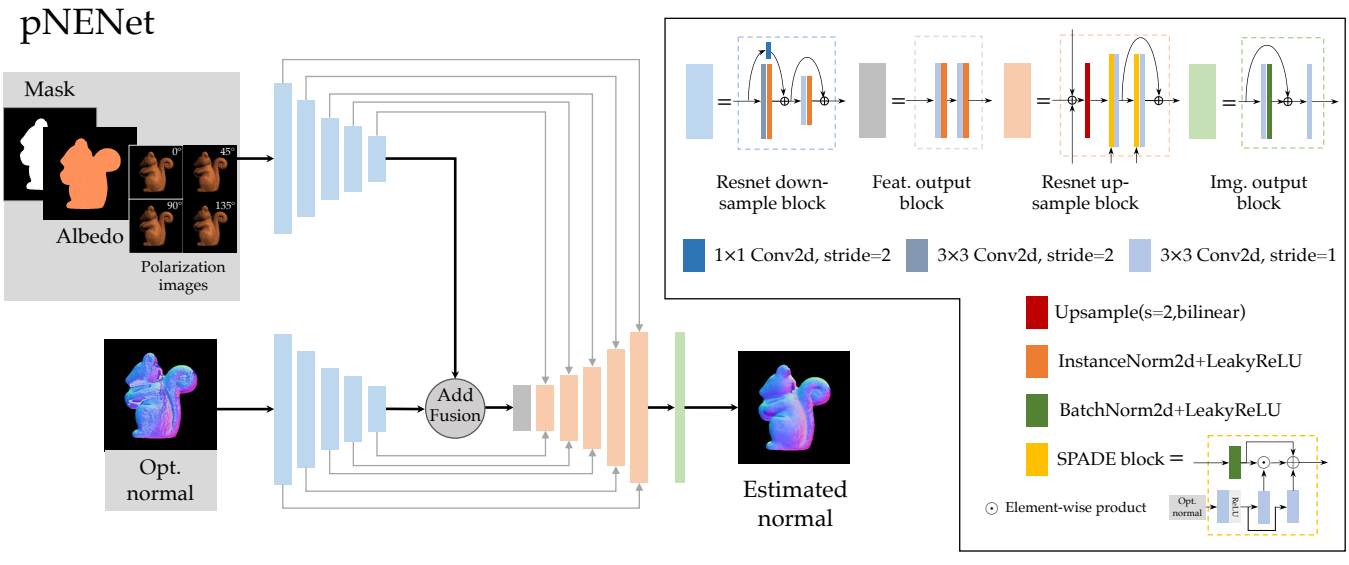


Fig. 8: Network architecture of pNENet.

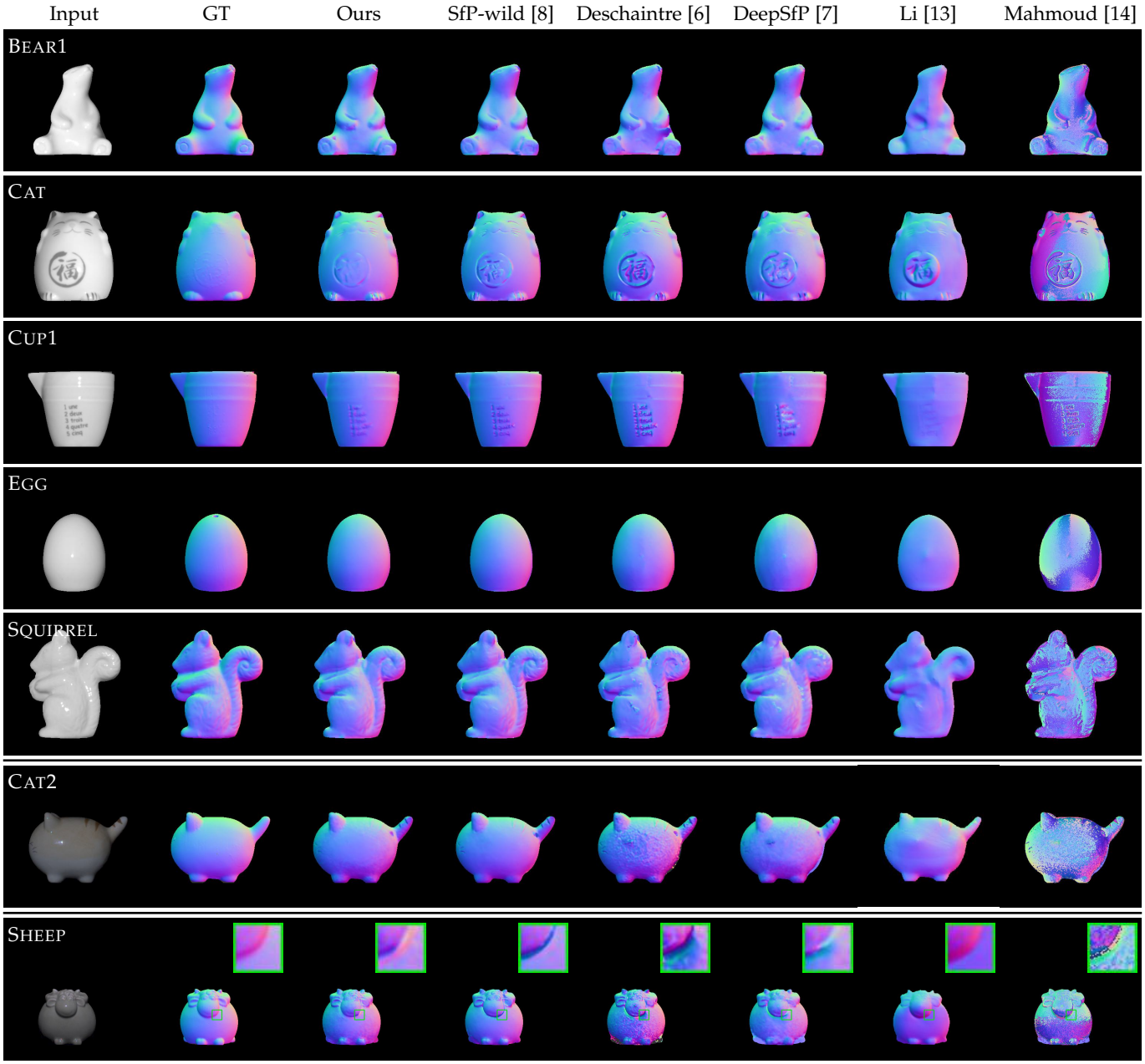


Fig. 9: Additional qualitative comparisons among the state-of-the-art approaches, SfP-wild [8], Deschaintre *et al*. [6], DeepSfP [7], Li *et al*. [13], and Mahmoud *et al*. [14], evaluated on the rest of the objects of SONY dataset [5] and our real dataset.

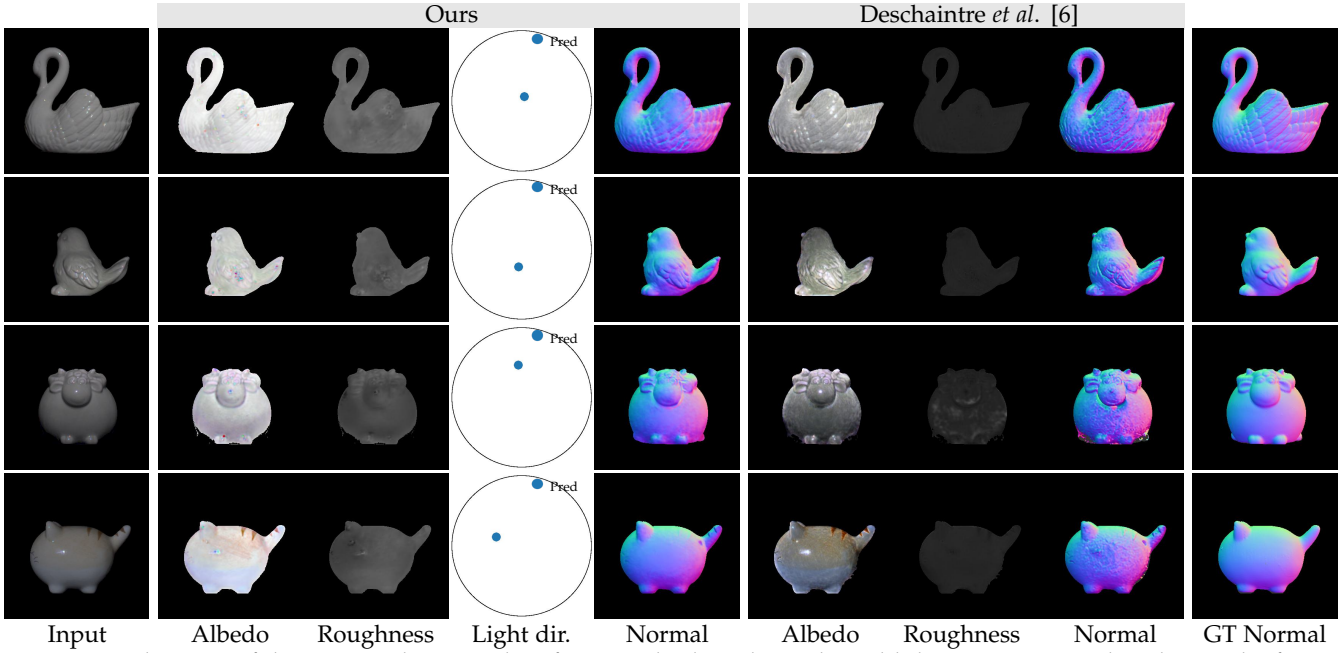


Fig. 10: Visualization of the intermediate results of our method on the real-world dataset, compared to the results from the SVBRDF approach Deschaintre *et al*. [6].

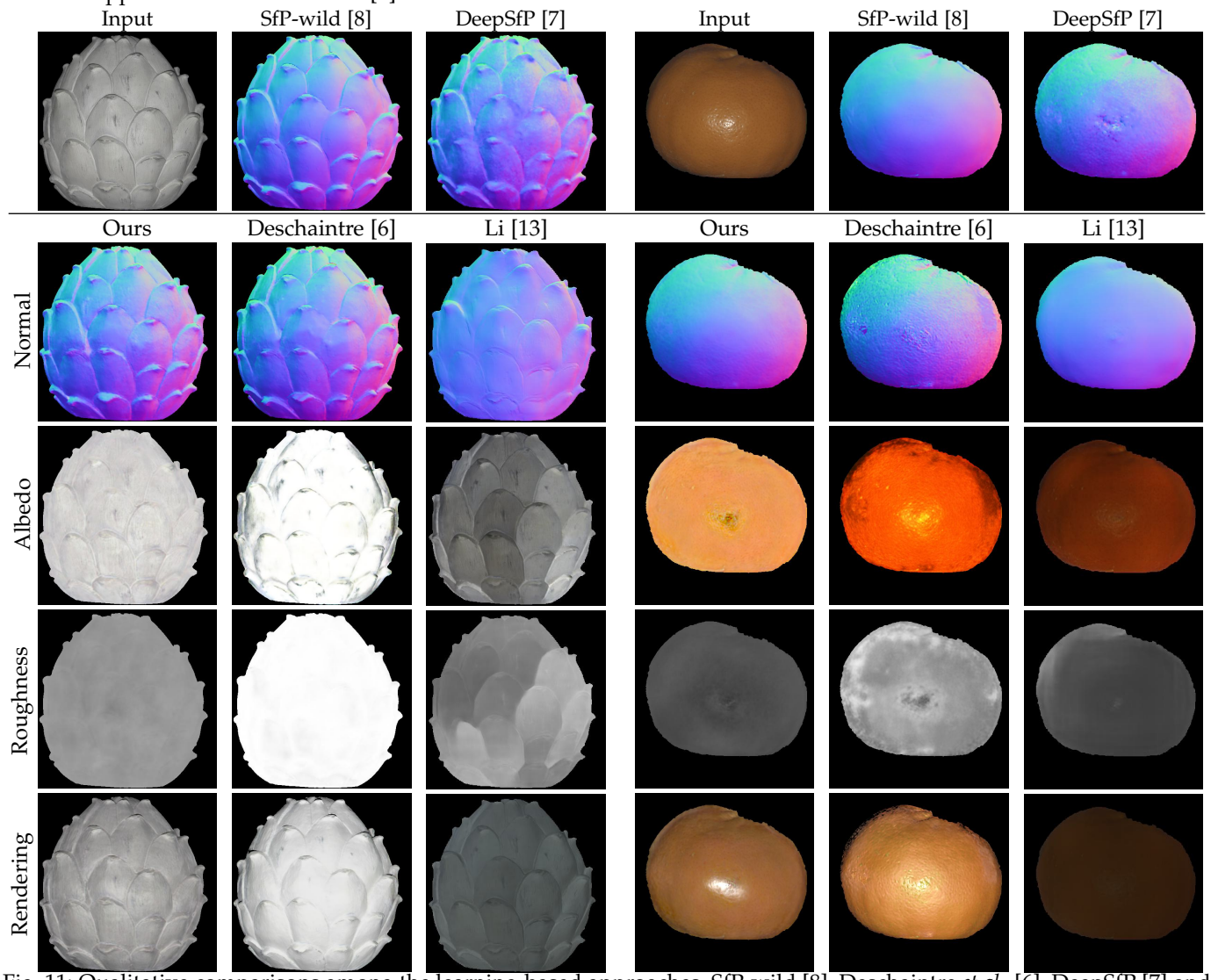


Fig. 11: Qualitative comparisons among the learning-based approaches, SfP-wild [8], Deschaintre *et al*. [6], DeepSfP [7] and Li *et al*. [13], evaluated on the real data, ARTICHOKE and CLEMENTINE, released by Deschaintre *et al*. [6].

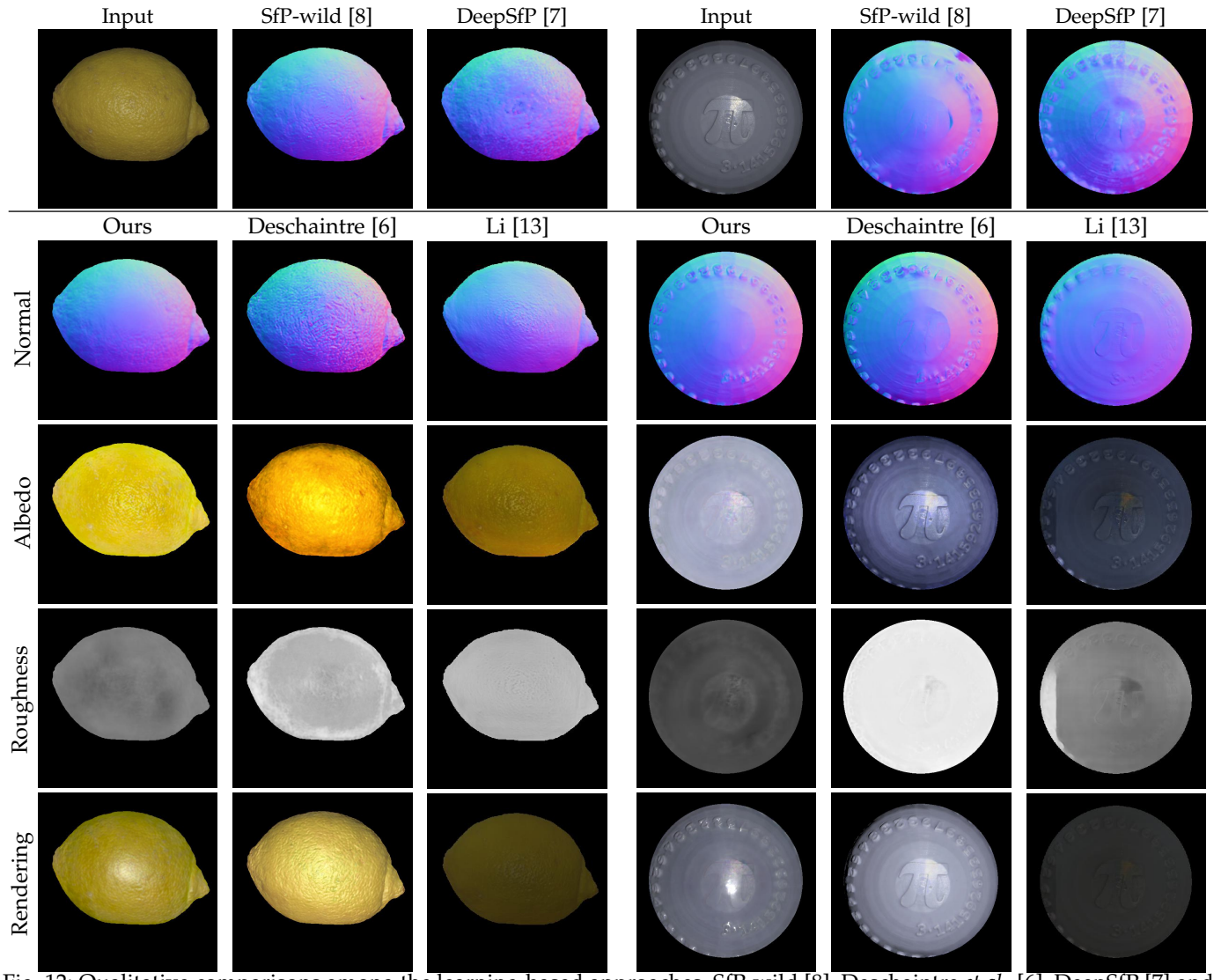


Fig. 12: Qualitative comparisons among the learning-based approaches, SfP-wild [8], Deschaintre *et al*. [6], DeepSfP [7] and Li *et al*. [13], evaluated on the real data, LEMON and PIBALL, released by Deschaintre *et al*. [6].

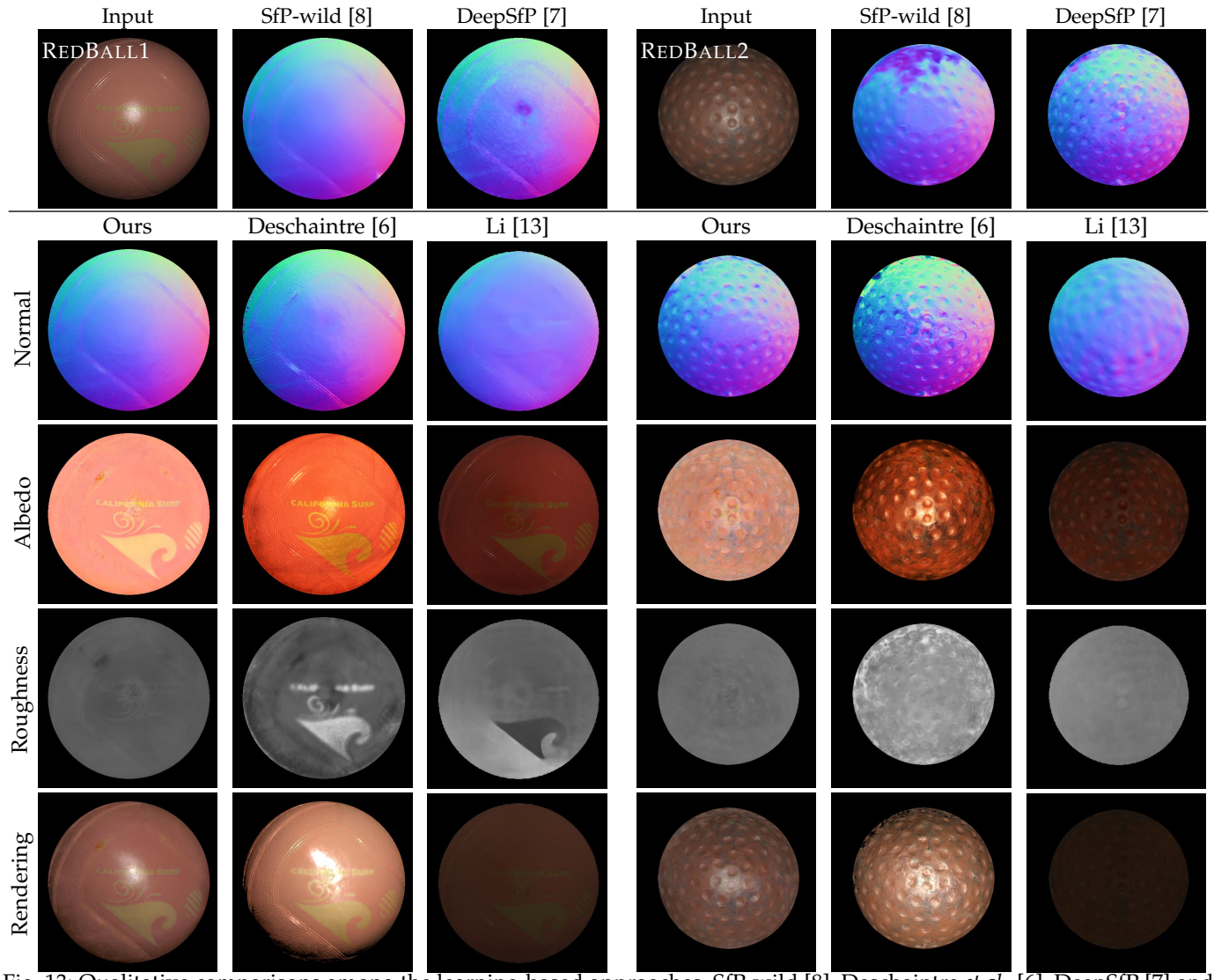


Fig. 13: Qualitative comparisons among the learning-based approaches, SfP-wild [8], Deschaintre *et al.* [6], DeepSfP [7] and Li *et al.* [13], evaluated on the real data, REDBALL1 and REDBALL2, released by Deschaintre *et al.* [6].

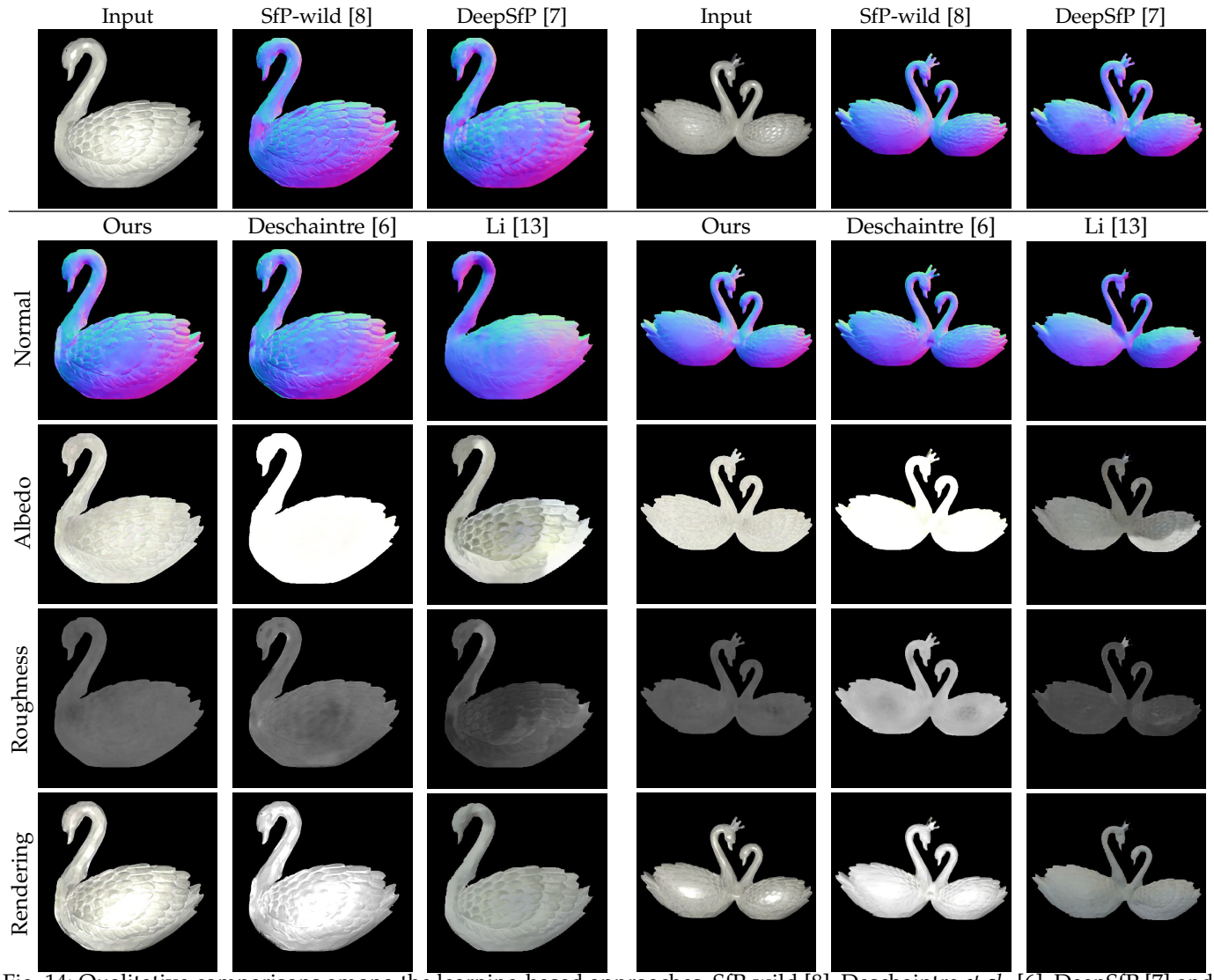


Fig. 14: Qualitative comparisons among the learning-based approaches, SfP-wild [8], Deschaintre *et al*. [6], DeepSfP [7] and Li *et al*. [13], evaluated on the real data, SWAN1 and SWAN2, released by Deschaintre *et al*. [6].

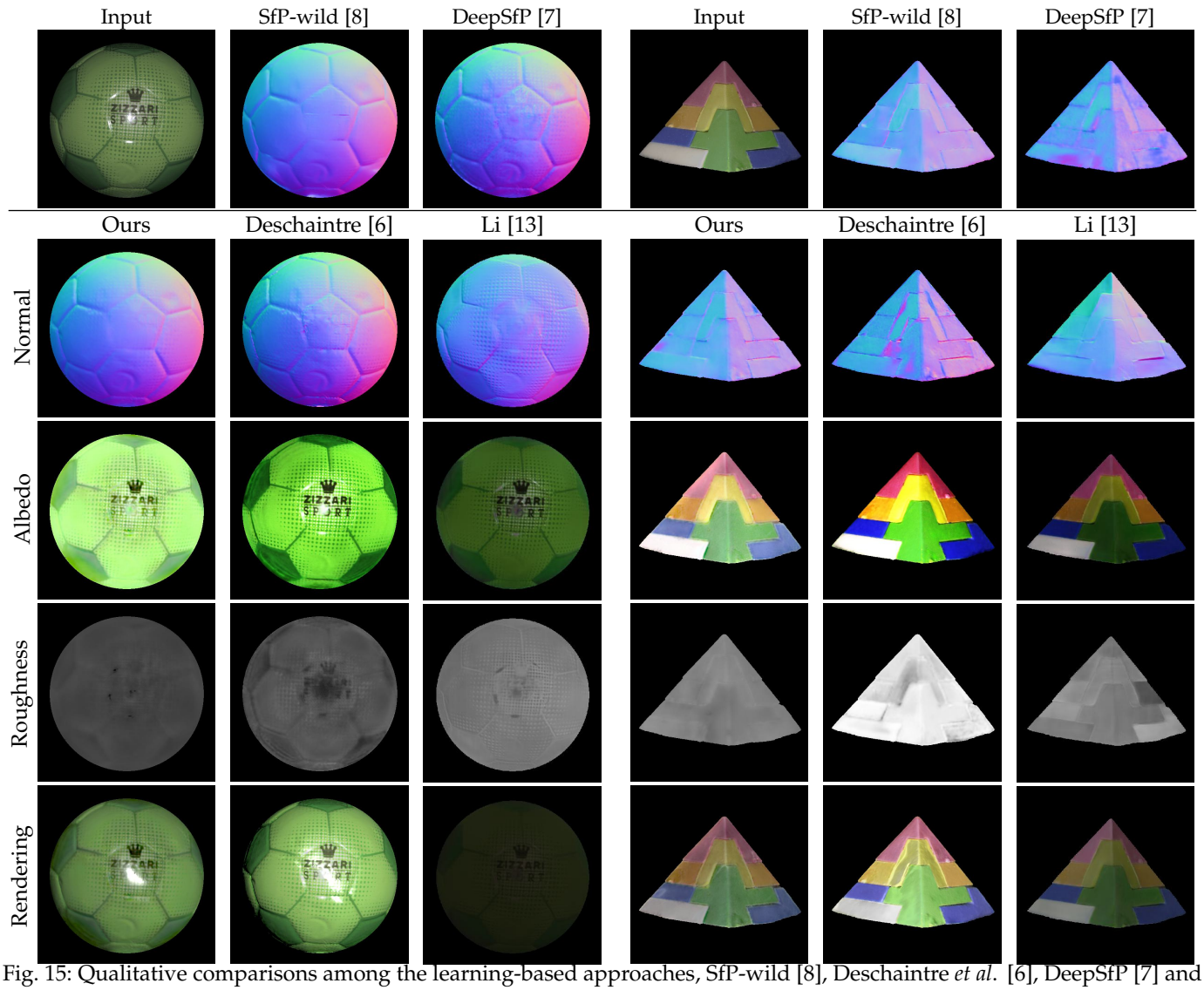


Fig. 15: Qualitative comparisons among the learning-based approaches, SfP-wild [8], Deschaintre *et al.* [6], DeepSfP [7] and Li *et al.* [13], evaluated on the real data, GREENBALL and PYRAMID, released by Deschaintre *et al.* [6].

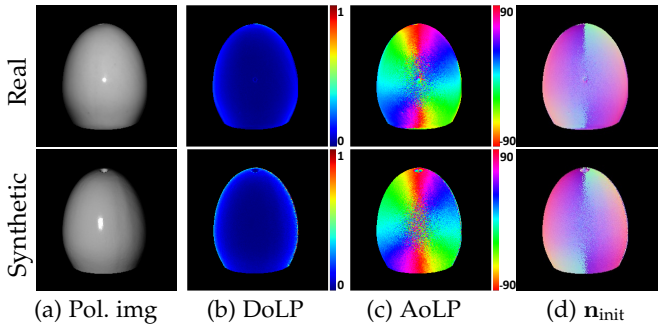


Fig. 16: An example of synthetic data. (a) The rendered polarization image $I(0^\circ)$ compared to the real one from SONY dataset [5]. (b) The DoLP calculated from the polarization images. (c) The AoLP calculated from the polarization images. (d) The initial normal calculated from the corresponding DoLP and AoLP.