Pacific Graphics (2023)
R. Chaine, Z. Deng, and M. H. Kim (Editors)

Short Paper

SS-SfP: Neural Inverse Rendering for Self Supervised Shape from (Mixed) Polarization

Ashish Tiwari and Shanmuganathan Raman

Computer Vision, Imaging and, Graphics Lab, Indian Institute of Technology Gandhinagar, India

Abstract

We present a novel inverse rendering-based framework to estimate the 3D shape (per-pixel surface normals and depth) of objects and scenes from single-view polarization images, the problem popularly known as Shape from Polarization (SfP). The existing physics-based and learning-based methods for SfP perform under certain restrictions, i.e., (a) purely diffuse or purely specular reflections, which are seldom in the real surfaces, (b) availability of the ground truth surface normals for direct supervision that are hard to acquire and are limited by the scanner's resolution, and (c) known refractive index. To overcome these restrictions, we start by learning to separate the partially-polarized diffuse and specular reflection components, which we call reflectance cues, based on a modified polarization reflection model and then estimate shape under mixed polarization through an inverse-rendering based self-supervised deep learning framework called SS-SfP, guided by the polarization data and estimated reflectance cues. Furthermore, we also obtain the refractive index as a non-linear least squares solution. Through extensive quantitative and qualitative evaluation, we establish the efficacy of the proposed framework over simple single-object scenes from DeepSfP dataset and complex in-the-wild scenes from SPW dataset in an entirely self-supervised setting. To the best of our knowledge, this is the first learning-based approach to address SfP under mixed polarization in a completely self-supervised framework. Code will be made publicly available.

CCS Concepts

• ***Computing methodologies*** → *Computer Vision; Image-based Rendering;*

1. Introduction

The polarization state of the light reflected off a surface depends on the shape of the underlying surface, which is the basic principle of the Shape from Polarization (SfP) problem. The prime objective is to estimate surface normals from a set of polarization images. The physics involved in SfP arises from Fresnel Equations [Col05] and is among the most complex problems in addressing computer vision tasks such as 3D reconstruction. Although several attempts have addressed SfP by providing physics-based solutions, researchers have only recently started looking at SfP through the lens of deep learning [BGW*20, LQX*22]. However, since a physics-only or a learning-only solution fails to achieve the best performance individually, Ba *et al.* [BGW*20] propose combining physical priors with a deep network to address the SfP problem. While other photometric methods for 3D reconstruction, such as Photometric Stereo (PS) [Woo80] and Shape from Shading (SfS) [PF05] require known or estimated lighting directions for shape estimation, SfP can be used under completely passive lighting conditions without knowing lighting directions. Consider, for example, a color poster on a wall. Unlike the RGB-based methods, such as PS [Woo80], SfP does not get distracted by mere image semantics (of the poster) to provide erroneous surface normal estimates of the underlying surface (the wall). The polarization cues reflect the changes in the polarization state of the reflected light instead of just the RGB measurement. Further, it applies to a broad category of materials [MCZ*22] such as translucent, transparent, dielectrics, and metals with applications such as image segmentation [SDMF12], robot navigation [BVM17], image enhancement [Sch11], and underwater imaging [Sch15]. The underlying Fresnel equations [Col05] in modeling SfP pose a different set of ambiguities arising due to model mismatch and periodicity in the sinusoidal behaviour of the reflected polarized light intensity with respect to the polarizer angle (detailed in the supplementary). While researchers have used additional information such as shading constraints, surface convexity, and coarse depth maps to address such ambiguities, they have not succeeded sufficiently [AH07, KTSR15]. The requirement of a known refractive index brings another challenge as it introduces *refractive distortion* in the estimated shape if the refractive index is incorrect. The assumption of the diffuse-only or specular-only reflections from surface points is highly restrictive. A broad class of methods (discussed in Section 2) use computational and optical techniques to split the image into specular-only or diffuse-only images. However, in the real world, surfaces are neither purely diffuse nor purely specular but somewhere in-between, i.e., they tend to exhibit mixed

DOI: 10.2312/pg.20231265

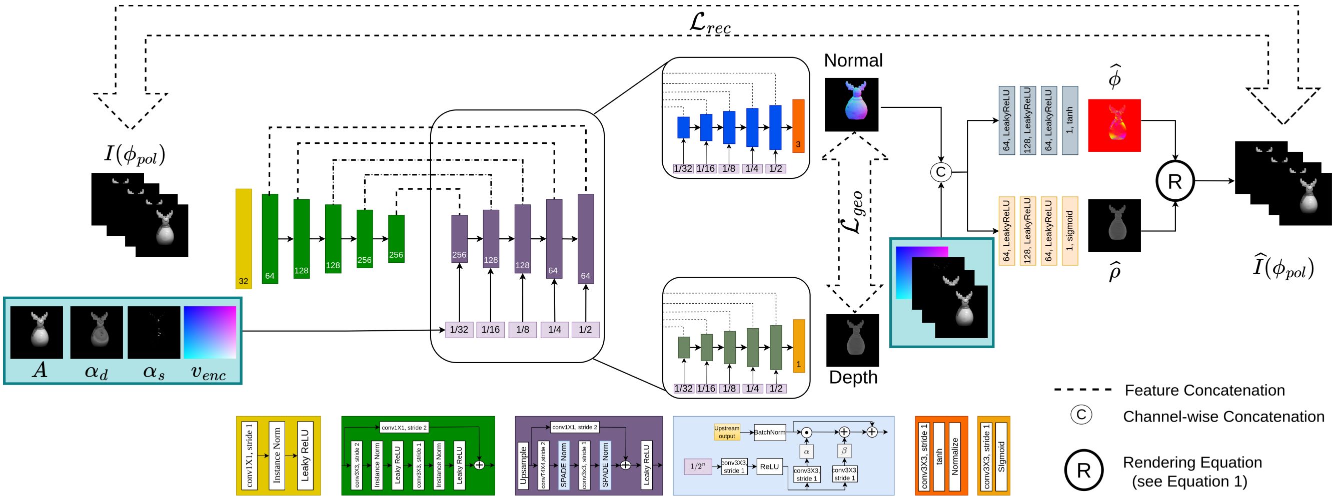

Figure 1: *Layer-wise detailed description of the proposed framework - SS-SfP.*

polarization where the camera receives both diffusely and specularly reflected light. The existing reflectance decomposition techniques (specific to the SfP setting) have primarily relied on image intensity or color information [NZI01, MHP*07, LLK*02, TI08], making them susceptible to artefacts. While some have used polarization cues, they needed active illumination to perform the separation [MHP*07, GFT*11].

Contributions. The following is a summary of the key contributions of the work.

- We analyze the polarization image formation model under mixed reflection and estimate diffuse and specular reflection components at each pixel, calling them the *reflectance cues*, using the least-squares approach.
- We propose a neural inverse rendering based framework, called SS-SfP - **S**elf **S**upervised **S**hape **f**rom (Mixed) **P**olarization, to estimate the shape (per-pixel surface normal and depth) over surfaces exhibiting mixed polarization using the raw polarization images and reflectance cues. Further, we also obtain refractive index through a non-linear least squares formulation.
- To the best of our knowledge, we are the first to address the problem of Shape from (Mixed) Polarization using a deep learning framework under a self-supervised setting.

2. Related Work

In this section, we shall review some of the physics-only and learning-based SfP methods, including methods for reflectance separation.

Shape from Polarization. Several traditional physics-based methods have used a variety of cues such as coarse shading from two views [AH07], depth maps [KTSR15], reciprocal image pairs [DJZ*21], two-view stereo [FKNN21, ZS19], multi-view stereo [CGS*17, MSB*16], or front-flash illumination [DLG21], active lighting [MFSG06] to resolve the underlying ambiguities in SfP. Others have combined photometric stereo [Woo80] and shading information [SRT18, MEMF12] with SfP. Baek *et al.* [BJTK18] choose to perform joint optimization of appearance, normals, and refractive index. Yu *et al.* [YZS17] directly estimate height from the polarization images through nonlinear least squares formulation. Moreover, Mecca *et al.* [MLC17, LMSC19] propose differential level-set-based geometric characterizations to address SfP from two-light polarimetric imaging. Guarnera *et al.* [GPDG12] have addressed SfP under a single spherical incident lighting condition that is either unpolarized or circularly polarized by considering the specularly reflected light. Further, other methods have used deep networks to disambiguate SfP. While Ba *et al.* [BGW*20] train a CNN to obtain normals from polarization over a real-world object-level dataset, Kondo *et al.* [KOS*20] have done the same over a synthetic dataset of polarization images with a new polarimetric BRDF model. Further, Lei *et al.* [LQX*22] study scene-level SfP over complex scenes in the wild. Recently, Muglikar *et al.* [MBMS23] approached SfP through event cameras instead of conventional RGB cameras.

Reflectance separation. Early SfP methods have assumed pure reflection types and materials to constrain the problem. For instance, Rahmann *et al.* [RC01] assume pure specular reflection, and others [MTHI03, Atk17] assume pure diffuse reflection. Other methods perform reflectance separation using only image intensity [NZI01] or color information [MHP*07, LLK*02, TI08] and hence, are susceptible to artefacts. Another set of approaches to separate diffuse and specular components using polarization rely on active illumination such as spherical gradient illumination [MHP*07, GFT*11] to achieve photorealistic reconstructions. Under passive illumination, Nayar *et al.* [NFB97] introduce a separation technique using polarization images and color cues limited by the smoothness assumption. Huynh *et al.* [HRKH10] enforce a generative model on the material dispersion equations to obtain refractive index. Taamazyan *et al.* [TKR16] propose a joint optimization method to estimate shape, perform diffuse-specular separation, and per-pixel refractive indices. Ghosh *et al.* [GCP*10] achieve the same by analyzing Stokes reflectance field of circularly polarized spherical illumination. Recently, Dave *et al.* [DZV22] used a learning-based framework to achieve the same except the refractive index estimation. They additionally estimate the illumination incident on the objects using polarization cues and perform the multi-view 3D reconstruction. However, our work primarily

Proceedings published by Eurographics - The European Association for Computer Graphics.

focuses on single-view reconstruction. Furthermore, Kajiyama *et al.* [KPKO23] achieve reflectance separation based on the polarization and dichromatic reflection models. However, they consider three-channel (RGB) color polarization images leading to twelve radiance values per pixel, unlike our approach that considers single-channel images with just four radiance values per pixel.

3. Shape from Polarization: Background

The polarization data can be obtained by physically rotating a linear polarizer at different angles in front of the camera. Nowadays, it is achieved through commercially available polarization cameras (such as PHX05XS-P) at four angles $\phi_{pol} = \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$ in a single shot. The irradiance $I(\phi_{pol})$ measured by the polarization camera after an incident unpolarized light is reflected from a single scene point shows a sinusoidal variation with respect to the rotation angle of the polarizer, as described in Equation 1.

$$I(\phi_{pol}) = \mathrm{A} + \mathrm{B}\cos(2\phi_{pol} - 2\phi) \tag{1}$$

Here, $A = \frac{I_{max}+I_{min}}{2}$ and $B = \frac{I_{max}-I_{min}}{2}$ with I_{max} and I_{min} being the maximum and minimum intensity values (respectively) obtained by rotating the polarizer at different angles. One can easily think of A as the DC component and B as the amplitude of the reflected polarized light. Three physical quantities - the unpolarized light intensity (A), the phase angle (ϕ), and the degree of polarization $(\rho = \frac{B}{A})$ describe entirely the polarization state of the light. The three unknowns (A, ρ, ϕ) can be obtained by sampling $I(\phi_{pol})$ at minimum three different values of ϕ_{pol}. We delve deeper into this model to discuss a more general case for mixed polarization in Section 4. The surface normal parametrized by the azimuth angle (φ) and the zenith angle (θ) at any surface point is given as follows.

$$\mathbf{n} = \begin{bmatrix} n_x & n_y & n_z \end{bmatrix}^T = \begin{bmatrix} \sin\theta\cos\varphi & \sin\theta\sin\varphi & \cos\theta \end{bmatrix}^T \tag{2}$$

4. Method

In this work, we aim to estimate the 3D shape (per-pixel surface normals and depth) of an object or a scene solely from single-view polarization images - the problem popularly known as Shape from Polarization (SfP). The key reason for using a deep learning framework is to overcome the inherent ambiguities and compensate for limitations in SfP by modeling the physics of polarization through a completely data-driven approach. The learning-based SfP methods such as [BGW*20, LQX*22] thus far have shown improved performance over the classical physics-based methods. Interestingly, while Ba *et al.* [BGW*20] show that a deep network guided by physics-based priors on surface normals has been successful in obtaining optimal solutions from the polarization information, Lei *et al.* [LQX*22] have shown improved performance without including any such physics-based priors. However, both of these methods have not explicitly handled mixed reflections. Moreover, since the physical priors on surface normals solely rely on the kind of reflection, the fundamental task is to estimate somehow the nature and the extent of reflection (diffuse or specular) at each pixel.

4.1. Overview

To address this, we first describe the modified polarization image formation under mixed polarization to obtain the reflectance components - diffuse component (A_d) and specular component (A_s) along with the respective degrees of polarization (ρ_d and ρ_s) and the respective phase angles (ϕ_d and ϕ_s). Next, we use an encoder-decoder-based architecture to estimate the surface normals and depth (through two separate decoder branches) using the polarization images, the reflectance cues ($\alpha_d = A_d/A$ and $\alpha_s = A_s/A$), and the view encoding (v_{enc}). We think of α_d and α_s as the extent of diffuse and specular reflections at each pixel. The existing methods use polarization information (A, ρ, ϕ) along with the raw polarization images $\{I(\phi^i_{pol})\}$ where $i \in \{1,2,3,4\}$ as input which is seemingly redundant, since the polarization information can be derived from the raw polarization images, and can then regress to surface normals $\hat{\mathbf{n}}$. Unlike the method by Ba *et al.* [BGW*20] that uses physics-based priors on normals, we use the derived reflectance cues to guide the network for shape estimation. These physics-based priors are often ambiguous and time-consuming to compute ($\sim$ 1.2 seconds, see discussion in Section 6 and Table 2). View encoding accounts for non-orthographic projections for scene-level SfP, as described in [LQX*22]. Further, we use these learned surface normals (and depth derivatives) to recover the complete polarization information $\{I(\phi_{pol}), \rho, \phi\}$ under mixed polarization, as per Equation 1. It is to be noted that for the latter to be recovered correctly, we expect the surface normal estimates to be unambiguous and error-free. In due course, the network learns the surface properties (reflectance and refractive index) under mixed polarization by attempting to handle the underlying ambiguities in a completely self-supervised setting.

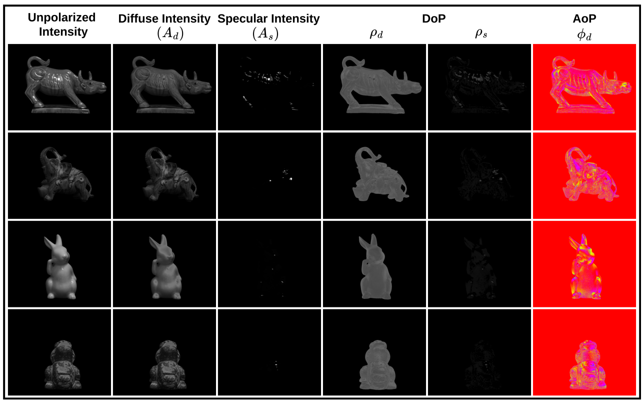

Figure 2: *Qualitative results reflectance separation (A_d, A_s) and recovered polarization information - Angle of Polarization (ϕ_d) and Degree of Polarization (ρ_d, ρ_s).*

4.2. Reflection Separation under Mixed Polarization

Consider an object or a scene illuminated by an unpolarized light source. The reflected light observed on the object/scene surface consists of partially polarized diffuse and specular reflection components. Following Equation 1, the intensity of the partially polarized diffuse reflection component can be obtained as per Equation 3.

$$I_d(\phi_{pol}) = A_d + B_d cos(2\phi_{pol} - 2\phi_d) \tag{3}$$

As discussed earlier, A_d and B_d are the DC component and amplitude of the diffuse reflection component, respectively. ϕ_d and

$\rho_d = B_d/A_d$ are the phase angle and degree of polarization for the diffuse component, respectively. Along similar lines, we describe the intensity of the partially polarized specular reflection with A_s, B_s ϕ_s, and $\rho_s = B_s/A_s$ being DC component, the amplitude, the phase angle, and degree of polarization of the specular reflection component, respectively, as described in Equation 4.

$$I_s(\phi_{pol}) = A_s + B_s cos(2\phi_{pol} - 2\phi_s) \tag{4}$$

Furthermore, the phase angles of diffuse and specular reflection components are related as per Equation 5.

$$\phi_s = \phi_d \pm \frac{\pi}{2} \tag{5}$$

Finally, combining the polarimetric properties in Equation 3, 4, and 5 such that $I(\phi_{pol}) = I_d(\phi_{pol}) + I_s(\phi_{pol})$ [TKR16], we can represent the radiance at a pixel for a particular polarizer angle (ϕ_{pol}) as described in Equation 6.

$$\begin{aligned} I(\phi_{pol}) &= [A_d + B_d cos(2\phi_{pol} - 2\phi_d)] \\ &\quad + [A_s - B_s cos(2\phi_{pol} - 2\phi_d)] \end{aligned} \tag{6}$$

Therefore, we can rewrite Equation 6 to obtain Equation 7 as follows.

$$\begin{aligned} I(\phi_{pol}) &= (A_d + A_s) + (B_d - B_s) cos(2\phi_{pol} - 2\phi_d) \\ &= A_m + B_m cos(2\phi_{pol} - 2\phi_d) \end{aligned} \tag{7}$$

Here, A_m and B_m are the DC component and amplitude of the reflected polarized light under mixed polarization, respectively, such that the Equation 8 and 9 holds.

$$A_m = A_d + A_s \tag{8}$$

$$B_m = B_d - B_s = \rho_d A_d - \rho_s A_s \tag{9}$$

4.2.1. Estimating A_m, B_m, and ϕ_d

We fit a cosine curve to the pixel values observed through four polarization angles, e.g. $0°, 45°, 90°$, and $135°$ to obtain A_m, B_m, and ϕ_d (upto the π-ambiguity) as the least squares solution to a system of linear equations. This approach better incorporates robustness to noise [HRKH10]. It is worth noting that the phase angle estimation in the case of color polarization images (as in [KPKO23]) becomes a little more tedious since each color channel could have different phase angles due to the noise present in pixel values. Rewriting Equation 7 in the vectorized form we obtain Equation 10.

$$I(\phi^i_{pol}) = \begin{bmatrix} 1 \\ \cos(2\phi^i_{pol}) \\ \sin(2\phi^i_{pol}) \end{bmatrix}^T \begin{bmatrix} A_m \\ B_m \cos(2\phi_d) \\ B_m \sin(2\phi_d) \end{bmatrix} = f_i^T x \tag{10}$$

Using this, we obtain the linear system of equations in Equation 11.

$$\mathbb{I} = \mathbb{A}\mathbf{x} \tag{11}$$

such that $\mathbb{I} = \begin{bmatrix} I(0) \\ I(\frac{\pi}{4}) \\ I(\frac{\pi}{2}) \\ I(\frac{3\pi}{4}) \end{bmatrix}$ and $\mathbb{A} = \begin{bmatrix} f_1^T \\ f_2^T \\ f_3^T \\ f_4^T \end{bmatrix}$.

Solving Equation 11, we get $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^T$ such that,

$$A_m = x_1, \; B_m = \sqrt{x_2^2 + x_3^2}, \text{ and } \phi_d = \frac{1}{2}\tan^{-1}\left(\frac{x_3}{x_2}\right) \tag{12}$$

As per Equation 5, we can obtain $\phi_s = \phi_d \pm \frac{\pi}{2}$.

4.2.2. Estimating A_d, A_s, ρ_d, and ρ_s

Given that we have the values of A_m and B_m for each pixel, we solve for the other unknowns (A_d, A_s, ρ_d, and ρ_s) in Equation 8 and 9 by nonlinear minimization as described in Equation 13.

$$\min_{A_d, A_s, \rho_d, \rho_s} \sum_{\text{all pixels}} \left[[A_m - (A_d + A_s)]^2 + [B_m - (\rho_d A_d - \rho_s A_s]^2 \right] \tag{13}$$

We solve Equation 13 via alternating least squares estimation. We first compute A_d and A_S by solving Equation 8 through the least squares. Here, we impose the non-negative constraints $A_d \geq 0$ and $A_s \geq 0$ since the DC components of the diffuse and specular reflection components are non-negative. Similarly, we then compute B_d and B_s in Equation 9 through the least squares estimation under non-negative constraints ($B_d \geq 0$ and $B_s \geq 0$). However, since the phase angle is estimated with π-ambiguity, we solve the modified version of Equation 9 such that we get Equation 14,

$$\pm B_m = B_d - B_s \tag{14}$$

Finally, we obtain $\rho_d = B_d/A_d$ and $\rho_s = B_s/A_s$. Figure 2 shows the diffuse and specular reflectance components for a few objects from the DeepSfP dataset [BGW*20].

4.3. Network Design

The detailed architecture of the proposed inverse rendering-based framework - SS-SfP is shown in Figure 1. We employ the widely adopted encoder-decoder architecture similar to the existing learning-based works [DLG21, BGW*20, LQX*22]. The encoder comprises five convolutional residual blocks, each with two convolutional layers. We use instance normalization in the encoder layers for better convergence (empirical evaluation is provided in the supplementary material). The encoder takes four polarization images $\{I(0), I(\frac{\pi}{4}), I(\frac{\pi}{2}), I(\frac{3\pi}{4})\}$ as input. The decoder is specialized through SPADE normalization [PLWZ19] (inspired by [BGW*20]) and is split into two branches - one for surface normal and the other for depth estimation. Each decoder branch takes the same input from the encoder. Additionally, we inject the unpolarised intensity (A), reflectance cues (α_d, α_s), and view encoding (v_{enc}) through SPADE normalization block into the decoder blocks to better model the surface properties and guide the network for shape estimation. For view encoding, we first compute the displacement field (du, dv) at each pixel, representing the displacement of each pixel from the center of the image grid, and then normalize to $[-1, 1]$ such that $v_{enc} = \begin{bmatrix} du & dv & 1 \end{bmatrix}^T$. The estimated surface normal maps along with A, α_d, α_s, and v_{enc} are then passed through two separate MLPs to recover ϕ and ρ, respectively, which are then passed through the rendering equation (Equation 1) to reconstruct the input raw polarization images $(\widehat{I}(\phi^i_{pol}))$.

4.3.1. Refractive Index Estimation

The learning-based approach implicitly learns the material properties (such as reflectance and refractive index) from the polarization

Proceedings published by Eurographics - The European Association for Computer Graphics.

Datasets	Miyazaki *et al.* [MTHI03]	Mahmoud *et al.* [MEMF12]	Smith *et al.* [SRT18]	Supervised (S)			Self-Supervised (SS)		
				Ba *et al.* [BGW*20]	Lei *et al.* [LQX*22]	SS-SfP	Ba *et al.* [BGW*20]	Lei *et al.* [LQX*22]	SS-SfP
DeepSfP	43.94	51.79	45.39	18.52	14.68	14.61	27.74	21.59	16.89
SPW	55.34	52.14	50.42	28.43	17.86	18.69	32.69	29.81	19.77

Table 1: *Quantitative comparison of SS-SfP in terms of mean angular of the estimated surface normals (in degree) with the baseline methods both evaluated under supervised (S) and self-supervised (SS) setting. The metrics reported are over the test sets of DeepSfP [BGW*20] and SPW [LQX*22] datasets. GREEN and YELLOW represent the best and the second best performance, respectively.*

data instead of using an explicit polarisation model for ρ with refractive index as a parameter. As described in Section 4, we estimate the shape (surface normal and depth) in a refractive index invariant manner and obtain the zenith angle as $\theta = cos^{-1}(n_z)$. With the zenith angle (θ), ρ_d, and ρ_s, we can obtain an optimal refractive index (η_{opt}) by solving a nonlinear least squares problem, as described in Equation 15.

$$\eta_{opt} = arg\min_{\eta} \sum_x \left(\left|\left| \alpha_d(x) \left(\rho_d(x) - \widehat{\rho}_d(\theta(x), \eta) \right) \right|\right|_2^2 + \left|\left| \alpha_s(x) \left(\rho_s(x) - \widehat{\rho}_s(\theta(x), \eta) \right) \right|\right|_2^2 \right) \quad (15)$$

Here, $\widehat{\rho}_d$ and $\widehat{\rho}_s$ are reflection-specific DoPs for diffuse and specular polarization [Col05] (described in the supplementary material). Interestingly, we rely on single-view polarimetric images for refractive index recovery, unlike previous attempts that have either considered multi-spectral images [HRKH10, HRKH13] or two source polarimetric images [TZS*21]. We assume a uniform refractive index over the entire object, as considered in earlier works [BGW*20, LQX*22] and utilize the fact that the dependency of ρ on η is weak [AH06]. Our refractive index estimate falls well within the range 1.3 to 1.6 (as prescribed for most dielectric materials) with a *mean* 1.486 and a *standard deviation* ± 0.26 across objects in the DeepSfP test set.

4.3.2. Optimization Objective

The entire network is optimized using a combination of reconstruction loss, geometric constraint, and polarization angle ratio constraint.

Reconstruction loss. We minimize the L2-loss over input ($I(\phi^i_{pol})$) and recovered ($\widehat{I}(\phi^i_{pol})$) polarization images and the desired (ρ) and the recovered ($\widehat{\rho}$) DoP, and L1-loss over the desired (ϕ) and the recovered ($\widehat{\phi}$) AoP .

$$\mathcal{L}_{rec} = \lambda_1 \sum_{i=1}^{4} ||\widehat{I}(\phi^i_{pol}) - I(\phi^i_{pol})||_2^2 + \lambda_2 ||\widehat{\rho} - \rho||_2^2 + \lambda_3 ||\widehat{\phi} - \phi||_1 \quad (16)$$

Here, $\lambda_1 = 1.0, \lambda_2 = 2.5$, and $\lambda_3 = 2.5$.

Geometric Constraint. To ensure smooth gradient flow from $\mathcal{L}_{rec}$ especially when there is no direct supervision for surface normals, we deploy the geometry constraint between the estimated surface normals and the estimated depth map.

$$\mathcal{L}_{geo} = \sum_{\text{all pixels}} (1 - \mathbf{n}^T \mathbf{z}_d) \quad (17)$$

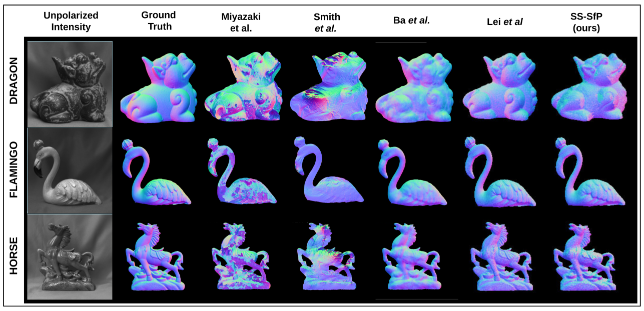

Figure 3: *Qualitative results on surface normal estimation over a few objects from DeepSfP dataset [BGW*20].*

Here, $\mathbf{z}_d = \frac{\begin{bmatrix} -z_x & -z_y & 1 \end{bmatrix}^T}{\sqrt{z_x{}^2 + z_y{}^2 + 1}}$ with $z_x = \frac{\partial z}{\partial x}$ and $z_y = \frac{\partial z}{\partial y}$ being the first-order derivatives of the depth map in x and y directions, respectively.

Polarizer Angle Ratio Constraint. We also constrain the estimation of surface normal by taking the ratio of $I(\phi_{pol})$ at polarizer angles $\phi_{pol} = 0$ and $\frac{\pi}{4}$ separately for diffuse and specular dominant regions. The following ratio constraint is inspired by the finding of Mecca *et al.* [MLC17, LMSC19].

$$\mathcal{L}_{ratio} = ||\alpha_d(F_d z_x - (-G_d) z_y)||_1 + ||\alpha_s(F_s z_y - G_s z_x)||_1 \quad (18)$$

Here, $F_\star = (-I(\frac{\pi}{4}) + A_\star)$ and $G_\star = (I(0) - A_\star + \rho_\star A_\star)$ with $\star \in \{d, s\}$, representing diffuse or specular case. Please refer to the supplementary material for detailed derivations of the constraints under diffuse and specular polarization dominance.

The overall optimization objective is thus described as per Equation 19.

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_{geo} \mathcal{L}_{geo} + \lambda_{ratio} \mathcal{L}_{ratio} \quad (19)$$

Here, $\lambda_{geo} = 1.0$ and $\lambda_{ratio} = 1.0$.

5. Experimental Evaluation

In this section, we analyze the performance of the proposed framework and compare it against three physics-only methods [SRT18, MTHI03, MEMF12] and two learning-based methods [BGW*20, LQX*22] for SfP trained over two datasets, DeepSfP [BGW*20] and SPW [LQX*22]. The baseline methods have either assumed diffuse reflection [SRT18], known lightings and/or albedos [SRT18, MEMF12], or use ground truth normals for supervision [BGW*20, LQX*22]. We report the widely used mean angular error (MAE) score (in degree) for quantitative analysis. Training of SS-SfP is not required as it learns at the test time through optimization.

Objects	Box	Dragon	Christmas	Flamingo	Horse	Vase	Whole Set	Time(s)
Ba *et al.* [BGW*20] (S)	23.51	21.55	13.50	20.19	22.27	10.32	18.52	1.181
Lei *et al.* [LQX*22] (S)	16.21	18.01	10.19	17.11	18.29	8.25	14.68	0.153
SS-SfP	19.57	19.14	13.29	18.62	20.05	10.61	16.89	0.183

Table 2: *Quantitative comparison over each object in the test sets of DeepSfP dataset [BGW*20] in terms of mean angular of the estimated surface normals (in degree). Last column: pre-processing time for polarization representation (1024×1224) provided as input to different methods. Tested on single thread Intel Core i7 processor clocking at 3.60GHz. GREEN and YELLOW represent the best and the second best performance, respectively.*

Lighting conditions	Miyazaki *et al.* [MTHI03] (S)	Mahmoud *et al.* [MEMF12] (S)	Smith *et al.* [SRT18] (S)	Ba *et al.* [BGW*20] (S)	Lei *et al.* [LQX*22] (S)	SS-SfP
Indoor	47.66	50.06	38.82	17.97	13.71	15.79
Outdoor Cloudy	43.24	46.76	36.51	16.36	13.07	15.12
Outdoor Sunny	51.63	51.27	48.99	21.24	17.25	19.76

Table 3: *Effect of varying lighting conditions over DeepSfP dataset [BGW*20] in terms of mean angular of the estimated surface normals (in degree). (S) and (SS) represent evaluation under supervised and self-supervised settings, respectively. GREEN and YELLOW represent the best and the second best performance, respectively.*

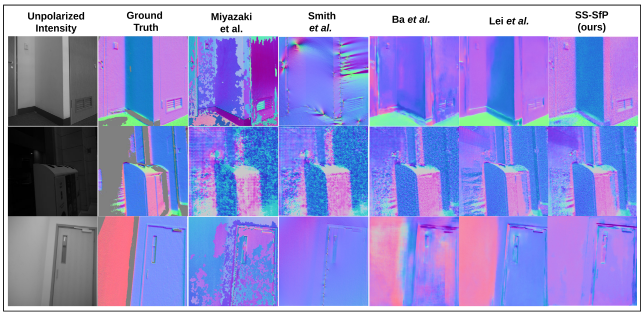

Figure 4: *Qualitative results on surface normal estimation on a few scenes from SPW dataset [LQX*22]*

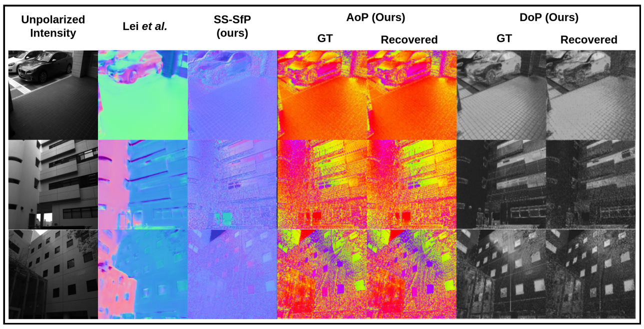

Figure 5: *Qualitative results on far-field outdoor scenes. Since ground truth normals are unavailable, we validate the efficacy of estimated normals through the recovered AoP ($\widehat{\phi}$) and DoP ($\widehat{\rho}$).*

5.1. Quantitative Evaluation

Evaluation Setting. Our prime focus is to highlight the strength of the SS-SfP for self-supervised shape estimation. However, since there is no self-supervised framework available and for a fair comparison, we evaluate the learning-based baseline methods [BGW*20, LQX*22] and the proposed SS-SfP under two settings: Supervised (S) and Self-supervised (SS) and compare them in the respective settings. Kindly refer to the supplementary material for details of training and evaluation under each setting.

We find three key observations from Table 1. (a) The proposed SS-SfP performs either the best or second best in either of the training settings. (b) SS-SfP outperforms all the baseline methods (across both the settings) except that of Lei *et al.* [LQX*22] (under a fully supervised setting), where it falls short by a small margin owing to the benefit of direct supervision. Furthermore, while Lei *et al.* [LQX*22] use the known camera intrinsics-based view encoding for handling non-orthographic projection (designed specifically for scene-level SfP), we employ a relatively simpler displacement field-based view-encoding. (c) The reduced performance of methods proposed by Ba *et al.* [BGW*20] and Lei *et al.* [LQX*22] under a self-supervised setting can be attributed to the fact that the network presumably finds an easier way to recover ρ and ϕ directly from the input (remember, ρ and ϕ are part of the input to [BGW*20, LQX*22]) and focus less on the normal estimation since there is no direct supervision. Figure 5 shows a qualitative comparison of the method by Lei *et al.* [LQX*22] and SS-SfP under self-supervised setting since the ground truth normals are not present for those far-field outdoor scenes.

Table 2 examines the performance of each object in the DeepSfP dataset. Table 3 shows the robustness of the proposed method to different lighting conditions (also, Figure 5). We find that SS-SfP exhibits only slight variation in MAE across indoor and outdoor overcast lighting. However, it shows a little higher MAE over outdoor solar lighting. This is primarily because the observed diffuse reflection dominates over the specular ones due to the high intensity of incoming light from the sun, leading to a slightly poor judgement about the reflection-dependent normal orientation. Furthermore, while we assume unpolarized incident illumination, the illumination incident on the scene under outdoor lighting could be partially polarized due to the reflections from other surfaces (acting as potential light sources). Overall, the framework yields an average SSIM of **0.836** and **0.714** on the recovered DoP, an average SSIM of **0.917** and **0.822** on the recovered polarization images, and a mean absolute error of **0.961°** and **1.772°** on the recovered AoP over the DeepSfP [BGW*20] and SPW [LQX*22] test sets, respectively.

ID	Encoder Input					Decoder Input				Depth and Normal	MAE	
	Raw pol images	(A,ρ,ϕ)	Normal Priors	(A,α_d,α_s)	VE	Encoder out	Normal Priors	(A,α_d,α_s)	VE	#Branches	DeepSfP	SPW
1	✓	✗	✗	✗	✗	✓	✗	✗	✗	2	30.26	40.75
2	✓	✓	✗	✗	✗	✓	✗	✗	✗	2	29.14	39.18
3	✓	✗	✓	✗	✗	✓	✗	✗	✗	2	20.98	32.75
4	✓	✗	✗	✓	✗	✓	✗	✗	✗	2	20.51	30.94
5	✓	✗	✗	✓	✓	✓	✗	✗	✗	2	20.22	21.63
6	✓	✗	✗	✗	✗	✓	✓	✗	✓	2	17.91	20.87
7	✓	✗	✗	✗	✗	✓	✗	✓	✓	2	**16.89**	**19.77**
8	✓	✗	✗	✗	✗	✓	✗	✓ (w/o SPADE)	✓ (w/o SPADE)	2	20.96	23.71
9	✓	✗	✗	✗	✗	✓	✗	✓	✓	1	21.29	27.19
10	SS-SfP: Without instance normalization in the encoder										18.29	21.38
11	SS-SfP: Decoder with Self-Attention (as proposed in [LQX*22])										19.69	21.78
12	SS-SfP: without geometric constraint ($\mathcal{L}_{geo}$)										22.13	29.05
13	SS-SfP: without polarization angle ratio constraint ($\mathcal{L}_{ratio}$)										19.97	27.16

Table 4: *Summary of ablation study over various design choices (ID 1-9) and architectural variations (ID 10-13) for the proposed framework.*

5.2. Qualitative Evaluation

Figure 3 and 4 show the qualitative comparison of SS-SfP over the other baseline methods. In Figure 3, we show the qualitative results on three complex objects: DRAGON, FLAMINGO, and HORSE in the DeepSfP dataset. We observe that SS-SfP is not too far from a completely supervised method by Lei *et al.* and performs better for some scenes (row 1, Figure 4) and nearly equal for objects in rows 2 and 3, Figure 3. SS-SfP can also scale to far-field outdoor scenes (see Figure 5) with distances far beyond the depth captured in the SPW dataset, leveraging the fact that the relationship between polarized light and surface normals remains unaffected by distance. While the normal estimates exhibit limited directional variations despite better AoP and DoP reconstructions (one of the limitations), they capture the finer geometrical variations in the scenes, such as tile partitions on the floor, glass shields on the car, and glass windows on the building (Figure 5), which are not prominent in the results obtained by Lei *et al.* [LQX*22]. This indicates the ability of the network to learn the shape from polarization images, especially when ground truth capture is difficult and inaccurate, and validates that SfP does not deviate from mere image semantics. Moreover, we observe that extreme high-frequency variations in the shapes are not captured well compared to their supervised counterparts (another limitation), leading to higher MAE in Table 1 and 2 and some blocky artefacts, especially in HORSE and DRAGON of the DeepSfP dataset (Figure 3). This behavior is generally observed for optimization-based self-supervised approaches.

6. Ablation and Discussion

Table 4 summarizes the effect of several design choices and potential variants of the proposed framework.

Effect of reflectance cues and view encoding. We obtain lower MAE when reflectance cues and view encoding are injected into the decoder than when they are provided as input to the encoder (compare IDs 5 and 7, Table 4). Since such deep residual architectures tend to wash away necessary information obtained from the input [HSL*16, BGW*20], it becomes essential to introduce (or re-introduce) guiding information to the decoder. View encoding subsumes the impact of viewing direction on the polarization data (note the difference of 9.31° among IDs 4 and 5).

Effect of SPADE normalization. We use SPADE normalization for two reasons: (a) it is a generalization of several other existing normalization schemes [PLWZ19] and (b) it is suited for better image understanding and synthesis through semantic layouts (diffuse and specular separation, in our case). We find that SPADE normalization helps the decoder better manage the reflectance cues and handle ambiguities for surface normal estimation through adaptive modulation parameters. In general, we observe better performance with SPADE normalization, irrespective of the information fed at its input (compare IDs 5, 6, 8 with 7, Table 4).

Physics-based normal priors vs reflectance cues. We find that the performance with reflectance cues is better than that when using normal priors (Table 4 (ID 6 and 7)) although the difference in MAE is just around 1°. Moreover, we find additional concerns with such handcrafted priors on normals. (a) Normal priors derived using reflectance-specific DoP [Col05] inherently have ambiguous angles since no such model characterizes the priors on mixed reflections. (b) They are sensitive to noise present in the raw polarization images. Merely shifting the azimuth angles by π or $\pi/2$ will not recover proper surface normals from noisy images. (c) Their computation is time-consuming and leads to slower inference (see Table 2, last column).

Effect of estimating depth. The depth is estimated to devise a geometrical constraint (used in literature) for better normal estimates (ID 7, Table 4). However, estimating depth directly shows poor performance (see ID 9, Table 4) and is mainly attributed to the discontinuities offered by the differentiation step in the depth estimates (and thus, the surface normal map) [YZS17].

Additional ablations studies, implementation details, qualitative results, and other necessary information are detailed in the supplementary material.

7. Conclusion

We proposed SS-SfP - a neural inverse rendering-based self-supervised framework to address SfP under mixed reflections by obtaining per-pixel diffuse and specular reflectance components and refractive index estimation in a completely self-supervised manner. We evaluated SS-SfP under both supervised and self-supervised settings with a heavy emphasis on self-supervised learning. Further, SS-SfP is also shown to extend to far-field outdoor scenes and opens up a direction toward in-the-wild geometry reconstruction from polarization images. We believe this work would generate more traction towards self-supervised learning methods to address SfP under mixed reflections without needing 3D ground truth.

Acknowldgement. This work is supported by the Prime Minister Research Fellowship (PMRF) awarded to Ashish Tiwari and Jibaben Patel Chair in AI awarded to Shanmuganathan Raman.

Proceedings published by Eurographics - The European Association for Computer Graphics.

References

[AH06] ATKINSON G. A., HANCOCK E. R.: Recovery of surface orientation from diffuse polarization. *IEEE transactions on image processing 15*, 6 (2006), 1653–1664. 5

[AH07] ATKINSON G. A., HANCOCK E. R.: Shape estimation using polarization and shading from two views. *IEEE transactions on pattern analysis and machine intelligence 29*, 11 (2007), 2001–2017. 1, 2

[Atk17] ATKINSON G. A.: Polarisation photometric stereo. *Computer Vision and Image Understanding 160* (2017), 158–167. 2

[BGW*20] BA Y., GILBERT A., WANG F., YANG J., CHEN R., WANG Y., YAN L., SHI B., KADAMBI A.: Deep shape from polarization. In *European Conference on Computer Vision* (2020), Springer, pp. 554–571. 1, 2, 3, 4, 5, 6, 7

[BJTK18] BAEK S.-H., JEON D. S., TONG X., KIM M. H.: Simultaneous acquisition of polarimetric svbrdf and normals. *ACM Trans. Graph. 37*, 6 (2018), 268–1. 2

[BVM17] BERGER K., VOORHIES R., MATTHIES L. H.: Depth from stereo polarization in specular scenes for urban robotics. In *2017 IEEE international conference on robotics and automation (ICRA)* (2017), IEEE, pp. 1966–1973. 1

[CGS*17] CUI Z., GU J., SHI B., TAN P., KAUTZ J.: Polarimetric multi-view stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 1558–1567. 2

[Col05] COLLETT E.: Field guide to polarization. Spie Bellingham, WA. 1, 5, 7

[DJZ*21] DING Y., JI Y., ZHOU M., KANG S. B., YE J.: Polarimetric helmholtz stereopsis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 5037–5046. 2

[DLG21] DESCHAINTRE V., LIN Y., GHOSH A.: Deep polarization imaging for 3d shape and svbrdf acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 15567–15576. 2, 4

[DZV22] DAVE A., ZHAO Y., VEERARAGHAVAN A.: Pandora: Polarization-aided neural decomposition of radiance. *arXiv preprint arXiv:2203.13458* (2022). 2

[FKNN21] FUKAO Y., KAWAHARA R., NOBUHARA S., NISHINO K.: Polarimetric normal stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 682–690. 2

[GCP*10] GHOSH A., CHEN T., PEERS P., WILSON C. A., DEBEVEC P.: Circularly polarized spherical illumination reflectometry. In *ACM SIGGRAPH Asia 2010 papers*. 2010, pp. 1–12. 2

[GFT*11] GHOSH A., FYFFE G., TUNWATTANAPONG B., BUSCH J., YU X., DEBEVEC P.: Multiview face capture using polarized spherical gradient illumination. In *Proceedings of the 2011 SIGGRAPH Asia Conference* (2011), pp. 1–10. 2

[GPDG12] GUARNERA G. C., PEERS P., DEBEVEC P., GHOSH A.: Estimating surface normals from spherical stokes reflectance fields. In *Computer Vision–ECCV 2012. Workshops and Demonstrations: Florence, Italy, October 7-13, 2012, Proceedings, Part II 12* (2012), Springer, pp. 340–349. 2

[HRKH10] HUYNH C. P., ROBLES-KELLY A., HANCOCK E.: Shape and refractive index recovery from single-view polarisation images. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2010), IEEE, pp. 1229–1236. 2, 4, 5

[HRKH13] HUYNH C. P., ROBLES-KELLY A., HANCOCK E. R.: Shape and refractive index from single-view spectro-polarimetric images. *International journal of computer vision 101*, 1 (2013), 64–94. 5

[HSL*16] HUANG G., SUN Y., LIU Z., SEDRA D., WEINBERGER K. Q.: Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14* (2016), Springer, pp. 646–661. 7

[KOS*20] KONDO Y., ONO T., SUN L., HIRASAWA Y., MURAYAMA J.: Accurate polarimetric brdf for real polarization scene rendering. In *European Conference on Computer Vision* (2020), Springer, pp. 220–236. 2

[KPKO23] KAJIYAMA S., PIAO T., KAWAHARA R., OKABE T.: Separating partially-polarized diffuse and specular reflection components under unpolarized light sources. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2023), pp. 2549–2558. 3, 4

[KTSR15] KADAMBI A., TAAMAZYAN V., SHI B., RASKAR R.: Polarized 3d: High-quality depth sensing with polarization cues. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 3370–3378. 1, 2

[LLK*02] LIN S., LI Y., KANG S. B., TONG X., SHUM H.-Y.: Diffuse-specular separation and depth recovery from image sequences. In *European conference on computer vision* (2002), Springer, pp. 210–224. 2

[LMSC19] LOGOTHETIS F., MECCA R., SGALLARI F., CIPOLLA R.: A differential approach to shape from polarisation: A level-set characterisation. *International Journal of Computer Vision 127* (2019), 1680–1693. 2, 5

[LQX*22] LEI C., QI C., XIE J., FAN N., KOLTUN V., CHEN Q.: Shape from polarization for complex scenes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 12632–12641. 1, 2, 3, 4, 5, 6, 7

[MBMS23] MUGLIKAR M., BAUERSFELD L., MOEYS D. P., SCARAMUZZA D.: Event-based shape from polarization. *arXiv preprint arXiv:2301.06855* (2023). 2

[MCZ*22] MINGQI S., CHONGKUN X., ZHENDONG Y., JUNNAN H., XUEQIAN W.: Transparent shape from single polarization images. *arXiv preprint arXiv:2204.06331* (2022). 1

[MEMF12] MAHMOUD A. H., EL-MELEGY M. T., FARAG A. A.: Direct method for shape recovery from polarization and shading. In *2012 19th IEEE International Conference on Image Processing* (2012), IEEE, pp. 1769–1772. 2, 5, 6

[MFSG06] MOREL O., FERRATON M., STOLZ C., GORRIA P.: Active lighting applied to shape from polarization. In *2006 International Conference on Image Processing* (2006), IEEE, pp. 2181–2184. 2

[MHP*07] MA W.-C., HAWKINS T., PEERS P., CHABERT C.-F., WEISS M., DEBEVEC P. E., ET AL.: Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. *Rendering Techniques 2007*, 9 (2007), 10. 2

[MLC17] MECCA R., LOGOTHETIS F., CIPOLLA R.: A differential approach to shape from polarization. 2, 5

[MSB*16] MIYAZAKI D., SHIGETOMI T., BABA M., FURUKAWA R., HIURA S., ASADA N.: Surface normal estimation of black specular objects from multiview polarization images. *Optical Engineering 56*, 4 (2016), 041303. 2

[MTHI03] MIYAZAKI D., TAN R. T., HARA K., IKEUCHI K.: Polarization-based inverse rendering from a single view. In *Computer Vision, IEEE International Conference on* (2003), vol. 3, IEEE Computer Society, pp. 982–982. 2, 5, 6

[NFB97] NAYAR S. K., FANG X.-S., BOULT T.: Separation of reflection components using color and polarization. *International Journal of Computer Vision 21*, 3 (1997), 163–186. 2

[NZI01] NISHINO K., ZHANG Z., IKEUCHI K.: Determining reflectance parameters and illumination distribution from a sparse set of images for view-dependent image synthesis. In *Proceedings Eighth IEEE international conference on computer vision. ICCV 2001* (2001), vol. 1, IEEE, pp. 599–606. 2

[PF05] PRADOS E., FAUGERAS O.: Shape from shading: a well-posed problem? In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (2005), vol. 2, IEEE, pp. 870–877. 1

[PLWZ19] PARK T., LIU M.-Y., WANG T.-C., ZHU J.-Y.: Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 2337–2346. 4, 7

[RC01] RAHMANN S., CANTERAKIS N.: Reconstruction of specular surfaces using polarization imaging. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001* (2001), vol. 1, IEEE, pp. I–I. 2

[Sch11] SCHECHNER Y. Y.: Inversion by p 4: polarization-picture post-processing. *Philosophical Transactions of the Royal Society B: Biological Sciences 366*, 1565 (2011), 638–648. 1

[Sch15] SCHECHNER Y. Y.: Self-calibrating imaging polarimetry. In *2015 IEEE International Conference on Computational Photography (ICCP)* (2015), IEEE, pp. 1–10. 1

[SDMF12] SHABAYEK A. E. R., DEMONCEAUX C., MOREL O., FOFI D.: Vision based uav attitude estimation: Progress and insights. *Journal of Intelligent & Robotic Systems 65*, 1 (2012), 295–308. 1

[SRT18] SMITH W. A., RAMAMOORTHI R., TOZZA S.: Height-from-polarisation with unknown lighting or albedo. *IEEE transactions on pattern analysis and machine intelligence 41*, 12 (2018), 2875–2888. 2, 5, 6

[TI08] TAN R. T., IKEUCHI K.: Separating reflection components of textured surfaces using a single image. In *Digitally Archiving Cultural Objects*. Springer, 2008, pp. 353–384. 2

[TKR16] TAAMAZYAN V., KADAMBI A., RASKAR R.: Shape from mixed polarization. *arXiv preprint arXiv:1605.02066* (2016). 2, 4

[TZS*21] TOZZA S., ZHU D., SMITH W. A., RAMAMOORTHI R., HANCOCK E. R.: Uncalibrated, two source photo-polarimetric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence 44*, 9 (2021), 5747–5760. 5

[Woo80] WOODHAM R. J.: Photometric method for determining surface orientation from multiple images. *Optical engineering 19*, 1 (1980), 139–144. 1, 2

[YZS17] YU Y., ZHU D., SMITH W. A.: Shape-from-polarisation: a nonlinear least squares approach. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (2017), pp. 2969–2976. 2, 7

[ZS19] ZHU D., SMITH W. A.: Depth from a polarisation+ rgb stereo pair. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 7586–7595. 2