Deep Shape from Polarization

Yunhao Ba[1], Alex Gilbert[1⋆], Franklin Wang[1⋆], Jinfa Yang[2], Rui Chen[2],
Yiqin Wang[1], Lei Yan[2], Boxin Shi[2⋆⋆], and Achuta Kadambi[1⋆⋆]

[1] University of California, Los Angeles
{yhba, alexrgilbert}@ucla.edu, franklinxzw@gmail.com, achuta@ee.ucla.edu
[2] Peking University
{jinfayang, shiboxin}@pku.edu.cn

Abstract. This paper makes a first attempt to bring the Shape from Polarization (SfP) problem to the realm of deep learning. The previous state-of-the-art methods for SfP have been purely physics-based. We see value in these principled models, and blend these physical models as priors into a neural network architecture. This proposed approach achieves results that exceed the previous state-of-the-art on a challenging dataset we introduce. This dataset consists of polarization images taken over a range of object textures, paints, and lighting conditions. We report that our proposed method achieves the lowest test error on each tested condition in our dataset, showing the value of blending data-driven and physics-driven approaches.

Keywords: Shape from Polarization, 3D Reconstruction, Physics-based Deep Learning

1 Introduction

While deep learning has revolutionized many areas of computer vision, the deep learning revolution has not yet been studied in context of Shape from Polarization (SfP). The SfP problem is fascinating because, if successful, shape could be obtained in completely passive lighting conditions without estimating lighting direction. Recent progress in CMOS sensors has spawned machine vision cameras that capture the required polarization information in a single shot [42], making the capture process more relaxed than photometric stereo.

This SfP problem can be stated simply: light that reflects off an object has a polarization state that corresponds to shape. In reality, the underlying physics is among the most optically complex of all computer vision problems. For this reason, previous SfP methods have high error rates (in context of mean angular error (MAE) of surface normal estimation), and limited generalization to mixed materials and lighting conditions.

[⋆] Equal contribution.

[⋆⋆] Corresponding authors.
Project page: https://visual.ee.ucla.edu/deepsfp.htm

Table 1. Deep SfP vs Previous Methods. We compare the input constraints and result quality of the proposed hybrid of physics and learning compared to previous, physics-based SfP methods.

Method	Inputs	Mean Angular Error	Robustness to Texture-Copy	Lighting Invariance
Miyazaki [37]	Polarization Images	High	Strong	Moderate
Mahmoud [33]	Polarization Images	High	Not Observed	Moderate
Smith [52]	Polarization Images Lighting Estimate	Moderate	Strong	Moderate
Proposed	Polarization Images	Lowest	Strong	Strong

The physics of SfP are based on the Fresnel Equations. These equations lead to an underdetermined system— the so-called *ambiguity problem*. This problem arises because a linear polarizer cannot distinguish between polarized light that is rotated by π radians. This results in two confounding estimates for azimuth angle at each pixel. Previous work in SfP has used additional information to constrain the ambiguity problem. For instance, Smith *et al.* [51] use both polarization and shading constraints as linear equations when solving object depth, and Mahmoud *et al.* [33] use shape from shading constraints to correct the ambiguities. Other authors assume surface convexity to constrain the azimuth angle [4, 37] or use a coarse depth map to constrain the ambiguity [21, 22]. There are also additional binary ambiguities based on reflection type, as discussed in [4, 33]. Table 1 compares our proposed technique with prior work.

Another contributing factor to the underdetermined nature of SfP is the *refractive problem*. SfP needs knowledge of per-pixel refractive indices. Previous work has used hard-coded values to estimate the refractive index of scenes [37]. This leads to a relative shape recovered with refractive distortion.

Yet another limitation of the physical model is particular susceptibility to *noise*. The polarization signal is very subtle for fronto-parallel geometries so it is important that the input images are relatively noise-free. Unfortunately, a polarizing filter reduces the captured light intensity by 50 percent, worsening the effects of Poisson shot noise, encouraging a noise tolerant SfP algorithm.[1]

In this paper, we address these SfP pitfalls by moving away from a physics-only solution, toward the realm of data-driven techniques. While it is tempting to apply traditional deep learning models to the SfP problem, we find this approach does not maximize performance. Instead, we propose a physics-based learning algorithm that not only outperforms traditional deep learning, but also outperforms three baseline comparisons to physics-based SfP. We summarize our contributions as follows:

- a first attempt to apply deep learning techniques to solve the SfP problem;
- incorporation of the existing physical model into the deep learning approach;
- demonstration of significant error reduction; and

[1] For a detailed discussion of other sources of noise please refer to Schechner [47].

- introduction of the first polarization image dataset with ground truth shape, laying a foundation for future data-driven methods.

Limitations: As a physics-based learning approach, our technique still relies on computing the physical priors for every test example. This means that the per-frame runtime would be the sum of the compute time for the forward pass and that of the physics-based prior. Our runtime details are in the supplement. Future work could parallelize compute of the physical prior. Another limitation pertains to the accuracy inherent to SfP. Our average MAE on the test set is 18.5 degrees. While this is the best SfP performer on our challenging dataset, the error is higher than with a more controlled technique like photometric stereo.

2 Related Work

Polarization cues have been employed for various tasks, such as reflectometry estimation [12], radiometric calibration [58], facial geometry reconstruction [13], dynamic interferometry [32], polarimetric spatially varying surface reflectance functions (SVBRDF) recovery [5], and object shape acquisition [14, 31, 43, 64]. This paper is at the seamline of deep learning and SfP, offering unique performance tradeoffs from prior work. Refer to Table 1 for an overview.

Shape from Polarization infers the shape (usually represented in surface normals) of a surface by observing the correlated changes of image intensity with the polarization information. Changes of polarization information could be captured by rotating a linear polarizer in front of an ordinary camera [2,60] or polarization cameras using a single shot in real time (e.g., PolarM [42] in [62]). Conventional SfP decodes such information to recover the surface normal up to some ambiguity. If only images with different polarization information are available, heuristic priors such as the surface normals along the boundary and convexity of the objects are employed to remove the ambiguity [4, 37]. Photometric constraints from shape from shading [33] and photometric stereo [1, 11, 39] complements polarization constraints to make the normal estimates unique. If multi-spectral measurements are available, surface normal and its refractive index could be estimated at the same time [16, 17]. More recently, a joint formulation of shape from shading and SfP in a linear manner is shown to be able to directly estimate the depth of the surface [51, 52, 59]. This paper is the first attempt at combining deep learning and SfP.

Polarized 3D involves stronger assumptions than SfP and has different inputs and outputs. Recognizing that SfP alone is a limited technique, the Polarized 3D class of methods integrate SfP with a low resolution depth estimate. This additional constraint allows not just recovery of shape but also a high-quality 3D model. The low resolution depth could be achieved by employing two-view [3, 6, 35], three-view [8], multi-view [9, 36] stereo, or even in real time by using a SLAM system [62]. These depth estimates from geometric methods are not reliable in textureless regions where finding correspondence for triangulation is difficult. Polarimetric cues could be jointly used to improve such unreliable depth estimates to obtain a more complete shape estimation. A depth sensor such as the Kinect can also provide coarse depth prior to disambiguate the ambiguous normal estimates given by SfP [21,22]. The key step that characterizes Polarized 3D is a holistic approach that rethinks both SfP and the depth-normal fusion process. The main limitation of Polarized 3D is the strong requirement of a coarse depth map, which is not true for our proposed technique.

Data-driven computational imaging approaches draw much attention in recent years thanks to the powerful modeling ability of deep neural networks. Various types of convolutional neural networks (CNNs) are designed to enable 3D imaging for many types of sensors and measurements. From single photon sensor measurements, a multi-scale denoising and upsampling CNN is proposed to refine depth estimates [28]. CNNs also show advantage in solving phase unwrapping, multipath interference, and denoising jointly from raw time-of-flight measurements [34,54]. From multi-directional lighting measurements, a fully-connected network is proposed to solve photometric stereo for general reflectance with a pre-defined set of light directions [45]. Then the fully convolutional network with an order-agnostic max-pooling operation [7] and the observation map invariant to the number and permutation of the images [18] are concurrently proposed to deal with an arbitrary set of light directions. Normal estimates from photometric stereo can also be learned in an unsupervised manner by minimizing reconstruction loss [57]. Other than 3D imaging, deep learning has helped solve several inverse problems in the field of computational imaging [30,46,55,56]. Separation of shape, reflectance and illuminance maps for wild facial images can be achieved with the CNNs as well [48]. CNNs also exhibit potential for modeling SVBRDF of a near-planar surface [10,25,26,63], and more complex objects [27]. The challenge with existing deep learning frameworks is that they do not leverage the unique physics of polarization.

3 Proposed Method

In this section, we first introduce basic knowledge of SfP, and then present our physics-based CNN. Blending physics and deep learning improves the performance and generalizability of the method.

3.1 Image formation and physical solution

Our objective is to reconstruct surface normals $\hat{\boldsymbol{N}}$ from a set of polarization images $\{\boldsymbol{I}_{\phi_1}, \boldsymbol{I}_{\phi_2}, ..., \boldsymbol{I}_{\phi_M}\}$ with different polarization angles. For a specific polarization angle ϕ_{pol}, the intensity at a pixel of a captured image follows a sinusoidal variation under unpolarized illumination:

$$I(\phi_{pol}) = \frac{I_{max} + I_{min}}{2} + \frac{I_{max} - I_{min}}{2} \cos(2(\phi_{pol} - \phi)), \tag{1}$$

where ϕ denotes the phase angle, and I_{min} and I_{max} are lower and upper bounds for the observed intensity. Equation (1) has a π-***ambiguity*** in context of ϕ: two phase angles, with a π shift, will result in the same intensity in the captured images. Based on the phase angle ϕ, the azimuth angle φ can be retrieved with $\frac{\pi}{2}$-***ambiguity*** as follows [9]:

$$\phi = \begin{cases} \varphi, & \text{if diffuse reflection dominates} \\ \varphi - \frac{\pi}{2}, & \text{if specular reflection dominates} \end{cases}. \tag{2}$$

The zenith angle θ is related to the degree of polarization ρ, which can be written as:

$$\rho = \frac{I_{max} - I_{min}}{I_{max} + I_{min}}. \tag{3}$$

When diffuse reflection is dominant, the degree of polarization can be expressed with the zenith angle θ and the refractive index n as follows [4]:

$$\rho_d = \frac{(n - \frac{1}{n})^2 \sin^2\theta}{2 + 2n^2 - (n + \frac{1}{n})^2 \sin^2\theta + 4\cos\theta\sqrt{n^2 - \sin^2\theta}}. \tag{4}$$

The dependency of ρ_d on n is weak [4], and we assume $n = 1.5$ throughout the rest of this paper. With this known n, Equation (4) can be rearranged to obtain a close-form estimation of the zenith angle for the diffuse dominant case.

When specular reflection is dominant, the degree of polarization can be written as [4]:

$$\rho_s = \frac{2\sin^2\theta\cos\theta\sqrt{n^2 - \sin^2\theta}}{n^2 - \sin^2\theta - n^2\sin^2\theta + 2\sin^4\theta}. \tag{5}$$

Equation (5) can not be inverted analytically, and solving the zenith angle with numerical interpolation will produce two solutions if there are no additional constraints. For real world objects, specular reflection and diffuse reflection are mixed depending on the surface material of the object. As shown in Figure 1, the ambiguity in the azimuth angle and uncertainty in the zenith angle are fundamental limitations of SfP. Overcoming these limitations through physics-based neural networks is the primary focus of this paper.

3.2 Learning with physics

A straightforward approach to estimating the normals, from polarization would be to simply take the set of polarization images as input, encode it into a feature map using a CNN, and feed the feature map into a normal-regression sub-network. Unsurprisingly, we find this results in normal reconstructions with higher MAE and undesirable lighting artifacts (see Figure 7). To guide the network towards more optimal solutions from the polarization information, one possible method is to force our learned solutions to adhere to the polarization equations described in Section 3.1, similar to the method used in [23]. However, it is difficult to use these physical solutions for SfP tasks due to the following reasons: **1.** Normals derived from the equations will inherently have ambiguous azimuth angles. **2.** Specular reflection and diffuse reflection coexist simultaneously, and determining the proportion of each type is complicated. **3.** Polarization images are usually noisy, causing error in the ambiguous normals, especially

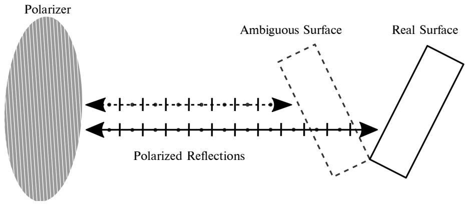

Fig. 1. SfP is underdetermined and one causal factor is the *ambiguity problem.* Here, two different surface orientations could result in exactly the same polarization signal, represented by dots and hashes. The dots represent polarization out of the plane of the paper and the hashes represent polarization within the plane of the board. Based on the measured data, it is unclear which orientation is correct. Ambiguities can also arise due to specular and diffuse reflections (which change the phase of light). For this reason, our network uses multiple physical priors.

when the degree of polarization is low. Shifting the azimuth angles by π or $\frac{\pi}{2}$ could not reconstruct the surface normals properly for noisy images.

Therefore, we propose directly feeding both the polarization images and ambiguous normal maps into the network, and leave the network to learn how to combine both of these inputs effectively from training data. The estimated surface normals can be structured as following:

$$\hat{\boldsymbol{N}} = f(\boldsymbol{I}_{\phi_1}, \boldsymbol{I}_{\phi_2}, ..., \boldsymbol{I}_{\phi_M}, \boldsymbol{N}_{diff}, \boldsymbol{N}_{spec1}, \boldsymbol{N}_{spec2}), \tag{6}$$

where $f(\cdot)$ is the proposed prediction model, $\{\boldsymbol{I}_{\phi_1},\ \boldsymbol{I}_{\phi_2},\ ...,\ \boldsymbol{I}_{\phi_M}\}$ is a set of polarization images, and $\hat{\boldsymbol{N}}$ is the estimated surface normals. We use the diffuse model in Section 3.1 to calculate $\boldsymbol{N}_{diff}$, and $\boldsymbol{N}_{spec1}, \boldsymbol{N}_{spec2}$ are the two solutions from the specular model. These ambiguous normals can implicitly direct the proposed network to learn the surface normal information from the polarization.

Our network structure is illustrated in Figure 2. It consists of a fully convolutional encoder to extract and combine high-level features from the ambiguous physical solutions and the polarization images, and a decoder to output the estimated normals, $\hat{\boldsymbol{N}}$. Although three polarization images are sufficient to capture the polarization information, we use images with a polarizer at $\phi_{pol} \in \{0°, 45°, 90°, 135°\}$. These images are concatenated channelwise with the ambiguous normal solutions as the model input.

Note that the fixed nature of our network input is not arbitrary, but based on the output of standard polarization cameras. Such cameras utilize a layer of polarizers above the photodiodes to capture these four polarization images in a single shot. Our network design is intended to enable applications using this current single-shot capture technology. Single-shot capture is a clear advantage of our method over alternative reconstruction approaches, such as photometric stereo, since it allows images to be captured in a less constrained setting.

After polarization feature extraction, there are five encoder blocks to encode the input to a $B \times 512 \times 8 \times 8$ tensor, where B is the minibatch size. The encoded

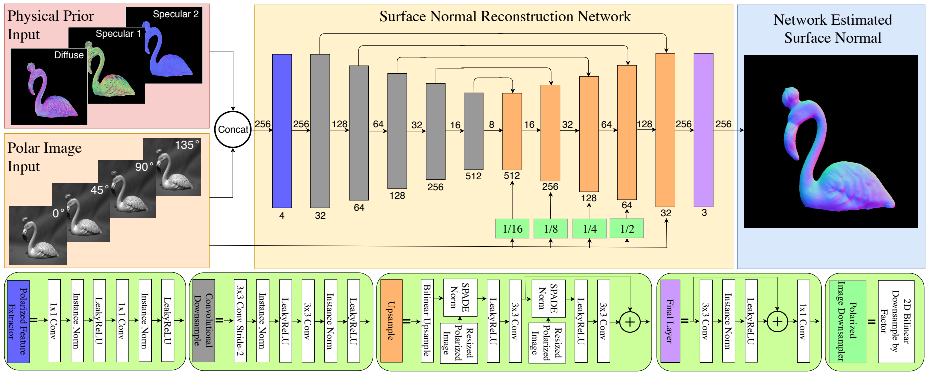

Fig. 2. Overview of our proposed physics-based neural network. The network is designed according to the encoder-decoder architecture in a fully convolutional manner. The blocks comprising the network are shown below the high-level diagram of our network pipeline. We use a block based on spatially-adaptive normalization as previously implemented in [40]. The numbers below the blocks refer to the number of output channels and the numbers next to the arrows refer to the spatial dimension.

tensor is then decoded by the same number of decoder blocks, with skip connections between blocks at the same hierarchical level as proposed in U-Net [44]. It has been noted that such deep architectures may wash away some necessary information from the input [15, 53], so we apply spatially-adaptive normalization (SPADE) [40] to address this problem. Motivated by their architecture, we replace the modulation parameters of batch normalization layers [19] in each decoder block with parameters learned from downsampled polarization images using simple, two-layer convolutional sub-networks. The details of our adaptations to the SPADE module are depicted in Figure 3. Lastly, we normalize the output estimated normal vectors to unit length, and apply the cosine similarity loss function:

$$L_{cosine} = \frac{1}{W \times H} \sum_{i}^{W} \sum_{j}^{H} (1 - \langle \hat{\boldsymbol{N}}_{ij}, \boldsymbol{N}_{ij} \rangle), \tag{7}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product, $\hat{\boldsymbol{N}}_{ij}$ is the estimated surface normal at pixel location (i, j), and $\boldsymbol{N}_{ij}$ is the corresponding ground truth surface normal. This loss is minimized when $\hat{\boldsymbol{N}}_{ij}$ and $\boldsymbol{N}_{ij}$ have identical orientation.

4 Dataset and Implementation Details

In what follows, we describe the dataset capture and organization as well as software implementation details. This is the first real-world dataset of its kind in the SfP domain, containing polarization images and corresponding ground truth surface normals for a variety of objects, under multiple different lighting condi-

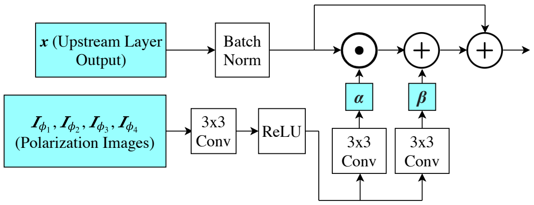
$\boldsymbol{x}$ (Upstream Layer Output)
Batch Norm
$\boldsymbol{I}_{\phi_1}, \boldsymbol{I}_{\phi_2}, \boldsymbol{I}_{\phi_3}, \boldsymbol{I}_{\phi_4}$ (Polarization Images)
3x3 Conv
ReLU
$\boldsymbol{\alpha}$
$\boldsymbol{\beta}$
3x3 Conv
3x3 Conv

Fig. 3. Diagram of SPADE normalization block. We use the polarization images to hierarchically inject back information in upsampling. The SPADE block, which takes a feature map $\boldsymbol{x}$ and a set of downsampled polarization images $\{\boldsymbol{I}_{\phi_1}, \boldsymbol{I}_{\phi_2}, \boldsymbol{I}_{\phi_3}, \boldsymbol{I}_{\phi_4}\}$ as the input, learns affine modulation parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The circle dot sign represents elementwise multiplication, and the circle plus sign represents elementwise addition.

tions. The Deep Shape from Polarization dataset can thus provide a baseline for future attempts at applying learning to the SfP problem.

4.1 Dataset

A polarization camera [29] with a layer of polarizers above the photodiodes (as described in Section 3.2) is used to capture four polarization images at angles $0°, 45°, 90°$ and $135°$ in a single shot. Then a structured light based 3D scanner [50] (with single shot accuracy no more than 0.1 mm, point distance from 0.17 mm to 0.2 mm, and a synchronized turntable for automatically registering scanning from multiple viewpoints) is used to obtain high-quality 3D shapes. Our real data capture setup is shown in Figure 4. The scanned 3D shapes are aligned from the scanner's coordinate system to the image coordinate system of the polarization camera by using the shape-to-image alignment method adopted in [49]. Finally, we compute the surface normals of the aligned shapes by using the Mitsuba renderer [20]. Our introduced dataset consists of 25 different objects, each object with 4 different orientations for a total of 100 object-orientation combinations. For each object-orientation combination, we capture images in 3 lighting conditions: indoors, outdoors on an overcast day, and outdoors on a sunny day. In total, we capture 300 images for this dataset, each with 4 polarization angles.[2]

4.2 Software implementation

Our model was implemented in PyTorch [41], and trained for 500 epochs with a batch size of 4. It took around 8 hours for the network to converge with a single NVIDIA GeForce RTX 2070. We used the Adam optimizer [24] with default parameters with a base learning rate of 0.01. We train our model on randomly cropped 256×256 image patches, which is relatively common in shape estimation tasks [38, 61] as a form of data augmentation. Further implementation details are in the supplement.

[2] The dataset is available at: https://visual.ee.ucla.edu/deepsfp.htm.

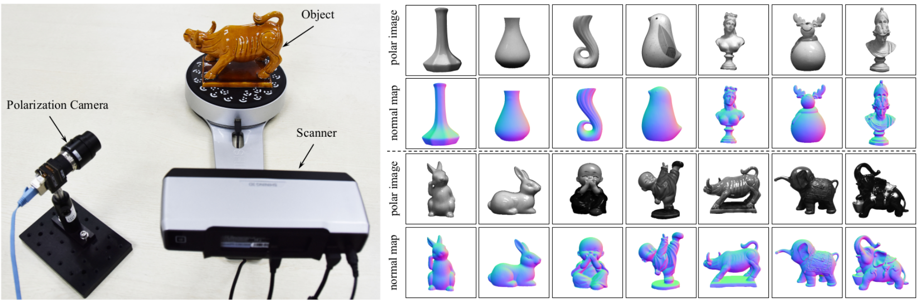

Fig. 4. This is the first dataset of its kind for the SfP problem. The capture setup and several example objects are shown above. We use a polarization camera to capture four gray-scale images of an object with four polarization angles in a single shot. The scanner is put next to the camera for obtaining the 3D shape of the object. The polarization images shown have a polarizer angle of 0 degrees. The corresponding normal maps are aligned below. For each object, the capture process was repeated for 4 different orientations (front, back, left, right) and under 3 different lighting conditions (indoor lighting, outdoor overcast, and outdoor sunlight).

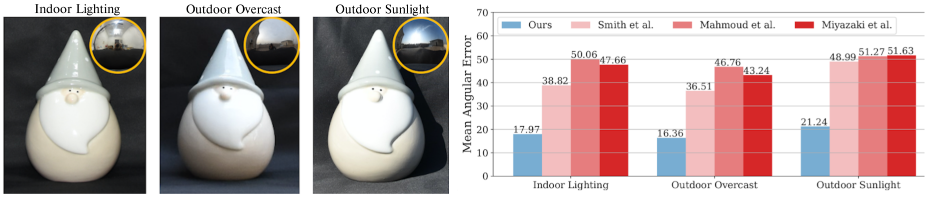

Fig. 5. The proposed method handles objects under varied lighting conditions. Note that our method has very similar mean angular error among all test objects across the three lighting conditions (bottom row). Please see supplement for further comparisons of lighting invariance.

5 Experimental Results

In this section, we evaluate our model with the presented challenging real-world scene benchmark, and compare it against three physics-only methods for SfP. All neural networks were trained on the same training data as discussed in Section 4.1. To quantify shape accuracy, we compute the widely used mean angular error (MAE) score on the surface normals.

5.1 Comparisons to physics-based SfP

We used a test dataset consisting of scenes that include BALL, HORSE, VASE, CHRISTMAS, FLAMINGO, DRAGON. On this test set, we implement three physics-based methods for SfP as a baseline: **1.** Smith *et al.* [52]. **2.** Mahmoud *et al.* [33]. **3.** Miyazaki *et al.* [37]. The first method recovers the depth map directly, and

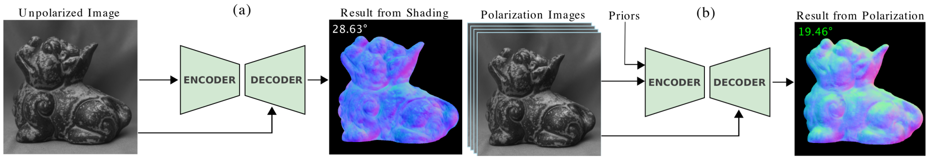

Fig. 6. Our network is learning from polarization cues, not just shading cues. An ablation study conducted on the Dragon scene. In (a) the network does not have access to polarization inputs. In (b) the network can learn from polarization inputs and polarization physics. Please refer to Figure 8, row c, for the ground truth shape of the Dragon.

we only use the diffuse model due to the lack of specular reflection masks. The surface normals are obtained from the estimated depth with bicubic fit. Both the first and the second methods require lighting input, and we use the estimated lighting from the first method during comparison. The second method also requires known albedo, and following convention, we assume a uniform albedo of 1. Note the method proposed in [37] is the same as that presented in [4]. We omit comparison with Tozza *et al.* [59], as it requires two unpolarized intensity images, with two different light source directions. To motivate a fair comparison, we obtained the comparison codes directly from Smith *et al.* [52]. [3]

5.2 Robustness to lighting variations

Figure 5 shows the robustness of the method to various lighting conditions. Our dataset includes lighting in three broad categories: (a) indoor lighting; (b) outdoor overcast; and (c) outdoor sunlight. Our method has the lowest MAE, over the three lighting conditions. Furthermore, our method is consistent across conditions, with only slight differences in MAE for each object between lightings.

5.3 Importance of polarization

An interesting question is how much of the shape information is learned from polarization cues as compared to shading cues. Figure 6 explores the benefit of polarization by ablating network inputs. We compare two cases. Figure 6(a) shows the resulting shape reconstruction when using a network architecture optimized for an unpolarized image input. The shape has texture copy and a high MAE of 28.63 degrees. In contrast, Figure 6(b) shows shape reconstruction from our proposed method of learning from four polarization images and a model of polarization physics. We observe that shape reconstruction using polarization cues is more robust to texture copy artifacts, and has a lower MAE of only 19.46 degrees. Although only one image is used in the shading network (as is typical for shape from shading), this image is computed using an average of the four polarization images. Thus the distinction between the two cases in Figure 6(a) and 6(b) is the polarization diversity, rather than improvements in photon noise.

[3] https://github.com/waps101/depth-from-polarisation

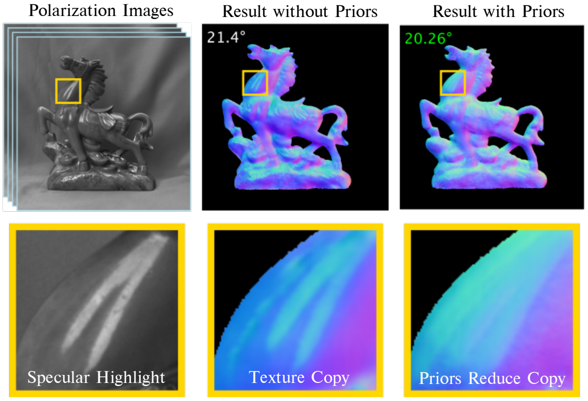

Fig. 7. Ablation test shows that the physics-based prior reduces texture copy artifacts. We see that the specular highlight in the input polarization image is directly copied into the normal reconstruction without priors. Note that our prior-based method shows stronger suppression of the copy artifact. Please see supplement for further examples of the effects of priors on texture copy.

5.4 Importance of physics revealed by ablating priors

Figure 7 highlights the importance of physics-based learning, as compared to traditional machine learning. Here, we refer to "traditional machine learning" as learning shape using only the polarization images as input. These results are shown in the middle column of Figure 7. Shape reconstructions based on traditional machine learning exhibit image-based artifacts, because the polarization images contain brightness variations that are not due to geometry, but due to specular highlights (e.g., the HORSE is shiny). Learning from just the polarization images alone causes these image-based variations to masquerade as shape variations, as shown in the zoomed inset of Figure 7. A term used for this is *texture copy*, where image texture is undesirably copied onto the geometry [21]. In contrast, the proposed results with physics priors are shown in the rightmost inset of Figure 7, showing less dependence on image-based texture (because we also input the geometry-based physics model).

5.5 Quantitative evaluation on our test set

We use MAE[4] to make a quantitative comparison between our method and the previous physics-based approaches. Table 2 shows that the proposed method has the lowest MAE on each object, as well as the overall test set. The two most

[4] MAE is the most commonly reported measure for surface normal reconstruction, but in many cases it is a deceptive metric. We find that a few outliers in high-frequency regions can skew the MAE for entire reconstructions. Accordingly, we emphasize the qualitative comparisons of the proposed method to its physics-based counterparts.

Table 2. Our method outperforms previous methods for each object in the test set. Numbers represent the MAE averaged across the three lighting conditions for each object. The best model is marked in magenta and the second-best is in blue.

Scene	**Proposed**	Smith [52]	Mahmoud [33]	Miyazaki [37]
Box	**23.31°**	**31.00°**	41.51°	45.47°
Dragon	**21.55°**	**49.16°**	70.72°	57.72°
Father Christmas	**13.50°**	39.68°	**39.20°**	41.50°
Flamingo	**20.19°**	**36.05°**	47.98°	45.58°
Horse	**22.27°**	55.87°	**50.55°**	51.34°
Vase	**10.32°**	**36.88°**	44.23°	43.47°
Whole Set	**18.52°**	**41.44°**	49.03°	47.51°

challenging scenes in the test set are the horse and the dragon. The former has intricate detail and specularities, while the latter has a mixed material surface. The physics-based methods struggle on these challenging scenes as all scenes have over 49 degrees of mean angular error. The method from Smith *et al.* [52] has the second-lowest error on the dragon scene, but the method from Miyazaki *et al.* [37] has the second-lowest error on the horse scene. On the overall test set, the physics-based methods are all clustered between 41.4 and 49.0 degrees, while the physics-based deep learning approach we propose achieves over a two-fold reduction in error to 18.5 degrees.

The reader may wonder why the physics-based methods perform poorly on tested scenes. The result from Smith *et al.* [52] assumes a reflection model and combinatorial lighting estimation, which do not appear to scale to unconstrained, real world environments, resulting in a normal map with a larger error. Mahmoud *et al.* [33] uses shading constraints that assume a distant light source, which is not the case for some of the tested scenes, especially the indoor ones. Finally, the large region-wise anomalies on many of the results from Miyazaki *et al.* [37] are due to the sensitive nature of their histogram normalization method.

5.6 Qualitative evaluation on our test set

Figure 8 shows qualitative and quantitative data for various objects in our test set. The RGB images in (row a) are not used as input, but are shown in the top row of the figure for context about material properties. The input to all the methods shown is four polarization images, shown in (row b) of Figure 8. The ground truth shape is shown in (row c), and corresponding shape reconstructions for the proposed method are shown in (row d). Comparison methods are shown in (row e) through (row g). It is worth noting that the physics-based methods particularly struggle with *texture copy* artifacts, where color variations masquerade as geometric variations. This can be seen in Figure 8, (row f), where the physics-based reconstruction of Mahmoud [33] confuses the color variation in the beak of the flamingo with a geometric variation. In contrast, our proposed method, shown in (row d), recovers the beak more accurately. Beyond texture copy, another limitation of physics-based methods lies in the difficulty of solving

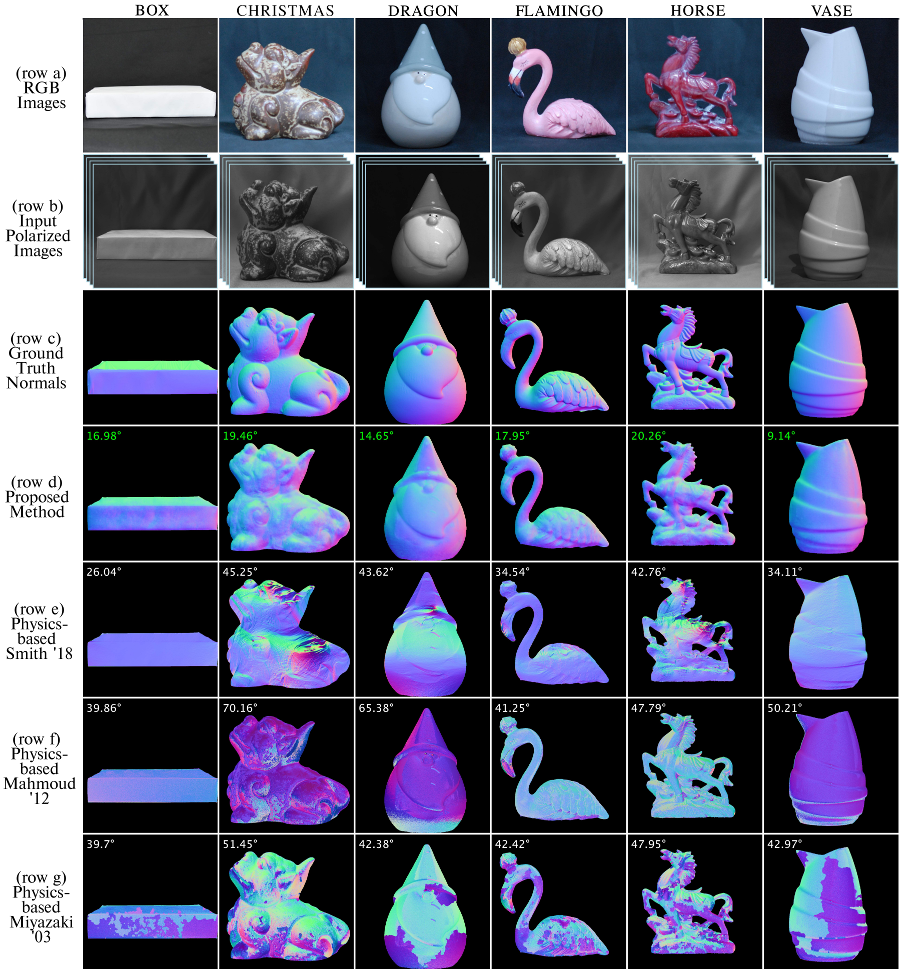

Fig. 8. The proposed method shows qualitative and quantitative improvements in shape recovery our test dataset. (row a) Shows the RGB scene photographs for context - these are not used as the input to any of the methods. (row b) The input to all methods are a stack of four polarization photographs at angles of 0°, 45°, 90°, and 135°(row c). The ground truth normals, obtained experimentally. (row d) The proposed approach for shape recovery. (row e-g) We compare with physics-based SfP methods by Smith *et al.* [51], Mahmoud *et al.* [33] and Miyazaki *et al.* [37]. (We omit the results from Atkinson *et al.* [4], which uses a similar method as [37]). Please see supplement for further comparisons.

the *ambiguity problem*, discussed earlier in this paper. In row g, the physics-based approach from Miyazaki *et al.* [37] has significant ambiguity errors. This can be seen as the fixed variations in color of normal maps, which are not due to random noise. Although less drastic, the physics-based method of Smith *et al.* [52] also shows such fixed pattern artifacts, due to the underdetermined nature of the problem. Our proposed method is fairly robust to fixed pattern error, and our deviation from ground truth is largely in areas with high-frequency detail. Although the focus of Figure 8 is to highlight qualitative comparisons, it is worth noting that the MAE in of the proposed method is the lowest for all these scenes (lowest MAE is highlighted in green font).

6 Discussion

In summary, we presented a first attempt re-examining SfP through the lens of deep learning, and specifically, physics-based deep learning. Table 2 shows that our network achieves over a two-fold reduction in shape error, from 41.4 degrees [52] to 18.5 degrees. An ablation test verifies the importance of using the physics-based prior in the deep learning model. In experiments, the proposed model performs well under varied lighting conditions, while previous physics-based approaches have either higher error or variation across lighting.

Future Work The framerate of our technique is limited both by the feed-forward pass, as well as the time required to calculate the physical prior (about 1 second per frame). Future work could explore parallelizing the physics-based calculations or using approximations for more efficient compute. As discussed in Section 5.5, the high MAE is largely due to a few regions with extremely fine detail. Finding ways to effectively weight these areas more heavily or add a refinement stage focused on these challenging regions, are promising avenues for future exploration. Moreover, identifying a metric better able to capture the quality of reconstructions than MAE would be valuable for continued study of learning-based SfP.

Conclusion We hope the results of this study encourage future explorations at the seamline of deep learning and polarization as well as the broader field of fusion of data-driven and physics-driven techniques.

Acknowledgements: The work of UCLA authors was supported by a Sony Imaging Young Faculty Award, Google Faculty Award, and the NSF CRII Research Initiation Award (IIS 1849941). The work of Peking University authors was supported by National Natural Science Foundation of China (61872012, 41842048, 41571432), National Key R&D Program of China (2019YFF0302902, 2017YFB0503004), Beijing Academy of Artificial Intelligence (BAAI), and Education Department Project of Guizhou Province.

References

1. Atkinson, G.A.: Polarisation photometric stereo. Computer Vision and Image Understanding (2017)
2. Atkinson, G.A., Ernst, J.D.: High-sensitivity analysis of polarization by surface reflection. Machine Vision and Applications (2018)
3. Atkinson, G.A., Hancock, E.R.: Multi-view surface reconstruction using polarization. ICCV (2005)
4. Atkinson, G.A., Hancock, E.R.: Recovery of surface orientation from diffuse polarization. IEEE TIP (2006)
5. Baek, S.H., Jeon, D.S., Tong, X., Kim, M.H.: Simultaneous acquisition of polarimetric SVBRDF and normals. ACM SIGGRAPH (TOG) (2018)
6. Berger, K., Voorhies, R., Matthies, L.H.: Depth from stereo polarization in specular scenes for urban robotics. ICRA (2017)
7. Chen, G., Han, K., Wong, K.Y.K.: PS-FCN: A flexible learning framework for photometric stereo. ECCV (2018)
8. Chen, L., Zheng, Y., Subpa-asa, A., Sato, I.: Polarimetric three-view geometry. ECCV (2018)
9. Cui, Z., Gu, J., Shi, B., Tan, P., Kautz, J.: Polarimetric multi-view stereo. CVPR (2017)
10. Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., Bousseau, A.: Single-image SVBRDF capture with a rendering-aware deep network. ACM SIGGRAPH (TOG) (2018)
11. Drbohlav, O., Sara, R.: Unambiguous determination of shape from photometric stereo with unknown light sources. ICCV (2001)
12. Ghosh, A., Chen, T., Peers, P., Wilson, C.A., Debevec, P.: Circularly polarized spherical illumination reflectometry. ACM SIGGRAPH (TOG) (2010)
13. Ghosh, A., Fyffe, G., Tunwattanapong, B., Busch, J., Yu, X., Debevec, P.: Multiview face capture using polarized spherical gradient illumination. ACM SIGGRAPH (TOG) (2011)
14. Guarnera, G.C., Peers, P., Debevec, P., Ghosh, A.: Estimating surface normals from spherical stokes reflectance fields. ECCV (2012)
15. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.: Deep networks with stochastic depth. CoRR (2016)
16. Huynh, C.P., Robles-Kelly, A., Hancock, E.R.: Shape and refractive index recovery from single-view polarisation images. CVPR (2010)
17. Huynh, C.P., Robles-Kelly, A., Hancock, E.R.: Shape and refractive index from single-view spectro-polarimetric images. IJCV (2013)
18. Ikehata, S.: CNN-PS: CNN-based photometric stereo for general non-convex surfaces. ECCV (2018)
19. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
20. Jakob, W.: Mitsuba renderer (2010), http://www.mitsuba-renderer.org
21. Kadambi, A., Taamazyan, V., Shi, B., Raskar, R.: Polarized 3D: High-quality depth sensing with polarization cues. ICCV (2015)
22. Kadambi, A., Taamazyan, V., Shi, B., Raskar, R.: Depth sensing using geometrically constrained polarization normals. IJCV (2017)
23. Karpatne, A., Watkins, W., Read, J., Kumar, V.: Physics-guided neural networks (PGNN): an application in lake temperature modeling. CoRR (2017)

24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
25. Li, X., Dong, Y., Peers, P., Tong, X.: Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. ACM SIGGRAPH (TOG) (2017)
26. Li, Z., Sunkavalli, K., Chandraker, M.: Materials for masses: SVBRDF acquisition with a single mobile phone image. ECCV (2018)
27. Li, Z., Xu, Z., Ramamoorthi, R., Sunkavalli, K., Chandraker, M.: Learning to reconstruct shape and spatially-varying reflectance from a single image. ACM SIGGRAPH Asia (TOG) (2018)
28. Lindell, D.B., O'Toole, M., Wetzstein, G.: Single-photon 3D imaging with deep sensor fusion. ACM SIGGRAPH (TOG) (2018)
29. Lucid Vision Phoenix polarization camera: https://thinklucid.com/product/phoenix-5-0-mp-polarized-model/ (2018)
30. Lyu, Y., Cui, Z., Li, S., Pollefeys, M., Shi, B.: Reflection separation using a pair of unpolarized and polarized images. In: Advances in Neural Information Processing Systems. pp. 14559–14569 (2019)
31. Ma, W.C., Hawkins, T., Peers, P., Chabert, C.F., Weiss, M., Debevec, P.: Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. Eurographics Conference on Rendering Techniques (2007)
32. Maeda, T., Kadambi, A., Schechner, Y.Y., Raskar, R.: Dynamic heterodyne interferometry. ICCP (2018)
33. Mahmoud, A.H., El-Melegy, M.T., Farag, A.A.: Direct method for shape recovery from polarization and shading. ICIP (2012)
34. Marco, J., Hernandez, Q., Munoz, A., Dong, Y., Jarabo, A., Kim, M.H., Tong, X., Gutierrez, D.: Deeptof: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. ACM SIGGRAPH (TOG) (2017)
35. Miyazaki, D., Kagesawa, M., Ikeuchi, K.: Transparent surface modeling from a pair of polarization images. PAMI (2004)
36. Miyazaki, D., Shigetomi, T., Baba, M., Furukawa, R., Hiura, S., Asada, N.: Surface normal estimation of black specular objects from multiview polarization images. International Society for Optics and Photonics, Optical Engineering (2016)
37. Miyazaki, D., Tan, R.T., Hara, K., Ikeuchi, K.: Polarization-based inverse rendering from a single view. ICCV (2003)
38. Mo, Z., Shi, B., Lu, F., Yeung, S.K., Matsushita, Y.: Uncalibrated photometric stereo under natural illumination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2936–2945 (2018)
39. Ngo, T.T., Nagahara, H., Taniguchi, R.: Shape and light directions from shading and polarization. CVPR (2015)
40. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. CVPR (2019)
41. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. NIPS-W (2017)
42. PolarM polarization camera: http://www.4dtechnology.com/products/polarimeters/polarcam/ (2017)
43. Riviere, J., Reshetouski, I., Filipi, L., Ghosh, A.: Polarization imaging reflectometry in the wild. ACM SIGGRAPH (TOG) (2017)
44. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. MICCAI (2015)

45. Santo, H., Samejima, M., Sugano, Y., Shi, B., Matsushita, Y.: Deep photometric stereo network. ICCV Workshops (2017)
46. Satat, G., Tancik, M., Gupta, O., Heshmat, B., Raskar, R.: Object classification through scattering media with deep learning on time resolved measurement. OSA Optics Express (2017)
47. Schechner, Y.Y.: Self-calibrating imaging polarimetry. ICCP (2015)
48. Sengupta, S., Kanazawa, A., Castillo, C.D., Jacobs, D.W.: SfSnet: Learning shape, reflectance and illuminance of faces in the wild. CVPR (2018)
49. Shi, B., Mo, Z., Wu, Z., Duan, D., Yeung, S.K., Tan, P.: A benchmark dataset and evaluation for non-Lambertian and uncalibrated photometric stereo. PAMI (2019)
50. SHINING 3D scanner: https://www.einscan.com/einscan-se-sp (2018)
51. Smith, W.A.P., Ramamoorthi, R., Tozza, S.: Linear depth estimation from an uncalibrated, monocular polarisation image. ECCV (2016)
52. Smith, W.A.P., Ramamoorthi, R., Tozza, S.: Height-from-polarisation with unknown lighting or albedo. PAMI (2018)
53. Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway networks. CoRR (2015)
54. Su, S., Heide, F., Wetzstein, G., Heidrich, W.: Deep end-to-end time-of-flight imaging. CVPR (2018)
55. Tancik, M., Satat, G., Raskar, R.: Flash photography for data-driven hidden scene recovery. arXiv preprint arXiv:1810.11710 (2018)
56. Tancik, M., Swedish, T., Satat, G., Raskar, R.: Data-driven non-line-of-sight imaging with a traditional camera. OSA Imaging and Applied Optics (2018)
57. Taniai, T., Maehara, T.: Neural inverse rendering for general reflectance photometric stereo. ICML (2018)
58. Teo, D., Shi, B., Zheng, Y., Yeung, S.K.: Self-calibrating polarising radiometric calibration. CVPR (2018)
59. Tozza, S., Smith, W.A.P., Zhu, D., Ramamoorthi, R., Hancock, E.R.: Linear differential constraints for photo-polarimetric height estimation. ICCV (2017)
60. Wolff, L.B.: Polarization vision: A new sensory approach to image understanding. Image Vision Computing (1997)
61. Xiong, Y., Chakrabarti, A., Basri, R., Gortler, S.J., Jacobs, D.W., Zickler, T.: From shading to local shape. IEEE transactions on pattern analysis and machine intelligence **37**(1), 67–79 (2014)
62. Yang, L., Tan, F., Li, A., Cui, Z., Furukawa, Y., Tan, P.: Polarimetric dense monocular SLAM. CVPR (2018)
63. Ye, W., Li, X., Dong, Y., Peers, P., Tong, X.: Single image surface appearance modeling with self-augmented cnns and inexact supervision. Wiley Online Library Computer Graphics Forum (2018)
64. Zhu, D., Smith, W.A.P.: Depth from a polarisation + RGB stereo pair. CVPR (2019)