Transparent Shape from a Single View Polarization Image

Mingqi Shao Chongkun Xia Zhendong Yang Junnan Huang Xueqian Wang*
Tsinghua Shenzhen International Graduate School
{smq21, yangzd21, hjn21}@mails.tsinghua.edu.cn {xiachongkun, wang.xq}@sz.tsinghua.edu.cn

Abstract

This paper presents a learning-based method for transparent surface estimation from a single view polarization image. Existing shape from polarization(SfP) methods have the difficulty in estimating transparent shape since the inherent transmission interference heavily reduces the reliability of physics-based prior. To address this challenge, we propose the concept of physics-based prior confidence, which is inspired by the characteristic that the transmission component in the polarization image has more noise than reflection. The confidence is used to determine the contribution of the interfered physics-based prior. Then, we build a network(TransSfP) with multi-branch architecture to avoid the destruction of relationships between different hierarchical inputs. To train and test our method, we construct a dataset for transparent shape from polarization with paired polarization images and ground-truth normal maps. Extensive experiments and comparisons demonstrate the superior accuracy of our method. Our cdataset and code are publicly available at https://github.com/shaomq2187/TransSfP

1. Introduction

Surface normals provide detailed 3D information about the surface of objects. However, estimating the high-quality surface normal of transparent objects is still an open challenge. The complex rays interactions in transparent objects lead to the difficulty of standard 3D sensors in acquiring accurate surface information[23]. Besides, transparent objects lack texture of their own, adopting instead the texture of the background, making the regular image-based methods extremely ill-posed and can not produce satisfied estimation[12]. Compared with regular 3D and RGB sensors, the polarization sensor can acquire information about the surface normal from the reflected light on transparent surface. In particular, the commercial polarization image sensors[20] that appeared in recent years have made it possible to capture polarization images at a single shot, making the acquisition of polarization images as easy as conventional sensors. Hence, in this paper, we focus on estimating the surface normals of transparent objects from a single-view polarization image.

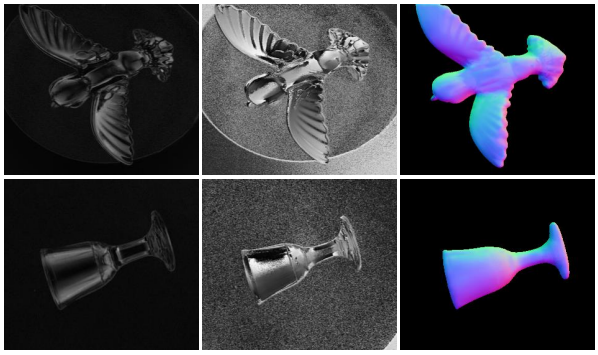

Figure 1. **Examples of our estimation.** Our method is capable of estimating the normal map of various transparent surface from a single view polarization image

The key barrier to using polarization information to estimate the surface normals of transparent objects is the existence of transmission interference. Fresnel specular reflection model(will be described in Sec.3.2) is commonly used to solve the possible surface normals, i.e. the physics-based prior in SfP, from the observed polarization state. However, due to the high transmittance of transparent objects, the specular reflection energy only accounts for about 4% of the incident light energy[26]. The interference transmitted from the background is superimposed on the observed polarization information, leading to the high estimation error of the physics-based SfP methods.

Aiming at the problem that the reliability of the Fresnel specular reflection model is reduced, we find that the polarization observation of the transmission component has more noise. This characteristic is helpful to distinguish the interfered areas and motivates us to propose the concept of physics-based prior confidence. The confidence indicates the reliability of the physics-based prior and hence serves as the weight of physics-based prior and polarization angle loss. The polarization angle loss is designed to force

*Corresponding author

our network to learn the prior knowledge defined by Fresnel specular reflection model. In addition, we notice that previous learning-based methods[2, 5] directly concatenating physics-based prior and raw polarization images will destroy the relationship between different hierarchies and lead to a decline in the performance. Therefore, we build a network called *TransSfP* with multi-branch architecture to handle different hierarchical inputs. To demonstrate the performance of our method, we construct a dataset for transparent shape polarization, which consists of both real-world and synthetic collections. Benefiting from the above designs, our method can estimate normal maps of various transparent surfacess with lower error as shown in Fig.1. We summarize our contributions as:

- We propose a novel learning-based method for transparent surface normal estimation from a single view polarization image.
- We contribute the first dataset for transparent shape from polarization, which consists of both real-world and synthetic collections.
- Technically, we introduce three novel designs for transparent SfP problem: the concept of physics-based prior confidence, a multi-branch architecture, and a self-supervised polarization angle loss.

2. Related Work

2.1. Shape from Polarization

When a beam of unpolarized incident light is reflected from the surface of an object, the polarization state of the light will change according to the incident angle, the relative refractive index, and the surface normal vector of the surface. Early works relied on the physical model of Fresnel equations to recover surface normals. Due to the inherent ambiguity of SfP, other constraints were required to determine a unique surface normal, including rotation measurements[15], active lighting[18], boundary prior[17] and shading cues[13], but these methods are limited due to strict assumptions and calibrations. With the vigorous development of deep learning in many areas of computer vision, the combination of SfP and deep learning has attracted the attention of researchers. Ba et al.[2] first proposed using deep learning to solve the SfP problem and established a real-world object-level dataset. Some works extended 'Sfp + deep learning' to other tasks, such as scene-level normal estimation[9] and human body shape estimation[31]. However, these approaches and datasets all focus on opaque objects. As far as we know, there are no publicly available polarization datasets and deep learning models for transparent objects. To address this issue, we propose the first dataset and deep learning model for transparent shape from polarization.

2.2. Shape Estimation for Transparent Objects

Shape estimation for transparent objects is a challenging problem[27]. Active lighting and capture devices are used in many methods to estimate the shape of transparent objects, such as shape from the distortion of the calibration pattern by transparent objects[25], shape from the reflection of transparent objects to known ambient light[28], shape from the corresponding relationship between the incident and reflected light[27]. These methods all require expensive experimental equipment and tedious reconstruction steps. Some researchers tried to use cheaper devices to estimate the surface shape of transparent objects. Li et al.[12] proposed a physics-based neural network, which can reconstruct 3D models of transparent objects from multi-view RGB images and achieve state-of-the-art results. These methods exploited multi-view images to achieve charming results, while our method aims to recover transparent shape from a single view image. There are some works focusing on RGB-D sensor depth completion[23, 30], which can complete the depth map of scenes containing transparent objects from a single view. But these works usually focused on transparent object grasping, their estimation accuracy is limited.

Shape from polarization has long been used to estimate the surface of transparent objects with the advantages of single-view, weak assumption of lighting, and passive imaging. Saito et al.[22] applied Fresnels's law to solve surface orientation by measuring the polarization information of the transparent object in an optical diffuser. To solve the ambiguity in SfP, Miyazaki et al.[16] proposed a method of rotating the transparent object with a tiny angle and matching the feature points of the two polarization states. Since the polarization state of transparent objects is seriously affected by internal reflection, refraction, and transmission, Miyazaki et al.[14] employed inverse raytracing to optimize transparent shapes. However, prior works only focused on simple transparent objects such as spheres and plates. We hope to estimate surface shape of more complex transparent objects by combining SfP with deep learning.

3. Method

3.1. Overview

Our goal is to estimate transparent surface normal from a single view polarization image. To this end, we present a data-driven approach as shown in Fig.2. We input the raw polarization images to our network, which has three components: intensity map I, degree of linear polarization(DoLP, ρ) map, and angle of linear polarization(AoLP, φ) map. Then, we compute the physics-based prior, the four normal maps $N_{phy0}, N_{phy1}, N_{phy2}, N_{phy3}$, by exploiting Fresnel's specular reflection model(Sec.3.2)) and they are input into our network. To handle the transmission interference, we

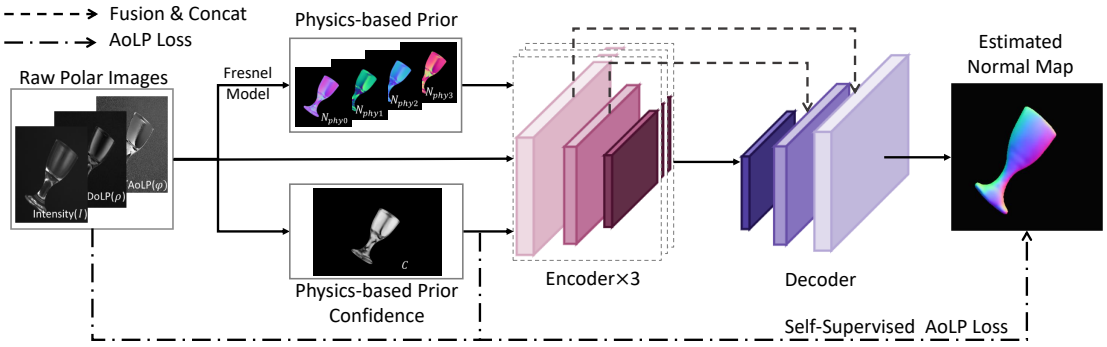

Figure 2. **Overview of our proposed method.** We use a multi-branch architecture to handle different hierarchical inputs rather than directly concatenating them. The physics-based prior confidence map is defined from raw polarization images and used for the weighted fusion of different inputs. In addition, we propose the self-supervised AoLP loss by exploiting the confidence map to force the network to learn the prior knowledge between the normal and AoLP

analyze the polarization observation characteristic on transparent surface and define the concept of the physics-based prior confidence C as an additional input(Sec.3.3). In addition, we propose a self-supervised AoLP loss based on the confidence to force the network to learn the prior knowledge between AoLP and surface normal(Sec.3.4). Finally, we propose separately feeding the different three inputs into the network and the estimated surface normal $\hat{\mathbf{n}}$ can be expressed as:

$$\hat{\mathbf{n}} = f(\rho, \varphi; N_{phy0}, N_{phy1}, N_{phy2}, N_{phy3}; C) \quad (1)$$

where $f(\cdot)$ represents the prediction model(Sec.3.4). In our method, we used the following assumptions: (a) Transparent surface is smooth and its refractive index is known. (b) The noise in AoLP map stems from the background, and other sources are ignored. (c) The reflection component is dominant on most areas of the transparent surface.

3.2. Polarization Image Formation of Transparent Shapes

The ray coming from the transparent surface is the superposition of the reflection component(I_r, ρ_r, φ_r) and the transmission component(I_t, ρ_t, φ_t). Hence, the intensity of a single pixel under a polarizer with an angle of α_{pol} can be expressed as follows:

$$\begin{aligned} I(\alpha_{pol}) =& I_r[1 + \rho_r \cos(2\alpha_{pol} - 2\varphi_r)] + \\ & I_t[1 + \rho_t \cos(2\alpha_{pol} - 2\varphi_t)] \quad (2) \\ =& I[1 + \rho \cos(2\alpha_{pol} - 2\varphi)] \quad (3) \end{aligned}$$

According to Fresnel's equations, the DoLP and AoLP in the specular reflection component have the following relationship with the zenith θ and azimuth angle ϕ of surface normal $\mathbf{n} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$:

$$\rho_r = \frac{2\sin^2\theta\cos\theta\sqrt{\eta^2 - \sin^2\theta}}{\eta^2 - \sin^2\theta - \eta^2\sin^2\theta + 2\sin^4\theta} \quad (4)$$

$$\varphi_r = \phi \pm \frac{\pi}{2} \quad (5)$$

where η is the refraction index and is set to 1.52 in this paper. Eq.4 and Eq.5 show that two zenith and two azimuth angles can be determined for a given DoLP and AoLP. Four normal maps, i.e. the physics-based prior, can be calculated from the raw polarization images. For detailed calculations of physics-based prior please refer to our supplementary material.

Since the polarization state of the transmission component is unknown, the acquisition of the physics-based prior implies the following approximations: $\rho \approx \rho_r$ and $\varphi \approx \varphi_r$, where ρ and φ are the captured DoLP and AoLP. The error of physics-based prior depends on the rationality of the approximations. To understand the impact of the transmission component, we solve ρ and φ and introduce the transmission coefficient T to strip the variables in I_r and I_t:

$$\rho = \frac{\sqrt{I_{r0}^2(1-T)^2\rho_r^2 + I_{t0}^2T^2\rho_t^2 + 2I_{r0}I_{t0}\rho_r\rho_tT(1-T)\cos(2\varphi_r - 2\varphi_t)}}{I_{r0}(1-T) + I_{t0}T} \quad (6)$$

$$\varphi = \frac{1}{2}\arctan\frac{I_{r0}\rho_r(1-T)\sin 2\varphi_r + I_{t0}\rho_tT\sin 2\varphi_t}{I_{r0}\rho_r(1-T)\cos 2\varphi_r + I_{t0}\rho_tT\cos 2\varphi_t} \quad (7)$$

where I_{r0} and I_{t0} can be regarded as constant values and satisfy the following relationship with I_r and I_t: $I_r = (1-T)I_{r0}$, $I_t = TI_{t0}$. Eq.6 and Eq.7 illustrate the formation of the polarization state on the transparent surface. The ρ_t is usually small if the background is diffuse and the $\rho_r \gg \rho_t$ since the specular reflection is highly polarized[6]. Hence, the interference of transmission on the DoLP ρ is weak.

Our experimental setting can ensure $\frac{I_{r0}}{I_{t0}} \approx 10$. Therefore, according to Eq.7, the transmission term interference

is significant only when T is very close to 1. In other words, φ will be disturbed by φ_t heavily when $T \to 1$; otherwise, $\varphi \approx \varphi_r$ is satisfied. This characteristic inspires us to propose the physics-based prior confidence and the AoLP loss.

3.3. Physics-based Prior Confidence

Using the physics-based prior in the areas with high transmittance will lead to a significant angular error. Although we can not directly obtain the transmittance of each point, fortunately, the characteristic of the transmission component's AoLP φ_t can help us to distinguish the areas with strong transmission interference.

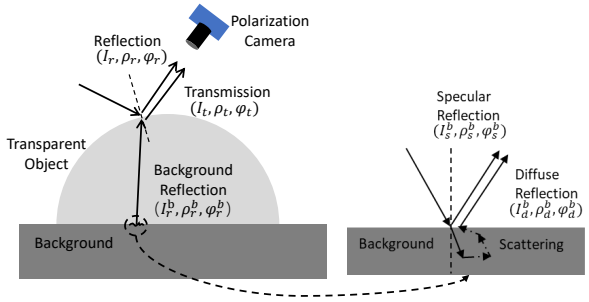

Figure 3. **Different components captured by polarization camera.** The captured ray consists of reflection and transmission components, where the transmission component is composed of two kinds of reflections from the rough background

Compared with the smooth transparent surface, the rays reflected from the rough background I_r^b contain specular I_s^b and diffuse reflection component I_d^b as shown in Fig.3, which can be expressed as follows:

$$\begin{aligned} I_r^b(\alpha_{pol}) =& I_s^b[1+\rho_s^b\cos(2\alpha_{pol}-2\varphi_s^b)]+ \\ & I_d^b[1+\rho_d^b\cos(2\alpha_{pol}-2\varphi_d^b)] \end{aligned} \tag{8}$$

$$= I_r^b[1+\rho_r^b\cos(2\alpha_{pol}-2\varphi_r^b)] \tag{9}$$

The shape from polarization theory gives us that the AoLPs of specular reflection and diffuse reflection are orthogonal[1]:

$$\varphi_s^b = \varphi_d^b \pm \frac{\pi}{2} \tag{10}$$

Then, we can solve for φ_t in terms of $I_s^b, \rho_s^b, \varphi_s^b, I_d^b, \rho_d^b$:

$$\varphi_r^b = \varphi_s^b - \frac{1-\mathrm{sign}(I_s^b\rho_s^b - I_d^b\rho_d^b)}{2}\pi \tag{11}$$

It can be seen that there is a sign function in the φ_r^b expression, which will lead to the discontinuity in the AoLP map. For the rough background, both specular and diffuse reflection exists and their dominance can not be determined, resulting in the noise in AoLP map. In fact, the polarization state of I_r^b will change according to the incidence angle θ_i of the interaction point when the ray passes through the transparent object. We denote this process as $\mathrm{M_t}(\cdot)$ and the following expression can be obtained:

$$I_t(\alpha_{pol}) = \mathrm{M_t}(I_r^b(\alpha_{pol}); \theta_i) \tag{12}$$

The incident angles of rays from a local area are approximately equal, which means that they will gain the same phase shift from $\mathrm{M_t}(\cdot)$. Therefore, the noise in φ_r^b also exists in φ_t. According to the above analysis, the AoLP observation of transparent surface has the characteristic that the area with higher transmittance contains more noise as shown in Fig.4.

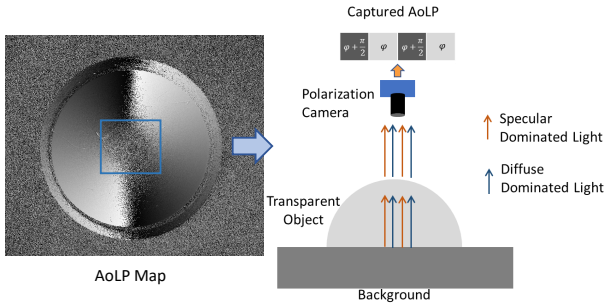

Figure 4. **The characteristic in the AoLP map of the transparent surface.** The noise in the AoLP map is derived from the transition between specular dominance and diffuse dominance on the background surface. More noise indicates higher transmittance and also lower reliability of physics-based prior

Above analysis illustrates that the noise in the AoLP map is related to the transmittance and physics-based prior reliability. Therefore, we define the concept **Physics-based Prior Confidence** by quantifying the noise in the AoLP map. The noise for a given pixel can be quantified by the distance between the pixel and its neighborhood pixels. We first define the distance $d_{K,m}(i,j)$ between the pixels in the $K \times K$ neighborhood of point (i,j):

$$d_{K,m}(i,j) = \sum_{p \in P_{i,j}} |p - \bar{p}_{i,j}|^m \tag{13}$$

where $P_{i,j}$ represents the set of pixel values in the $K \times K$ neighborhood of point (i,j), $\bar{p}_{i,j}$ is the mean of pixel values belonging to the set $P_{i,j}$ and m is the smoothing exponential term. K determines the consistency between the AoLP map and confidence map and its value in this paper is set to 9. The parameter m controls the mapping relationship between noise density and the value of confidence and its default value is 0.5. By normalizing the distance map, the physics-based prior confidence can be obtained:

$$C_{K,m}(i,j) = 1 - \frac{d_{K,m}(i,j)}{\max\limits_{0 \le x < W, 0 \le y < H} d_{K,m}(x,y)} \tag{14}$$

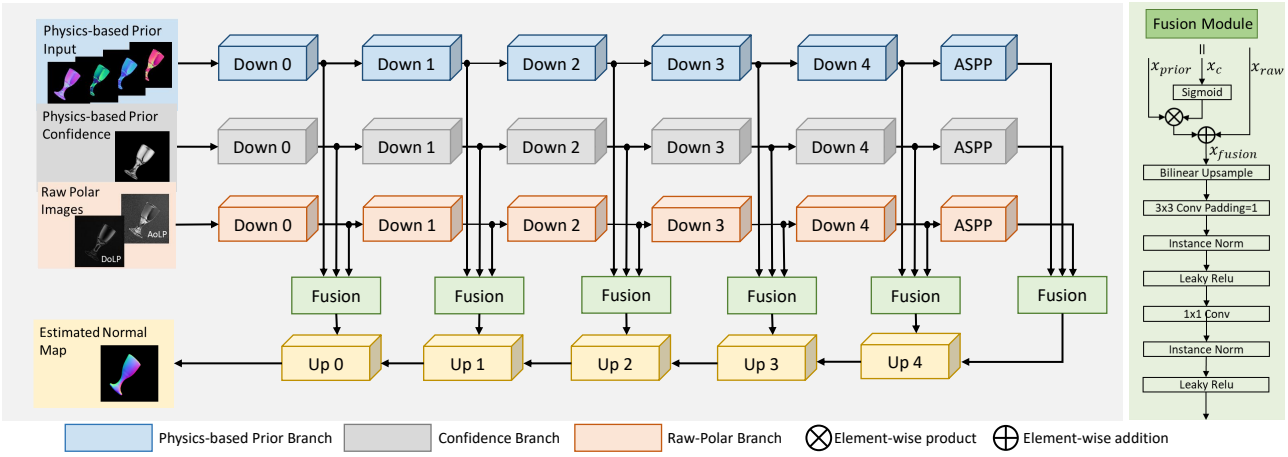

Figure 5. **Overview of TransSfP.** We regard raw polar images, physics-based prior, and confidence map as different hierarchical information and input them into three independent encoders. The confidence branch is used for weighting the features from physics-based prior branch, which is completed in our proposed fusion module

where H, W denote the height and width of the AoLP map respectively. The physics-based prior confidence defined by the above equation is also low at the junction of 0 and π in the AoLP map (according to Eq.3, the value range of AoLP is $[0, \pi]$). This situation is allowed because the lower confidence of the physics-based prior in such area is helpful to avoid the network being misguided by the jumped physics-based prior(see additional results which are provided in our supplementary material).

We first input the physics-based prior confidence map as an addition prior into the network and guide the fusion of raw polar images and physics-based prior to minimize prediction error. Then it is also used as the weight of the self-supervised AoLP loss.

3.4. Network and Optimization

A straightforward way to use deep learning to solve the SfP problem is to concatenate the raw polarization images and physics-based prior into the network for normal estimation[2, 9]. Different with previous methods, we regard transparent SfP as a multi-modal fusion problem, using separate encoders to extract features of the different inputs. Features from the raw polar branch and physics-based prior branch are fused at multiple levels according to the output of the confidence branch, and finally, a decoder is employed to complete the normal estimation.

As shown in Fig.5, our network *TransSfP* consists of three encoders and one decoder. The encoders are the physical prior branch, the physics-based prior confidence branch, and the raw polarization branch respectively. The structures of the encoders are completely the same, but the parameters are independent of each other. EPSANet50(Small) is adopted as the backbone of the Encoder, which is an improved backbone that replaces the Bottleneck in ResNet50 by ESPANet Block which exploits channel attention mechanism[29]. At the beginning of the Encoder, the module named 'Down 0' consisting of a 1×1 convolutional layer and a BatchNorm layer is used to force the network to extract the per-pixel features. At the end of the Encoder, we add an ASPP(Atrous Spatial Pyramid Pooling) module to enhance the network's ability to capture multi-scale information[4]. The detailed structure of each module is provided in the supplementary material.

To mitigate the adverse contribution of interfered physics-based prior, we propose a fusion module as shown in the right of Fig.5. Features from the confidence branch x_c are first converted to weights in the range $(0, 1)$ by passing through a Sigmoid function and then multiply with features from the physics-based prior branch x_{prior} to control the contribution of physics-based prior. Then, the U-Net style skip connection[21] is employed to introduce the fused features of different levels into the upsampling modules.

Lastly, we adopt the following loss function to optimize our network:

$$\mathcal{L}_{net} = \mathcal{L}_{sim} + \lambda \mathcal{L}_{aolp} \tag{15}$$

where $\mathcal{L}_{sim}$ is the cosine similarity loss and $\mathcal{L}_{aolp}$ is the AoLP loss. The default value of λ is set to 0.05.

The cosine similarity loss is commonly used in normal estimation, and its expression is as follows:

$$\mathcal{L}_{sim} = \sum_{i=0}^{W} \sum_{j=0}^{H} \left(1 - \frac{\mathbf{n}_{i,j} \cdot \hat{\mathbf{n}}_{i,j}}{\|\mathbf{n}_{i,j}\|_2 \, \|\hat{\mathbf{n}}_{i,j}\|_2}\right) \tag{16}$$

where $\mathbf{n}_{i,j}$ represents the ground-truth normal vector at the point (i, j), and $\hat{\mathbf{n}}_{i,j}$ is the estimated normal vector.

We propose the AoLP loss to force the network to learn the physical knowledge represented by Eq.5. Due to the

existence of transmission interference, Eq.5 can not be accurately satisfied. Hence, we take the physics-based prior confidence as the weight of the AoLP error to mitigate this interference:

$$\mathcal{L}_{aolp} = \sum_{i=0}^{W}\sum_{j=0}^{H} c_{i,j} \min(|\varphi_{i,j} + \frac{\pi}{2} - \hat{\phi}_{i,j}|, |\varphi_{i,j} - \frac{\pi}{2} - \hat{\phi}_{i,j}|) \tag{17}$$

where $c_{i,j}$ is the value of point (i, j) in the confidence map and $\hat{\phi}_{i,j}$ represents the azimuth of $\hat{\mathbf{n}}_{i,j}$. $\mathcal{L}_{aolp}$ is a self-supervised loss term since only AoLP and confidence map are used in the computation.

3.5. Dataset

We establish a dataset for transparent shape from polarization, which contains raw polarization images, physics-based prior(four normal maps), ground-truth masks and normal maps. Compared with prior polarization datasets[2, 9, 8], our dataset is the first dataset for transparent shape from polarization.

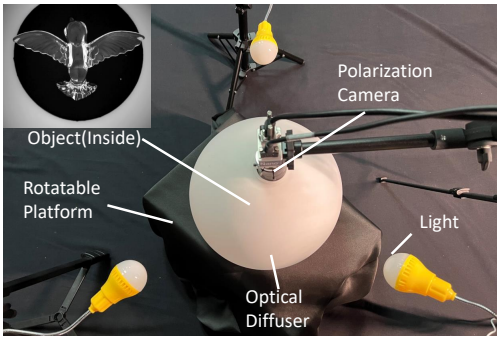

Figure 6. **Setup for acquiring real-world dataset**

Our dataset consists of two parts: the real-world dataset and the synthetic dataset. Fig.6 is our real-world setup for acquiring our dataset. As discussed before, the polarization information of the transmission component will bring adverse effect to normal estimation, hence, our dataset acquisition setup should increase the reflection and reduce the transmission as much as possible. An optical diffuser made of frosted glass with diameter of $30cm$ is used to simulate global illumination so that the transparent surface has strong and uniform reflections in all directions. To reduce the transmission from the background, the transparent object is placed on rough background. A DLASA G3-GM14-M2450 polarization camera is employed as our capture device, which can capture four polarization images(at θ_{pol} of $0°, 45°, 90°, 135°$) at a single shot. We coate transparent objects with powder and reconstructed their 3D models by using a 3D scanner with an accuracy of $0.1mm$, and then manually align them with the captured polarization images in the unity game engine to obtain the ground-truth mask and normal maps. Using the above steps, we build the real-world dataset of 10 different objects, each of which is rotated between $0°$ and $360°$ at $5°$-$10°$ intervals, resulting in a total of 486 samples.

The synthetic dataset is created since the real-world data acquisition requires manual alignment, resulting in the difficulty of acquiring enough real-world data for training. We establish a scene similar to the setup in Fig.6 and implement an integrator capable of outputting a camera-space normal map in the mitsuba2[19], a physically-based render that can track the full polarization state of light during a simulation. To simulate the diffuser's uniform lighting, we attach the diffuser's surface with an area light source. The background surface uses the diffuse BSDF with the reflectance of 0.1, and the transparent object adopts the dielectric BSDF with the refractive index of 1.52. The 3D models of 13 objects, collected from DiLiGenT-MV dataset[11] and Stanford 3D Scanning Repository[10], are placed in the scene we build to run polarization rendering. Each object is rotated 72 times from $0°$ to $360°$, resulting in a total of 936 samples in the synthetic dataset.

3.6. Implementation Details

We implement our model on PyTorch and the model is trained on an NVIDIA GeForce RTX 3090 GPU(24GB) with a batch size of 5. The Adam optimizer[7] with an initial learning rate of 1e-6 and a weight decay coefficient of 5e-4 is used for optimizing the network. We adopt the Step learning rate decreasing strategy, the learning rate times 0.1 every 9 epochs. To reduce memory usage, we crop the images to 512×512 patches in the data augmentation stage.

4. Experiments

4.1. Experimental Setup

We use the evaluation metrics commonly used in normal estimation, including mean angular error(*mean, MAE*), median angular error(*median*), and *Accuracy* $11.25°$, $22.5°$, $30°$, which represent the ratio of the number of pixels with an error lower than this value to the total number of valid pixels. All samples in the synthetic dataset are used for training, and some objects in the real-world dataset are also employed for training to make up for the difference between the synthetic and the real-world collections.

4.2. Comparisons to Baselines

In this paper, we aim to recover the transparent surface normals from polarization images in a single view. Hence, we employ two physics-based SfP methods, Miyazaki et al.[17], Mahmoud et al.[13], and three learning-based SfP methods, DeepSfP[2], Kondo et al.[8], SfP in the

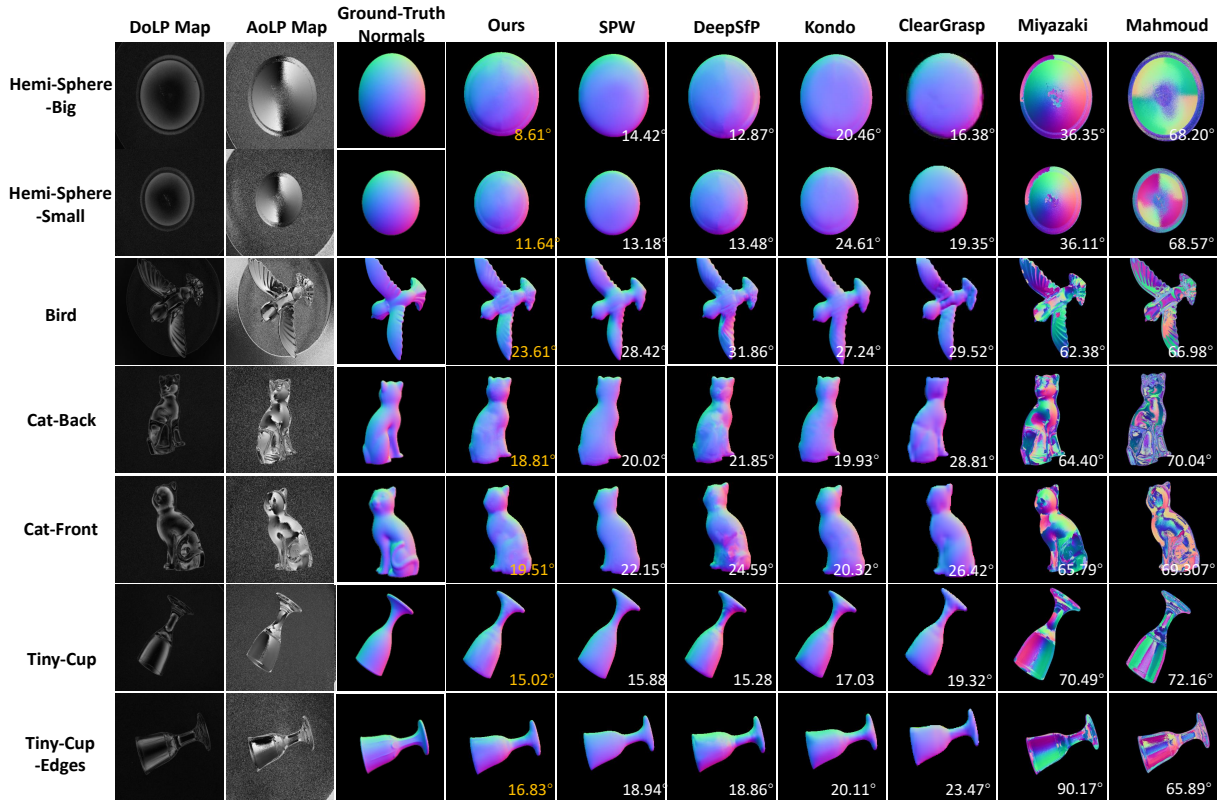

Figure 7. **Visual results of the comparisons to baselines.** Our method outperforms all baselines on all test objects

Table 1. **Quantitaive results of the comparisons to baselines.** The best results in the table are marked in **Bold**

Method	Mean Angular Error ↓							
	Hemi-Sph -Big	Hemi-Sph -Small	Bird	Cat-Back	Cat-Front	Tiny-Cup	Tiny-Cup -Edges	All
Miyazaki[17]	36.11°	36.35°	62.38°	64.40°	65.79°	70.49°	90.17°	55.62°
Mahmoud[13]	68.57°	68.20°	66.98°	70.04°	69.30°	72.16°	65.89°	68.72°
ClearGrasp[23]	19.35°	16.38°	29.52°	28.81°	26.42°	19.32°	23.47°	23.32°
DeepSfP[2]	13.48°	12.87°	31.86°	21.85°	24.59°	15.28°	18.86°	19.83°
SPW[9]	13.18°	14.42°	28.42°	20.02°	22.15°	15.88°	18.94°	19.00°
Kondo[8]	24.61°	20.46°	27.24°	19.93°	20.32°	17.03°	20.11°	21.39°
Ours	**8.61°**	**11.64°**	**23.61°**	**18.81°**	**19.51°**	**15.02°**	**16.83°**	**16.29°**

Wild(SPW)[9] as the SfP baselines of our method. In addition, ClearGrasp[23], whose normal prediction module can estimate the normals of transparent shapes using a single RGB image, is also employed as one of the baselines. We retained the learning based-SfP methods on our dataset. For the ClearGrasp, we performed fine-tuning on our dataset for the model pre-trained on the ClearGrasp dataset. Mahmoud's method needs to provide albedo and light source direction, we assume the albedo is uniform of 1, and the light source direction is estimated using the method proposed by Smith et al.[24].

Fig.7 lists the visual results of the comparison with baselines, Table 1 lists the quantitative results of the comparison. Our method achieves the best performance on all objects in the test set. A notable result is the comparison between ClearGrasp and other learning-based SfP methods. Though ClearGrasp uses large-scale synthesized RGB images for pretraining, its performance is still worse than any other learning method using polarization information, which illustrates the advantages of polarization sensor over regular RGB sensor.

4.3. Comparisons of Network Architectures

TransSfP uses the multi-branch architecture as shown in Fig.5. To illustrate the effectiveness of our architecture, we conduct the comparison with several commonly used semantic segmentations networks such as U-Net[21], DeepLabV3+[4] and models for normal estimation: PS-FCN[3], DeepSfP[2], SPW[9], Kondo[8] and single branch version of *TransSfP*. Our multi-branch architecture inputs the raw-polar input, physics-based prior input, and transmission confidence map separately, while other architectures concatenate them directly.

Table 2. **Comparisons of network architectures**

Network	Angular Error ↓		Accuracy ↑		
	Mean	Median	11.25°	22.5°	30°
U-Net[21]	25.20°	22.90°	16.06%	49.42%	69.07%
PS-FCN[3]	25.93°	22.42°	22.06%	52.79%	71.12%
DeepLabV3+[4]	21.90°	18.93°	21.04%	63.78%	81.84%
DeepSfP[2]	19.00°	14.91°	38.36%	77.36%	87.48%
SPW[9]	21.73°	18.43°	28.38%	61.24%	84.00%
Kondo[8]	24.40°	22.20°	18.95%	54.39%	75.09%
Ours(Single Branch)	18.73°	15.28°	33.72%	77.67%	88.65%
Ours	**16.29°**	**12.85°**	**48.31%**	**83.20%**	**93.90%**

The raw polarization information and the physics-based prior are converted through Frenel's physical model, hence they can be regarded as different hierarchies of polarization information. Direct concatenation will destroy the relationship between different hierarchies. The quantitative results in Table 2 prove that our consideration of treating them as different hierarchical information and then transforming them into a multi-modal fusion problem is effective.

Table 3. **Ablation experiment results.** The module with '*' indicates that the single branch architecture is used

Studied Module	Angular Error ↓		Accuracy ↑		
	Mean	Median	11.25°	22.5°	30°
W/o Polarization*	24.70°	22.99°	18.32%	52.84%	74.25%
With Polarization*	18.73°	15.29°	33.72%	77.67%	91.65%
W/o Confidence	17.72°	14.47°	40.02%	81.08%	92.06%
W/o AoLP Loss	17.17°	14.11°	41.81%	82.06%	93.14%
Full	**16.29°**	**12.85°**	**48.31%**	**83.20%**	**93.90%**

4.4. Ablation Studies

Importance of Polarization. In Section.4.2, the performance of ClearGrasp using RGB images is worse than any learning-based SfP method, which preliminarily illustrates the importance of polarization. To avoid the influence of network architecture, we conducted a further ablation experiment. We adopt the single branch architecture mentioned in Section.4.3, using light intensity and polarization information as input, respectively. The results listed in Table 3 show that the performance of using polarization information is much better than using only intensity since

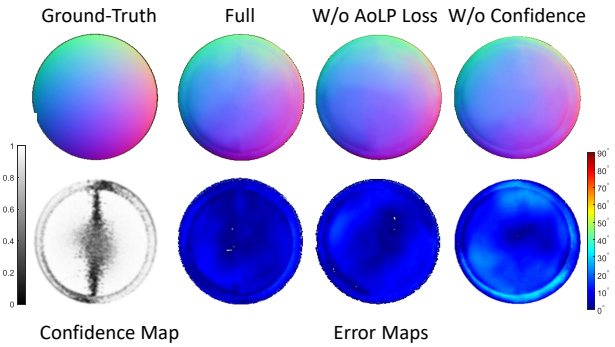

Figure 8. **Importance of AoLP loss and prior confidence**

polarization cues contain more information about the surface normal. Both the physical insights and the ablation study prove the superiority of the polarization sensor in the surface estimation of transparent objects.

Importance of Physics-based Prior Confidence. We set the confidence to constant 1 to remove the confidence branch. The quantitative and qualitative results in Fig.8 and Table 3 show that our proposed prior confidence is beneficial to transparent shape from polarization. The physics-based prior defined by the Eq.4 and Eq.5 provides rich information for normal estimation while the prior of some areas may have adverse effects due to the transmission interference. The confidence we proposed forces the network more conservative to use the physics-based prior of these areas and hence can reduce the adverse contribution significantly.

Importance of AoLP Loss. Fig.8 shows that the AoLP loss can reduce the error in areas with high confidence(or weak transmission) because the approximation of $\varphi \approx \phi \pm \frac{\pi}{2}$ is reliable in these areas. The AoLP loss can force the network to learn this prior knowledge and thus improving the prediction accuracy.

5. Conclusions

In this paper, we present a novel method for transparent shape from polarization. We demonstrate the performance of our method on our dataset, the first dataset for transparent shape from polarization. By introducing the physics-based prior, multi-branch network architecture, and self-supervised AoLP loss, the negative contribution of the interfered physics-based prior is effectively reduced and our method outperforms previous approaches on all objects in the test set. We hope our dataset and model can contribute to the community of transparent shape estimation.

Limitations Our method alleviates the transmission interference, however, when the most areas of transparnt surface are dominated by transmission component, our method will degenerate to a rgb-based method. In addition, we notice that the pure polarization information limits the overall accuracy of transparent shape estimation. Combining po-

larization information with other reconstruction methods is one of our future research directions.

Acknowledgement

This work was supported by the National Key R&D Program of Chin 2022YFB4701400/4701402), National Natural Science Foundation of Chin No. U21 B6002, 62203260, 92248304), Guangdong Basic and Applied Bas esearch Foundation (2023A1515011773), Science, Technology and Innova on Commission of Shenzhen Municipality (WDZC20200820160650001CYJ20200109143003935).

References

[1] Gary A Atkinson and Edwin R Hancock. Recovery of surface orientation from diffuse polarization. *IEEE transactions on image processing*, 15(6):1653–1664, 2006. 4

[2] Yunhao Ba, Alex Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. Deep shape from polarization. In *European Conference on Computer Vision*, pages 554–571. Springer, 2020. 2, 5, 6, 7, 8

[3] Guanying Chen, Kai Han, and Kwan-Yee K Wong. Ps-fcn: A flexible learning framework for photometric stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–18, 2018. 8

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 5, 8

[5] Valentin Deschaintre, Yiming Lin, and Abhijeet Ghosh. Deep polarization imaging for 3d shape and svbrdf acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15567–15576, 2021. 2

[6] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8602–8611, 2020. 3

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[8] Yuhi Kondo, Taishi Ono, Legong Sun, Yasutaka Hirasawa, and Jun Murayama. Accurate polarimetric brdf for real polarization scene rendering. In *European Conference on Computer Vision*, pages 220–236. Springer, 2020. 6, 7, 8

[9] Chenyang Lei, Chenyang Qi, Jiaxin Xie, Na Fan, Vladlen Koltun, and Qifeng Chen. Shape from polarization for complex scenes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12632–12641, 2022. 2, 5, 6, 7, 8

[10] Marc Levoy, J Gerth, B Curless, and K Pull. The stanford 3d scanning repository. *URL http://www-graphics. stanford. edu/data/3dscanrep*, 5(10), 2005. 6

[11] Min Li, Zhenglong Zhou, Zhe Wu, Boxin Shi, Changyu Diao, and Ping Tan. Multi-view photometric stereo: A robust solution and benchmark dataset for spatially varying isotropic materials. *IEEE Transactions on Image Processing*, 29:4159–4173, 2020. 6

[12] Zhengqin Li, Yu-Ying Yeh, and Manmohan Chandraker. Through the looking glass: Neural 3d reconstruction of transparent shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1262–1271, 2020. 1, 2

[13] Ali H Mahmoud, Moumen T El-Melegy, and Aly A Farag. Direct method for shape recovery from polarization and shading. In *2012 19th IEEE International Conference on Image Processing*, pages 1769–1772. IEEE, 2012. 2, 6, 7

[14] Daisuke Miyazaki and Katsushi Ikeuchi. Inverse polarization raytracing: estimating surface shapes of transparent objects. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 910–917. IEEE, 2005. 2

[15] Daisuke Miyazaki, Masataka Kagesawa, and Katsushi Ikeuchi. Determining shapes of transparent objects from two polarization images. In *MVA*, pages 26–31, 2002. 2

[16] Daisuke Miyazaki, Masataka Kagesawa, and Katsushi Ikeuchi. Transparent surface modeling from a pair of polarization images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):73–82, 2004. 2

[17] Daisuke Miyazaki, Robby T Tan, Kenji Hara, and Katsushi Ikeuchi. Polarization-based inverse rendering from a single view. In *Computer Vision, IEEE International Conference on*, volume 3, pages 982–982. IEEE Computer Society, 2003. 2, 6, 7

[18] Olivier Morel, Mathias Ferraton, Christophe Stolz, and Patrick Gorria. Active lighting applied to shape from polarization. In *2006 International Conference on Image Processing*, pages 2181–2184. IEEE, 2006. 2

[19] Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. Mitsuba 2: A retargetable forward and inverse renderer. *ACM Transactions on Graphics (TOG)*, 38(6):1–17, 2019. 6

[20] Jun Ohta. *Smart CMOS image sensors and applications*. CRC press, 2017. 1

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5, 8

[22] Megumi Saito, Yoichi Sato, Katsushi Ikeuchi, and Hiroshi Kashiwagi. Measurement of surface orientations of transparent objects by use of polarization in highlight. *JOSA A*, 16(9):2286–2293, 1999. 2

[23] Shreeyak Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3634–3642. IEEE, 2020. 1, 2, 7

[24] William AP Smith, Ravi Ramamoorthi, and Silvia Tozza. Height-from-polarisation with unknown lighting or albedo.

IEEE transactions on pattern analysis and machine intelligence, 41(12):2875–2888, 2018. 7

[25] Kenichiro Tanaka, Yasuhiro Mukaigawa, Hiroyuki Kubo, Yasuyuki Matsushita, and Yasushi Yagi. Recovering transparent shape from time-of-flight distortion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4387–4395, 2016. 2

[26] M Vedel, N Lechocinski, and S Breugnot. 3d shape reconstruction of optical element using polarization. In *Polarization: Measurement, Analysis, and Remote Sensing IX*, volume 7672, pages 21–33. SPIE, 2010. 1

[27] Bojian Wu, Yang Zhou, Yiming Qian, Minglun Gong, and Hui Huang. Full 3d reconstruction of transparent objects. *arXiv preprint arXiv:1805.03482*, 2018. 2

[28] Sai-Kit Yeung, Tai-Pang Wu, Chi-Keung Tang, Tony F Chan, and Stanley Osher. Adequate reconstruction of transparent objects on a shoestring budget. In *CVPR 2011*, pages 2513–2520. IEEE, 2011. 2

[29] Hu Zhang, Keke Zu, Jian Lu, Yuru Zou, and Deyu Meng. Epsanet: An efficient pyramid squeeze attention block on convolutional neural network. *arXiv preprint arXiv:2105.14447*, 2021. 5

[30] Luyang Zhu, Arsalan Mousavian, Yu Xiang, Hammad Mazhar, Jozef van Eenbergen, Shoubhik Debnath, and Dieter Fox. Rgb-d local implicit function for depth completion of transparent objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4649–4658, 2021. 2

[31] Shihao Zou, Xinxin Zuo, Sen Wang, Yiming Qian, Chuan Guo, and Li Cheng. Human pose and shape estimation from single polarization images. *IEEE Transactions on Multimedia*, 2022. 2