Shape from Polarization for Complex Scenes in the Wild
Supplementary Material

Anonymous CVPR submission

Paper ID 1557

Summary of the Supplementary Material

This supplementary document is organized as follows:

- Section 1 provides addition results.
- Section 2 presents the details of our implementation.
- Section 3 introduces the background of shape from polarization.

1. Additional Results

1.1. Comparison to non-polarization baselines

We choose the latest RGB-based normal estimation method [8] for comparison. According to the evaluation results of Yang et al. [8], their model obtains the best score on the NYU dataset when models under the same training setting. We retrain their model on the SPW dataset without using the polarization information.

Table 1 shows the quantitative results. Our results are significantly better than the results of Yang et al. [8]. Note that the results are for reference as the experimental setting is unfair: we use the extra polarization information; they use the pretrained weights on ImageNet [3].

Method	Angular Error ↓			Accuracy ↑		
	Mean	Median	RMSE	11.25°	22.5°	30.0°
TransDepth†	25.96	21.71	31.77	26.9	56.7	68.3
Ours	**17.86**	**14.20**	**22.72**	**44.6**	**76.3**	**85.2**

Table 1. **Quantitative evaluation on the SPW dataset.** Our approach outperforms rgb-based method TransDepth [8] by a large margin on all evaluation metrics. †: we retrain the model on the unpolarized intensity images in SPW dataset.

1.2. Ablation experiments

We report the quantitative results with various number of self-attention blocks in Table 2. On the SPW dataset, using 8 blocks obtains the best performance. An interesting phenomenon is that using only 1 block can also improve the performance substantially.

Blocks	Angular Error ↓			Accuracy ↑		
	Mean	Median	RMSE	11.25°	22.5°	30.0°
0	21.08	16.54	26.62	36.1	68.5	79.3
1	18.99	14.90	24.22	42.1	74.9	83.6
2	19.21	15.35	24.17	41.4	73.6	82.7
4	18.40	14.39	23.49	**45.0**	75.9	84.5
8	**17.86**	**14.20**	**22.72**	44.6	**76.3**	**85.2**
12	18.81	14.65	24.21	43.7	75.2	83.9

Table 2. **Ablation experiments for the number of self-attention blocks on the SPW dataset.** We choose 8 blocks in our model according to the quantitative results.

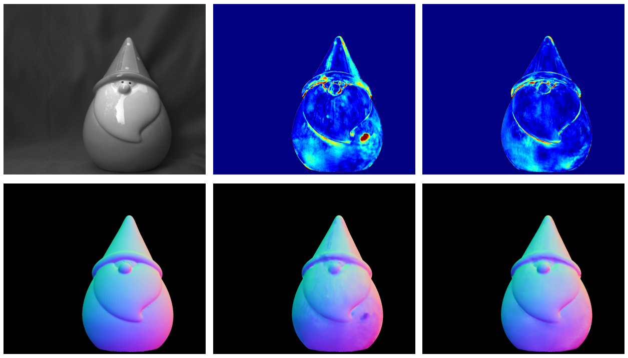

Input $\mathbf{I}_{un}$ and GT DeepSfP [2] Ours

Figure 1. Visual comparison results of estimated normals. We show error maps of DeepSfP [2] and ours.

1.3. Visualization on Deepsfp Dataset

In Fig. 1, we present more perceptual results of our method and DeepSfP [2] baseline on the DeepSfP [2] dataset. Our result is more accurate than the baseline since our hybrid architecture and polarization representation can handle the diffuse/specular-ambiguity better.

2. Implementation Details

2.1. Network architecture

Table 3 shows details of our architecture specification. “3×3, 64,” denotes a 2d convolution operation of kernel size 3, output channel 64. “BN, ReLU” denotes batch normalization [4] and ReLU activation [5]. “IN” denotes instance normalization [7], while “LN” denotes layer normalization [1]. In multi-head self-attention blocks, “dim 512 (head 8) MHA” indicates a 8-heads attention block each with head dimention 64. “2048-d MLP” denotes a MLP with a hidden layer of 2048 dimensions.

3. Shape from Polarization

In this section, we provide a detailed introduction to polarization. Table 4 shows all the used symbols and notations.

3.1. Polarization measurement

Given polarization images $I^0, I^{\pi/2}, I^{\pi/4}, I^{3\pi/4}$ obtained by different polarizer angles, the polarization information can be obtained through the following equation:

$$I = (I^0 + I^{\pi/2} + I^{\pi/4} + I^{3\pi/4})/2, \tag{1}$$

$$\rho = \frac{\sqrt{((I^0 - I^{\pi/2})^2 + (I^{\pi/4} - I^{3\pi/4})^2)}}{I}, \tag{2}$$

$$\phi = \frac{1}{2}\arctan\frac{I^{\pi/4} - I^{3\pi/4}}{I^0 - I^{\pi/2}}. \tag{3}$$

3.2. Preliminary

Coordinate system. We represent the surface normal and viewing direction in a global coordinate system, as shown in Fig. 2. The x-axis is rightward. The y-axis is upward. The z-axis is pointing out of paper. The original point of this coordinate system coincides with the camera’s Principle Point. The camera plane is perpendicular to the z-axis.

Normal representation. Surface normal can be represented by two angles θ and α:

$$\mathbf{n} = [\sin\theta\cos\alpha, \sin\theta\sin\alpha, \cos\theta]^\intercal, \tag{4}$$

where $\mathbf{n}$ is the surface normal, θ is the zenith angle, and α is the azimuth angle, as shown in Fig. 2.

3.3. Polarization under orthographic projection

More details about the relationship between the surface normal and polarization information are presented.

3.3.1 Zenith angle

The viewing angle $\theta_\mathbf{v}$ is the angle between viewing direction and surface normal. Under orthographic projection, the zenith angle θ equals to the viewing angle $\theta_\mathbf{v}$ according to Equation 4:

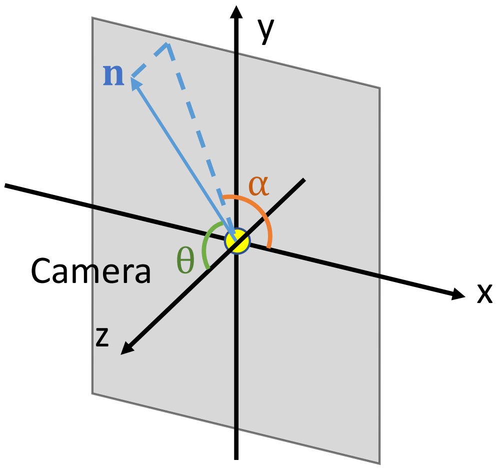


Figure 2. Our coordinate system.

$$\begin{aligned} \cos\theta_\mathbf{v} &= \mathbf{n}\cdot\mathbf{v} = n_x v_x + n_y v_y + n_z v_z, & (5) \\ &= \cos\theta,\ \text{ if }\ \mathbf{v} = [0,0,1] & (6) \end{aligned}$$

The viewing angle $\theta_\mathbf{v}$ influences the degree of polarization ρ directly. Specifically, given the refractive index η of the object, the degree of polarization ρ is decided by the viewing angle $\theta_\mathbf{v}$ with a function $\rho = g(\theta_\mathbf{v}; \eta)$. The function g is decided by many factors, such as the reflection type. For example, for specular reflection, we have:

$$\rho = \frac{2\sin^2\theta_\mathbf{v}\cos\theta_\mathbf{v}\sqrt{\eta^2 - \sin^2\theta_\mathbf{v}}}{\eta^2 - \sin^2\theta_\mathbf{v} - \eta^2\sin^2\theta_\mathbf{v} + 2\sin^4\theta_\mathbf{v}}. \tag{7}$$

For diffuse reflection, we have:

$$\rho = \frac{(\eta - \frac{1}{\eta})^2\sin^2\theta_\mathbf{v}}{2 + 2\eta^2 - (\eta + \frac{1}{\eta})^2\sin^2\theta_\mathbf{v} + 4\cos\theta_\mathbf{v}\sqrt{\eta^2 - \sin^2\theta_\mathbf{v}}}. \tag{8}$$

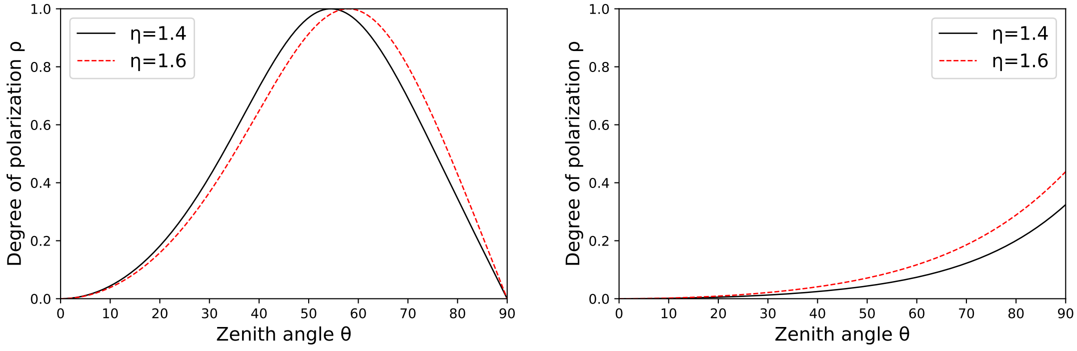
Equation 7 can be inverted to obtain an estimation of viewing angle from the degree of polarization ρ:

$$\cos\theta = \cos\theta_\mathbf{v} = \sqrt{\frac{\eta^4(1-\rho^2)+2\eta^2(2\rho^2+\rho-1)+\rho^2+2\rho-4\eta^3\rho\sqrt{1-\rho^2}+1}{(\rho+1)^2(\eta^4+1)+2\eta^2(3\rho^2+2\rho-1)}}. \tag{9}$$

As shown in Fig. 3, we can estimate the zenith angle θ given the degree of polarization ρ under a specific refractive index η and the reflection type.

stage	building block	output size
input convolution	$\begin{bmatrix} 3\times 3, 64 \\ \text{BN, ReLU} \end{bmatrix} \times 2$	$H \times W \times 64$
downsampling convolution1	$\begin{bmatrix} 3\times 3, 128 \\ \text{IN, ReLU} \end{bmatrix} \times 2$	$\frac{H}{2} \times \frac{W}{2} \times 128$
downsampling convolution2	$\begin{bmatrix} 3\times 3, 256 \\ \text{IN, ReLU} \end{bmatrix} \times 2$	$\frac{H}{4} \times \frac{W}{4} \times 256$
downsampling convolution3	$\begin{bmatrix} 3\times 3, 512 \\ \text{IN, ReLU} \end{bmatrix} \times 2$	$\frac{H}{8} \times \frac{W}{8} \times 512$
downsampling convolution4	$\begin{bmatrix} 3\times 3, 512 \\ \text{IN, ReLU} \end{bmatrix} \times 2$	$\frac{H}{16} \times \frac{W}{16} \times 512$
multi-head attention	$\begin{bmatrix} \text{LN, dim 512 (head 8) MHA} \\ \text{LN, 2048-d MLP} \end{bmatrix} \times 8$	$\frac{H}{16} \times \frac{W}{16} \times 512$
skip-connection and upsampling convolution1	$\begin{bmatrix} 3\times 3, 512 \\ \text{BN, ReLU} \\ 3\times 3, 256 \\ \text{BN, ReLU} \end{bmatrix} \times 1$	$\frac{H}{8} \times \frac{W}{8} \times 256$
skip-connection and upsampling convolution2	$\begin{bmatrix} 3\times 3, 256 \\ \text{BN, ReLU} \\ 3\times 3, 128 \\ \text{BN, ReLU} \end{bmatrix} \times 1$	$\frac{H}{4} \times \frac{W}{4} \times 128$
skip-connection and upsampling convolution3	$\begin{bmatrix} 3\times 3, 128 \\ \text{BN, ReLU} \\ 3\times 3, 64 \\ \text{BN, ReLU} \end{bmatrix} \times 1$	$\frac{H}{2} \times \frac{W}{2} \times 64$
skip-connection and upsampling convolution4	$\begin{bmatrix} 3\times 3, 64 \\ \text{BN, ReLU} \end{bmatrix} \times 2$	$H \times W \times 64$
output convolution	$1 \times 1, 3$	$H \times W \times 3$

Table 3. **Architectures of our hybrid model.** Building blocks are shown in brackets, with the numbers of blocks stacked. Downsampling is performed at the beginning of downsampling convolution layer using max pooling of stride 2. 2× bilinear upsampling and skip-connection with the encoder features are conducted at the beginning of upsampling convolution layer.



(a) Specular reflection

(b) Diffuse reflection

Figure 3. Degree of polarization changes differently for (a) specular and (b) diffuse reflection. η: refractive index.

Symbol	Description
$\mathbf{n}$	Surface normal
$\mathbf{P_i}$	Incidence plane.
$\mathbf{n_i}$	Normal of the incidence plane
$\mathbf{n_c}$	Normal of the camera plane
$\mathbf{v}$	Viewing direction
α	Azimuth angle
θ	Zenith angle
$\theta_\mathbf{v}$	Viewing angle, the angle between $\mathbf{n}$ and $\mathbf{v}$
η	Refractive index
ρ	Degree of polarization
ϕ	Angle of polarization
I_{un}	Unpolarized intensity
I	Intensity of incident light
$\mathbf{d}$	3D polarization direction
ϕ_{pol}	Polarizer angle
$\mathbf{\Phi}$	The vector representation of ϕ in the camera plane.

Table 4. Symbols and notations used in the paper.

3.3.2 Azimuth angle

The azimuth angle α is closely related to the polarization angle ϕ. Specifically, there are four possible solutions for α based on the measured ϕ under the *orthographic assumption* (i.e., $\mathbf{v} = [0, 0, 1]^\intercal$):

$$\alpha \in \{\phi, \phi + \pi, \phi + \pi/2, \phi - \pi/2\}, \;\; 0 \leq \alpha < 2\pi. \tag{10}$$

There are two ambiguities in the solution: π-ambiguity and $\pi/2$-ambiguity. The π-ambiguity is because ϕ is from 0 to π and there is no difference between ϕ and $\phi + \pi$. The $\pi/2$-ambiguity is decided by the reflection type. If diffuse reflection dominates, α equals to ϕ or $\phi + \pi$; if specular reflection dominates, there is a $\pi/2$ shift compared with ϕ.

3.3.3 Solutions for surface normal

As analyzed in Section 3.3.1 and Section 3.3.2, zenith angle θ and azimuth angle α can be estimated through the degree of polarization ρ and angle of polarization ϕ respectively. At last, we can obtain possible solutions using Equation 4 directly.

3.4. Polarization under perspective projection

Most equations in Section 3.3 do not hold under perspective projection since we cannot assume $\mathbf{v} = [0, 0, 1]^\intercal$ for all pixels. However, we can still derive other equations from utilizing the relationship between polarization and surface normal $\mathbf{n}$.

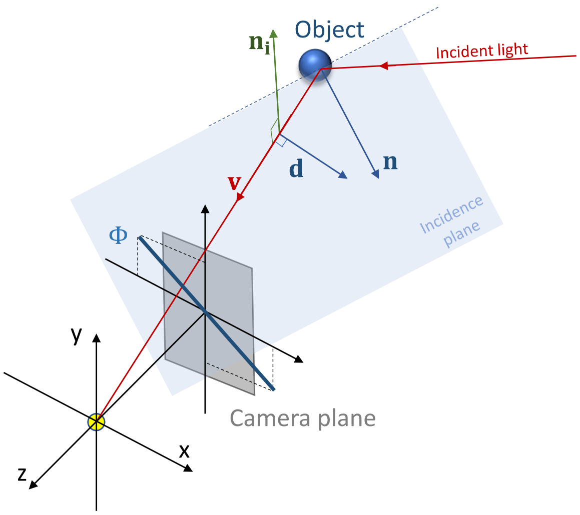


Figure 4. Shape from polarization under perspective projection.

3.4.1 Degree of polarization

Zhu et al. [9] extend the linear formulation of Smith et al. [6] to the perspective case for the zenith angle. Note the degree of polarization ρ is decided by the viewing angle $\theta_\mathbf{v}$ and refractive index η. Since $\theta_\mathbf{v}$ is angle between surface normal $\mathbf{n}$ and viewing direction $\mathbf{v}$:

$$\cos\theta_\mathbf{v} = \mathbf{n} \cdot \mathbf{v} = n_x v_x + n_y v_y + n_z v_z, \tag{11}$$

$$\cos\theta_\mathbf{v} = v_x \sin\theta\cos\alpha + v_y \sin\theta\sin\alpha + v_z \cos\theta. \tag{12}$$

Obviously, we do not have $\theta = \theta_\mathbf{v}$ under perspective projection ($\mathbf{v}$ is not $[0, 0, 1]^\intercal$) for all pixels. Hence, we cannot estimate the zenith angle θ through the degree of polarization ρ individually like Equation 9.

3.4.2 Angle of polarization

We extend the azimuth angle formulation to the perspective case here. The incidence plane is the plane that contains the surface normal, incident light, and viewing direction. Hence, the normal of incidence plane $\mathbf{n_i}$ is perpendicular to surface normal $\mathbf{n}$ and viewing direction $\mathbf{v}$:

$$\mathbf{n_i} = \mathbf{n} \times \mathbf{v}. \tag{13}$$

In terms of their physical properties, the polarization direction $\mathbf{d}$ has no difference with its opposite direction $-\mathbf{d}$. For brevity, we only consider one direction. The polarization direction $\mathbf{d}$ is perpendicular to the propagation direction of light. In addition, $\mathbf{d}$ is always parallel or perpendicular to the incidence plane. Hence, we have:

$$\mathbf{d} = \begin{cases} \mathbf{n_i}, & \mathbf{d} \perp \mathbf{P_i} \\ \mathbf{n_i} \times \mathbf{v}, & \mathbf{d} \parallel \mathbf{P_i}, \end{cases} \tag{14}$$

where $\mathbf{P_i}$ is the incidence plane.

When polarization direction $\mathbf{d}$ is projected on the camera plane, we have the following equation since this is an intersection between polarization direction $\mathbf{d}$ and camera plane:

$$\Phi = (\mathbf{d} \times \mathbf{v}) \times \mathbf{n_c} = [\cos\phi, \sin\phi, 0]^\intercal, \tag{15}$$

$$\phi = \arctan(\Phi_y/\Phi_x), \tag{16}$$

where $\mathbf{n_c} = [0, 0, 1]^\intercal$ is the surface normal of camera plane. At last, we can get the angle of polarization ϕ through Equation 16 directly.

To sum up, the Φ can be modeled as follows. When diffuse reflection dominates, we have:

$$\Phi = \mathbf{d} \times \mathbf{n_c} = \mathbf{n_i} \times \mathbf{n_c} = \mathbf{n} \times \mathbf{v} \times \mathbf{n_c}. \tag{17}$$

When specular reflection dominates, we have:

$$\Phi = \mathbf{d} \times \mathbf{n_c} = \mathbf{n_i} \times \mathbf{v} \times \mathbf{n_c} = \mathbf{n} \times \mathbf{v} \times \mathbf{v} \times \mathbf{n_c}. \tag{18}$$

We provide an example for Equation 17 and Equation 18. From Equation 17, when $\mathbf{v} = [0, 0, 1]^\intercal$, we have:

$$\begin{aligned}
\Phi_{[0,0,1]} &= \mathbf{n} \times \mathbf{v} \times \mathbf{n_c} \\
&= [n_y v_z - n_z v_y, n_z v_x - n_x v_z, n_x v_y - n_y v_x]^\intercal \times \mathbf{n_c} \\
&= [n_z v_x - n_x v_z, -(n_y v_z - n_z v_y), 0]^\intercal \\
&= [-n_x, -n_y, 0]^\intercal, && (19) \\
\phi &\in \{\alpha, \alpha + \pi\} && (20)
\end{aligned}$$

Similarly, from Equation 18,

$$\begin{aligned}
\Phi_{[0,0,1]} &= \mathbf{n} \times \mathbf{v} \times \mathbf{v} \times \mathbf{n_c} \\
&= [-n_x, -n_y, 0]^\intercal \times \mathbf{n_c} \\
&= [-n_y, n_x, 0]^\intercal, && (21) \\
\phi &\in \{\alpha + \pi/2, \alpha - \pi/2\} && (22)
\end{aligned}$$

Our result under $\mathbf{v} = [0, 0, 1]^\intercal$ is consistent with Equation 10.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 2

[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 *ECCV*, 2020. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[4] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 2

[5] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 2

[6] William A. P. Smith, Ravi Ramamoorthi, and Silvia Tozza. Height-from-polarisation with unknown lighting or albedo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(12):2875–2888, 2019. 4

[7] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv:1607.08022*, 2016. 2

[8] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *ICCV*, 2021. 1

[9] Dizhong Zhu and William A. P. Smith. Depth from a polarisation + RGB stereo pair. In *CVPR*, 2019. 4