Learning Accurate 3D Shape Based on Stereo Polarimetric Imaging (Supplementary Material)

Tianyu Huang[1*] Haoang Li[1,2*] Kejing He[1] Congying Sui[1] Bin Li[1] Yun-Hui Liu[1†]
[1]The Chinese University of Hong Kong, Hong Kong, China
[2]Technical University of Munich, Germany

Overview of Supplementary Material

This supplementary material is organized as follows:

- In Section 1, we introduce additional information regarding shape from polarization (SfP).
- In Section 2, we present details of acquiring ground truth normal and disparity maps.
- In Section 3, we introduce our network architecture and implementation details.
- In Section 4, we provide additional experimental results.

1. Additional Information Regarding SfP

We first introduce how to compute the polarization parameters, followed by illustrating the orthographic projection problem.

1.1. Polarization Parameters

Recall that in Section 3.1 of the main manuscript (line 259), we can use the polarization camera to measure the light intensities in different angles of polarizer ϕ_{pol}. We denote these intensities with respect to $0°$, $45°$, $90°$, and $135°$ by I_0, I_{45}, I_{90}, and I_{135}, respectively. By combing the measured intensities, we can solve the unknown polarization parameters through the following equations:

- Angle of Linear Polarization (AoLP) with π-ambiguity:

$$\phi = \frac{1}{2}\arctan\left(\frac{I_0 + I_{90} - 2I_{45}}{I_{90} - I_0}\right). \tag{1}$$

- Degree of Linear Polarization (DoLP):

$$\rho = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} = \frac{\sqrt{(I_0 - I_{90})^2 + (I_{45} - I_{135})^2}}{I_0 + I_{90}}. \tag{2}$$

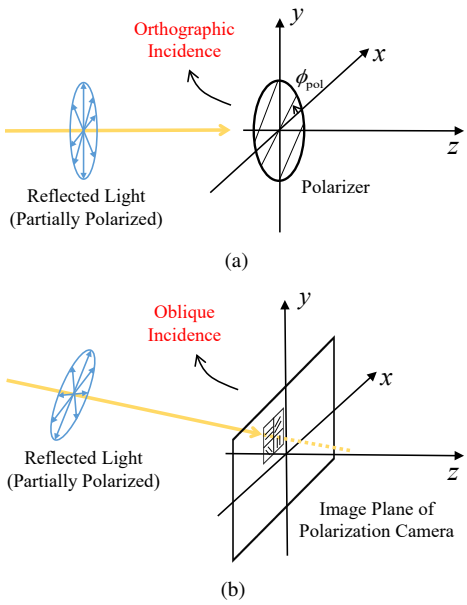


Figure 1. (a) Polarimetric measurement based on orthographic projection. (b) Non-orthographic projection in practice.

- Average intensity:

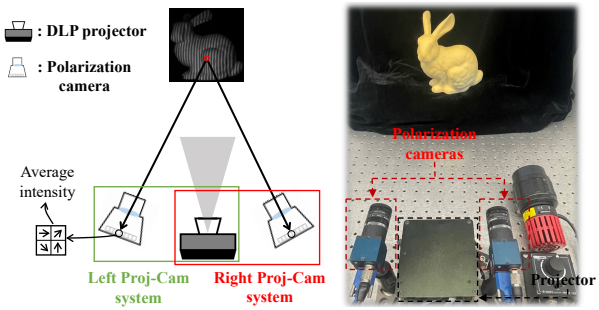
$$\bar{I} = \frac{I_0 + I_{45} + I_{90} + I_{135}}{2}. \tag{3}$$

The above parameters have been proved to be highly related to the surface normal [1, 3, 5, 18].

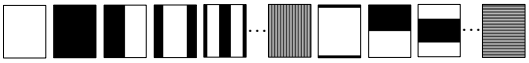
1.2. Orthographic Projection Problem

Recall that in Section 3.1 of the main manuscript (lines 242-249), the polarimetric measurement is based on orthographic incidence to the polarizer of the measured light [4] (see Fig. 1(a)). However, in practice, polarimetric measurement using the quad-Bayer polarization camera is based on non-orthographic projection. Specifically, the incident lights of the pixels away from the image center are oblique, as shown in Fig. 1(b). For problem simplification, most existing SfP methods [2, 17] assume orthographic projection

*Tianyu Huang and Haoang Li contributed equally to this work.
†Yun-Hui Liu is the corresponding author.



(a)



(b)

Figure 2. Our stereo structured-light system to acquire ground truth normal and disparity maps. (a) The projector and the stereo polarization camera rig are combined as a stereo structured-light system. (a) Structured-light patterns (i.e., Gray code).

for the whole image (i.e., viewing directions of all pixels are $[0, 0, 1]^\top$). This assumption affects the accuracy of polarimetric measurement and further the quality of shape recovery. We propose to integrate the viewing direction map with Transformer to solve this problem.

2. Acquisition of Ground Truth Maps

Recall that in Section 4 of the main manuscript (lines 517-524), we adopt a stereo structured-light system to acquire the ground truth normal and disparity maps of our dataset. This system is composed of a projector and a stereo polarization camera rig, as shown in Fig. 2(a).

The projector projects a sequence of structured light patterns (i.e., Gray code in Fig. 2(b)) on the object surface. Accordingly, each mini-patch of the object surface is associated with a unique code. We use the stereo camera rig to obtain images of the object surface. These images are used to acquire the stereo code maps [13]. We use these code maps to compute the ground truth normal and disparity maps as follows. For one thing, the code map of the left view and the projected patterns constitute a stereo pair. We use this pair to calculate the point cloud observed by the left camera. Given the point cloud, we use the least-squares normal estimation algorithm [15] to calculate the ground truth normal map of the left view. For another, the stereo code maps also constitute a stereo pair. We use this pair to calculate the ground truth disparity map.

Table 1. Details of our network architecture.

Network Part	Layer Setting	Output
Inputs		5×1024×1024
Feature extraction	Conv2d + BN2d + LeakyReLU, [7, 16, 2] Conv2d + BN2d + LeakyReLU, [5, 32, 2] Conv2d + BN2d + LeakyReLU, [5, 32, 1] Conv2d + BN2d + LeakyReLU, [3, 64, 2] Conv2d + BN2d + LeakyReLU, [3, 64, 1]	64×128×128
	CSWin-Transformer block with ***VDPE***	64×128×128
Cost volume	Concatenate	$128\times\frac{D+1}{8}\times128\times128$
Disparity part	Conv3d + BN3d + LeakyReLU, [3, 64, 1] Conv3d + BN3d + LeakyReLU, [3, 32, 1]	$32\times\frac{D+1}{8}\times128\times128$
	ConvTrans3d + BN3d + LeakyReLU, [3, 16, 2] Conv3d + BN3d + LeakyReLU, [3, 1, 1] Upample(2, 2, 2)	$1\times\frac{D+1}{2}\times512\times512$
	squeeze() + softmax() + sum()	1×512×512
Normal part	(CSWin-Transformer block with ***VDPE***) ×2	64×128×128
	(CSWin-Transformer block with ***VDPE***) ×2	64×128×128
	Conv2d + BN2d + LeakyReLU, [3, 32, 1] ConvTrans2d + BN2d + LeakyReLU, [3, 16, 2] ConvTrans2d + BN2d + LeakyReLU, [3, 8, 2] Conv2d + BN2d + LeakyReLU, [3, 3, 1]	3×512×512
	Normalize	3×512×512
SCMP (Disparity)	Conv3d + BN3d + LeakyReLU, [3, 1, 1]	$1\times\frac{D+1}{8}\times128\times128$
	squeeze() + softmax() + sum()	1×128×128
SCMP (Normal)	Conv2d + BN2d + LeakyReLU, [3, 32, 1] Conv2d + BN2d + LeakyReLU, [3, 16, 1] Conv2d + BN2d + LeakyReLU, [3, 3, 1]	3×128×128
	Normalize	3×128×128

3. Network Architecture and Implementation Details

3.1. Network Architecture

Table 1 shows the details of our network architecture. 2D features are expressed by channel × height × width and 3D features are expressed by channel × disparity × height × width. D represents the maximum disparity. "ConvTrans2d" and "ConvTrans3d" represent the transposed 2D and 3D convolution neural network layer, respectively. "BN2d" and "BN3d" represent 2D and 3D batch normalization operations, respectively. "[7, 16, 2]" denotes the kernel size of 7, the output channel of 16, and the stride of 2. "Upsample(2, 2, 2)" means 2×2×2 trilinear upsampling on the disparity, height, and width dimensions of 3D features. To reduce the computational cost, we adopt the Cross-Shaped Window-based self-attention mechanism (CSWin-Transformer) [6] instead of the classical full-attention Transformer [7]. The outputs labeled by yellow are fed to our first SCMP module, while those labeled by blue are fed to our second SCMP module.

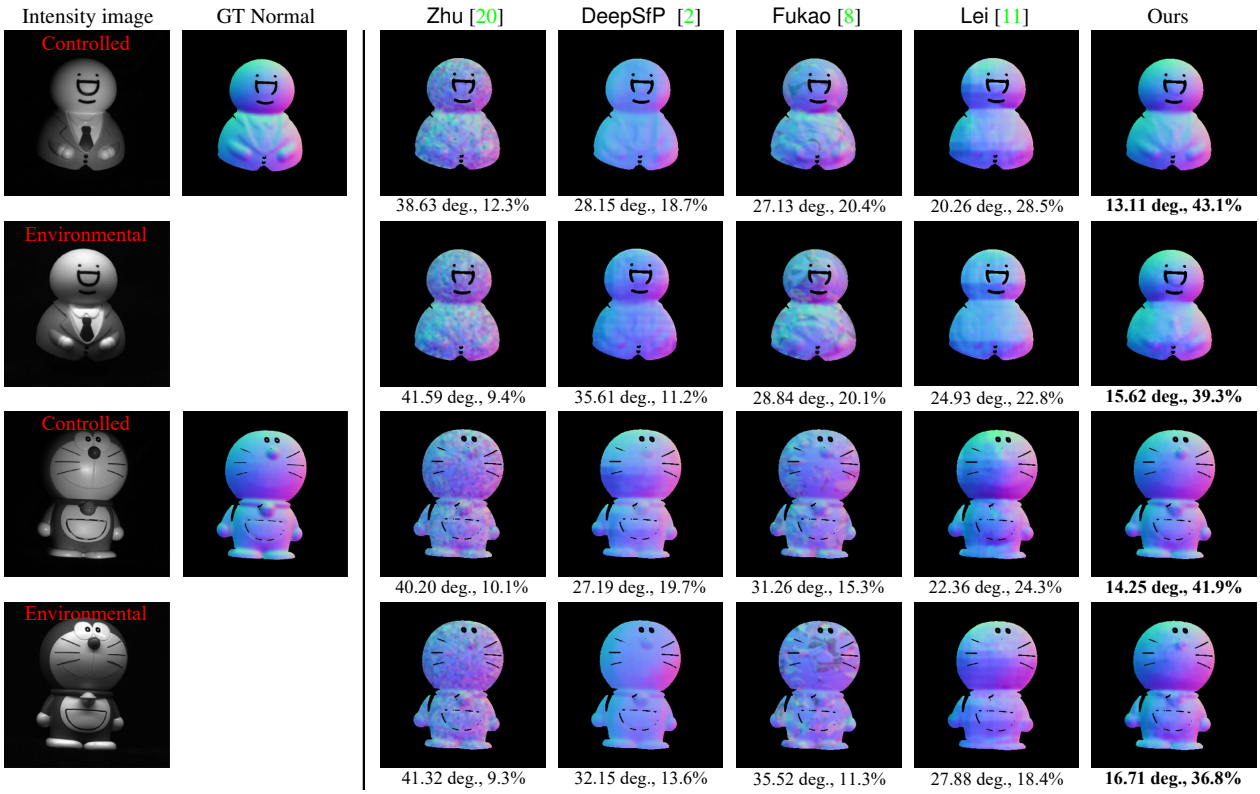


Figure 3. Representative accuracy comparisons between various SfP methods for normal recovery under the controlled and environmental illuminations. (a) Intensity images and ground truth (GT) normal maps. (b) The first and third rows show the estimated normal maps under *controlled* illumination. The second and fourth rows show the estimated normal maps under *environmental* illumination. A pair of numbers below each estimated normal map represents the mean of angular errors and the percentage of pixels with angular errors smaller than 11.25°.

3.2. Implementation Details

We implement our network in PyTorch [14]. To train our network, we adopt the Adam optimizer [9] and a cosine decay scheduler for the learning rate. We train our network for 400 epochs with the batch size of 24. All the experiments are conducted on a computer equipped with an Intel(R) i9-10900K CPU and four NVIDIA GeForce RTX 3090 GPUs.

4. Experiments

In Section 4.1, we present an extra test regarding environmental illumination. In Section 4.2, we provide additional results of the experiments introduced in the main manuscript.

4.1. Test Regarding Environmental Illumination

In the main manuscript, we only report the experimental results under the controlled illumination. In this section, we present an extra test regarding environmental illumination. The environmental illumination typically contains various polarized lights, which introduces noise to the polarimetric measurement. To obtain testing data, we illuminate 13 objects under the controlled and environmental illuminations, respectively. Under different illuminations, we obtain images from the same viewpoint. For each object, we use 5 random viewpoints. We use these images to test our method and the four state-of-the-art SfP approaches, i.e., Zhu [20], DeepSfP [2], Fukao [8], and Lei [11].

As shown in Table 2 and Fig. 3, for all the SfP methods, their accuracies under controlled illumination are higher than those under environmental illumination. The reason is that the controlled illumination can significantly reduce the noise in the polarimetric measurements. In addition, our method still achieves the highest accuracy on all the evaluation metrics.

4.2. Additional Comparison Results

Robustness to Light Variation. In the main manuscript, we only report the comparison results between Fukao, Lei, and our method. As shown in Fig. 4, we additionally pro-

Table 2. Accuracy comparisons between various SfP methods for normal recovery on all the images obtained under the *controlled* and *environmental* illuminations.

Method	Controlled Illumination						Environmental Illumination					
	Angular Error (deg.)			Pixel Percentage (%)			Angular Error (deg.)			Pixel Percentage (%)		
	Mean	Median	RMSE	11.25°	22.5°	30.0°	Mean	Median	RMSE	11.25°	22.5°	30.0°
Zhu [20]	38.17	33.98	44.21	11.7	32.5	42.9	43.25	36.71	48.13	8.9	24.7	36.5
DeepSfP [2]	27.03	25.62	32.91	18.9	40.3	60.8	33.18	28.37	39.62	12.8	30.7	42.6
Fukao [8]	30.89	27.66	36.98	17.3	38.1	55.6	32.67	29.07	39.21	15.6	32.3	47.5
Lei [11]	22.31	20.12	29.17	29.7	53.6	70.2	25.71	22.16	32.69	21.3	43.2	67.3
Ours	**13.51**	**11.83**	**17.92**	**43.7**	**75.8**	**91.3**	15.69	13.21	19.55	37.7	62.3	81.6

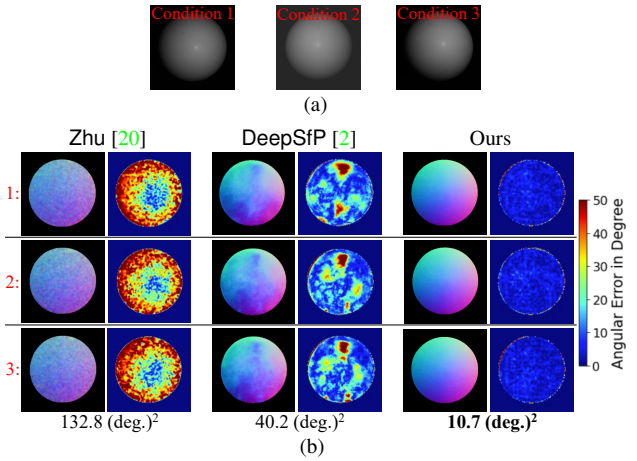


Figure 4. Representative comparisons regarding robustness to light variation between Zhu [20], DeepSfP [2], and our method. (a) Three different illumination conditions. (b) Each method corresponds to two columns. The first column shows the recovered normal maps under different illumination conditions. The second column shows the error maps. The number below each column pair represents the mean of variances of angular errors.

Table 3. Accuracy comparisons between various methods in terms of pixel percentage. †: the same network inputs as ours.

Method	Pixel Percentage (%)		
	11.25°	22.5°	30.0°
Long [12]†	26.7	51.3	68.3
Kusupati [10]†	34.2	62.7	76.9
Ours	**46.2**	**77.5**	**90.1**

vide the recovery results of Zhu and DeepSfP. Compared with these two methods, our approach still leads to higher robustness to light variation.

Joint Estimation of Disparity and Normal. In the main manuscript, we only report the results on two evaluation metrics, i.e., angular error and disparity error. As shown in Table 3, we provide the results on an extra metric, i.e., percentage of pixels with angular error smaller than 11.25°, 22.5°, and 30.0°. These results demonstrate that our method still achieves the highest accuracy.

Table 4. Ablation study regarding different network inputs. We report the results in terms of pixel percentage.

Network Inputs	Pixel Percentage (%)		
	11.25°	22.5°	30.0°
Intensity images	13.2	29.8	45.2
Raw polarization	33.6	60.7	78.3
AoLP&DoLP	35.4	66.7	83.6
Stokes maps	32.7	63.5	79.1
Inputs as DeepSfP [2]	38.6	69.5	85.3
Original (AoLP&DoLP Stokes maps)	**46.2**	**77.5**	**90.1**

Table 5. Ablation study regarding SCMP module. We report the results in terms of pixel percentage.

Network Design	Pixel Percentage (%)		
	11.25°	22.5°	30.0°
Without any SCMP	28.9	54.7	71.1
Without first SCMP	36.5	67.1	81.3
Without second SCMP	40.7	69.3	84.2
Original (with both SCMPs)	**46.2**	**77.5**	**90.1**

Table 6. Ablation study regarding VDPE design. We report the results in terms of pixel percentage.

Encoding Strategy	Pixel Percentage (%)		
	11.25°	22.5°	30.0°
APE [19]	31.6	61.9	78.3
RPE [16]	34.7	65.2	81.7
VE [11]	41.2	72.7	85.9
Original (VDPE)	**46.2**	**77.5**	**90.1**

Ablation Study. In the main manuscript, we only report the results on two evaluation metrics, i.e., angular error and disparity error. As shown in Tables 4, 5, and 6, we additionally provide the results on an extra metric, i.e., percentage of pixels. These results demonstrate the effectiveness of the proposed network inputs, SCMG module, and VDPE strategy, respectively.

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. 1

[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. 1, 3, 4

[3] Seung-Hwan Baek, Daniel S Jeon, Xin Tong, and Min H Kim. Simultaneous acquisition of polarimetric svbrdf and normals. *ACM Trans. Graph.*, 37(6):268–1, 2018. 1

[4] Edward Collett. Field guide to polarization. Spie Bellingham, WA, 2005. 1

[5] Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, and Jan Kautz. Polarimetric multi-view stereo. In *IEEE conference on computer vision and pattern recognition*, pages 1558–1567, 2017. 1

[6] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 2

[8] Yoshiki Fukao, Ryo Kawahara, Shohei Nobuhara, and Ko Nishino. Polarimetric normal stereo. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 682–690, 2021. 3, 4

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[10] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal assisted stereo depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2189–2199, 2020. 4

[11] Chenyang Lei, Chenyang Qi, Jiaxin Xie, Na Fan, Vladlen Koltun, and Qifeng Chen. Shape from polarization for complex scenes in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12632–12641, 2022. 3, 4

[12] Xiaoxiao Long, Lingjie Liu, Christian Theobalt, and Wenping Wang. Occlusion-aware depth estimation with adaptive normal constraints. In *European Conference on Computer Vision*, pages 640–657. Springer, 2020. 4

[13] Daniel Moreno and Gabriel Taubin. Simple, accurate, and robust projector-camera calibration. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 464–471, 2012. 2

[14] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 3

[15] Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011. IEEE. 2

[16] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. 4

[17] William AP Smith, Ravi Ramamoorthi, and Silvia Tozza. Linear depth estimation from an uncalibrated, monocular polarisation image. In *European Conference on Computer Vision*, pages 109–125. Springer, 2016. 1

[18] 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. 1

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017. 4

[20] Dizhong Zhu and William AP Smith. Depth from a polarisation+ rgb stereo pair. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7586–7595, 2019. 3, 4