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

Short Paper

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

Ashish Tiwari and Shanmuganathan Raman

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

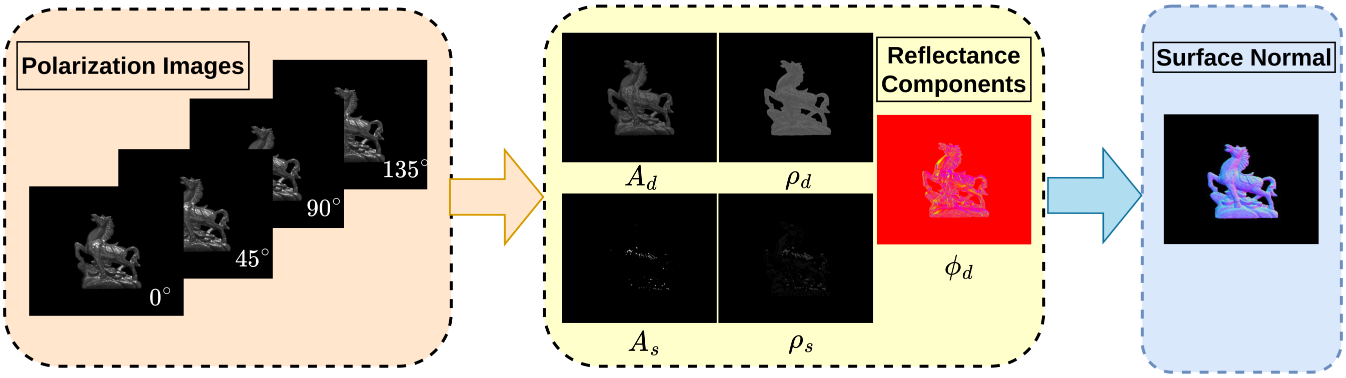

Figure 1: *Overview of the proposed self-supervised inverse rendering-based framework (SS-SfP) to obtain per-pixel surface normals under mixed polarization by decomposing the diffuse (A_d) and specular (A_s) reflection components from the raw polarization images.*

The following is the summary of the contents covered in the supplementary material.

1. Ambiguities in SfP
2. Implementation details
3. Dataset Details
4. Additional Ablation Experiments
5. Additional Qualitative Results
6. Polarization Angle Ratio Constraint - Derivation

1. Ambiguities in SfP

An unpolarized light striking a surface point exhibits diffuse and/or specular (mixed) reflection (see Figure 2 (a)). The estimation of ϕ and ρ depends on the surface reflectance model and directly relates to the azimuth (φ) and the zenith (θ) angles, respectively, described as per the coordinate system shown in Figure 2(b). The polarization image formation is generally given by Equation 1.

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

Equation 1 is manifested in the form of a Transmitted Radiance Sinusoid (TRS), as shown in Figure 2 (c).

(i) Azimuth Angle Ambiguities. As per Equation 1, two azimuth angles separated by π radians cannot be distinguished in polarization images, i.e., φ and $\varphi + \pi$ will have the same result. This is referred to as ***azimuthal angle ambiguity***. Consider T and R as the transmittance and reflectance coefficients either parallel ($||$) or perpendicular ($\perp$) to the incidence plane. Under diffuse reflection, a portion of the light enters the object and gets refracted and thus, partially polarized [WB93] with a greater magnitude in the direction parallel to the incidence plane ($T_{||} > T_{\perp}$). Therefore, maximum light intensity is observed for $\varphi = \phi$. Under specular reflection, the reflected light is predominantly polarized in the direction perpendicular to the incidence plane ($R_{\perp} > R_{||}$). Therefore, the maximum light intensity will be observed at $\varphi = \phi \pm \frac{\pi}{2}$. In short, for a general surface, when the type of reflectance is not known apriori, we are unsure if the estimated angle should be shifted $\pi/2$. This is called ***azimuthal model mismatch***.

(ii) Zenith Angle Ambiguities. The zenith angle relies on the degree of polarization (DoP) (ρ) and refractive index (η). Moreover, as in the case of azimuthal angle estimation, the type of reflectance model affects the zenith angle estimation as well and produces ***zenith model mismatch***, as described below. The DoP is de-

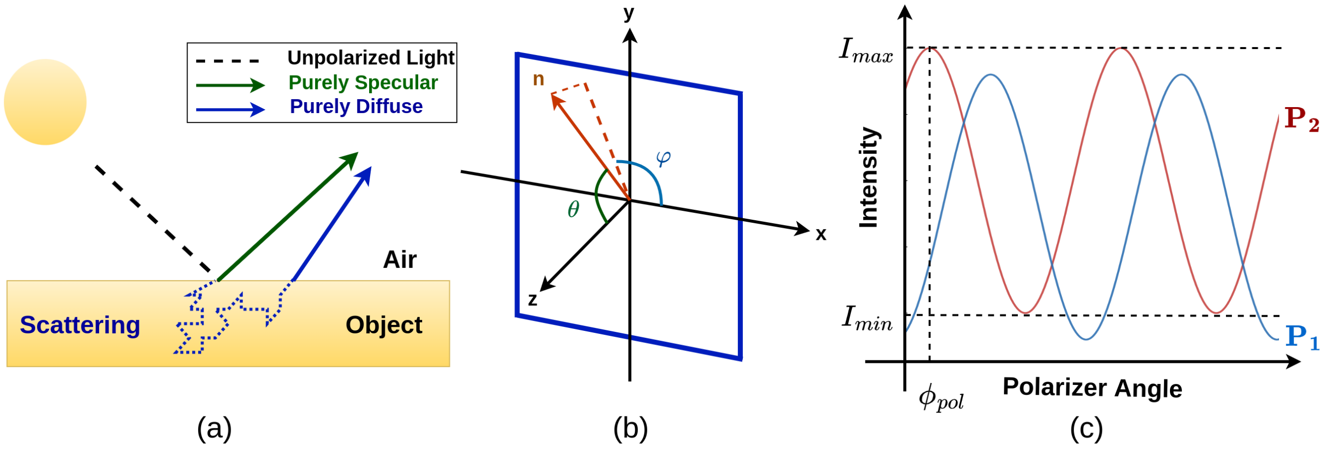

Figure 2: *(a) Mixed reflections (and polarization) off the surface. (b) Coordinate system for polarization imaging. (c) Transmitted Radiance Sinusoid (TRS) showing the observed intensities under varying polarizer angles for two pixels (P_1 and P_2) with different surface normals.*

scribed as per Equation 2 for diffuse reflection [Col05].

$$\rho = \frac{(\eta - \frac{1}{\eta})^2 \sin^2\theta}{2 + 2\eta^2 - (\eta + \frac{1}{\eta})^2 \sin^2\theta + 4\cos\theta\sqrt{\eta^2 - \sin^2\theta}} \qquad (2)$$

Similarly, the DoP is described in Equation 3 for specular reflection [Col05].

$$\rho = \frac{2\sin\theta\tan\theta\sqrt{\eta^2 - \sin^2\theta}}{\eta^2 - 2\sin^2\theta + \tan^2\theta} \qquad (3)$$

However, this relation applies to highly specular objects and has been used for metallic objects. The very requirement of known refractive index (η) imposes ***refractive distortion*** if an improper refractive index is used. Moreover, for regions having a zenith angle close to zero, DoP is small, and estimated surface normals are noisy due to low SNR. The readers are requested to kindly refer to [SYC*20] for more details.

2. Implementation Details

The network weights are randomly initialized just at the beginning, and the weights are subsequently updated through the loss functions for 2500 iterations. The framework is trained over 256 × 256 sized images and is implemented in PyTorch [PGC*17] over the NVIDIA RTX 5000 GPU with 16 GB memory. Each object or scene is optimized using Adam optimizer [KB14] with default parameters. Note that the initial learning rate for the optimization is set to 0.001. The network is optimized for 2500 iterations with a learning rate decay of 0.1 after every 250 iterations.

To train the baselines under a self-supervised setting, we use their respective networks for normal estimation and recover the polarization information (ρ, ϕ) from the estimated normals. We replace the normal supervision with the reconstruction error between estimated and ground truth AoP and DoP. We could not enforce the geometry and ratio constraints since the two methods [BGW*20, LQX*22] do not estimate depth. Further, under a supervised setting, SS-SfP is trained over the respective train sets under direct normal supervision and tested over the respective test sets. We stick to the same train-test split as originally given for the respective datasets for fair comparison so that they do not contain images from the same scene.

3. Dataset Details

DeepSfP Dataset [BGW*20] contains 33 objects in total out of which 25 objects are kept for the training while the remaining 8 belong to the test set. Each of the objects are imaged under 3 different lighting conditions (indoor, outdoor-sunny day, and outdoor-cloudy day) and 4 different orientations (front, back, left, and right) such that we have a total of 300 images in the train set.

SPW Dataset [LQX*22] contains the scene-level polarization data for the scenes in the wild. It consists of 522 images from 110 different scenes with diverse object materials and lighting conditions. It contains 403 images in the train set and 119 in the test set.

4. Additional Ablation Experiments

Table 2 reports the variation in the performance of the proposed framework with the number of layers in the encoder and the decoder. The network performance is best for 6 and 5 blocks (each for encoder and decoder) over the DeepSfP and SPW datasets, respectively. However, we finally resorted to 5 blocks for a lighter network.

Further, we chose to use instance normalization after observing a relatively smoother and faster convergence with instance normalization when compared to that with batch normalization, as shown in Figure 3. The values are averaged over the scenes in the test set of the SPW dataset [LQX*22]. Table 1 shows how instance normalization achieves better performance. Observing a slight fallback in performance compared to SPW [LQX*22]), we tried an interesting variant to use self-attention [CLY*21, YTD*21] (see ID 11, Table 1) in the decoder (inspired by SPW [LQX*22]). However, the performance still suffered when compared to the proposed SS-SfP.

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

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

#Encoder and Decoder	MAE (in deg.)	
Blocks	DeepSfP	SPW
2	28.57	33.07
3	19.18	26.35
4	18.02	21.64
5	16.89	**19.77**
6	**16.84**	20.14
8	18.23	21.78
10	19.11	22.16

Table 2: *Ablation experiments for the number of encoder and decoder blocks on the DeepSfP and the SPW datasets. We choose 5 blocks each for the encoder and the decoder in our model according to these quantitative results.*

Figure 4 shows the quality of surface normals estimates under different design choices. We find that the high-frequency details are blurred out if we do not inject the reflectance cues (Figure 4 (a)). Further, since the scenes are mostly diffuse-dominant, the network fails to estimate normals in the specular regions precisely. While the polarization angle ratio constraint does seem to help a bit (without reflectance cues), it also fails at the highly specular regions (Figure 4 (b)). Moreover, adding total variation loss smoothens out the surface normals (Figure 4 (c)). Therefore, we chose to inject the reflectance cues and used the geometric and ratio constraint for better surface normal estimates (Figure 4 (d)).

Why should we reconstruct ϕ and ρ? The simple reason is to allow the network to model a perfect relation between surface normals and the physically measurable quantities: DoP (ρ) and AoP (ϕ) under mixed polarization and handle the underlying ambiguities. Furthermore, the derived quantities - ϕ_d, ϕ_s, ρ_d, and, ρ_s, can only be measured if the surface is purely diffuse or specular, which is seldom the case, and that too with the π-ambiguity. While there are closed-form expressions to establish such a relation for diffuse and specular reflections individually (see Section 1), there is no such model concerning mixed polarization. By reconstructing ϕ and ρ from the surface normal estimates, we force the network to learn their inter-dependencies and further use them to reconstruct the polarization images as per the standard polarization image formation model, as described by Equation 1.

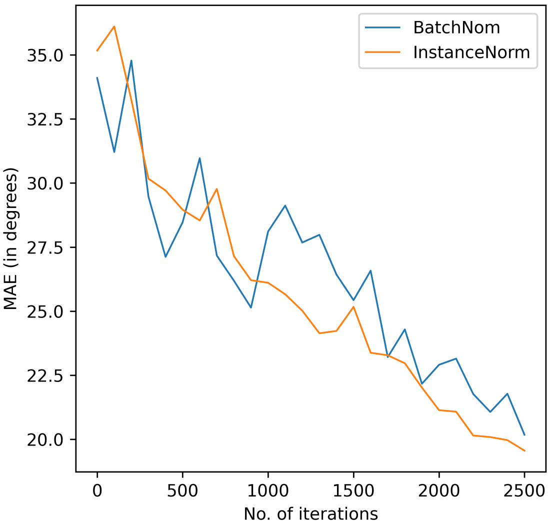

Figure 3: *Learning curve of the proposed framework over the SPW dataset [LQX*22]*

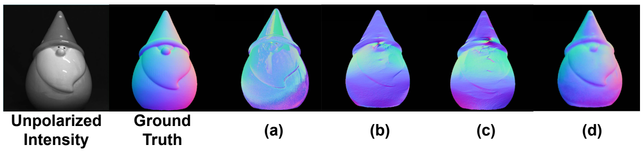

Figure 4: *Qualitative effect of specific design choices on the network performance. (a) without injecting the reflectance cues into the decoder. (b) without reflectance cues, with $\mathcal{L}_{ratio}$. (c) with total variation loss (d) with reflectance cues and $\mathcal{L}_{ratio}$ (ours). Note that $\mathcal{L}_{geo}$ is included in experiments (a), (b), (c), and (d).*

Why do we estimate both surface normals and depth? Since surface normals can be obtained from depth derivatives, we could have a single decoder in a deep network for surface normal esti-

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

mates. However, as discussed in the main paper, we observe poor performance quantitatively through just depth estimation (see ID 8, Table 1). This is attributed to the discontinuities offered by the differentiation step in the depth estimates (and thus, surface normal map) [YZS17]. One way could be to use smoothness constraints such as minimizing total variation. However, they were found to flatten out the normals (and smoothen out the high-frequency details), especially when there is no direct supervision for surface normals. Moreover, such self-supervised frameworks get unstable when applied to real data (such as spikes in the depth maps) [YZS17]. Therefore, we estimate surface normal and depth through two different branches and enforce geometric constraint ($\mathcal{L}_{geo}$) and reflection-dependent ratio constraint ($\mathcal{L}_{ratio}$) for better results (see IDs 12 and 13, Table 1).

5. Additional Qualitative Results

Figure 5 shows the qualitative results on three additional objects (BOX, VASE, and CHRISTMAS) of the DeepSfP dataset [BGW*20] that were not included in the main paper. Further, it also shows the results on the other three objects (DRAGON, FLAMINGO, and HORSE) observed from different views.

Figure 6 shows the results on some additional scenes chosen from the test set of the SPW dataset [LQX*22]. The proposed framework performs better than that of Lei *et al.* for scenes in rows 1 and 2) and almost equally well for the scenes in rows 4 and 5 of Figure 6. To validate the performance under a self-supervised setting, we also show the associated phase angle and degree of polarization in Figure 6.

6. Polarization Angle Ratio Constraint - Derivation

Let us start with the image formation model described in the main paper in Section 4.2 (Equation 7).

$$I(\phi_{pol}) = A_m + B_m cos(2\phi_{pol} - 2\phi) \tag{4}$$

We deploy the findings of [LMSC19,MLC17] through a differential formulation of SfP to model the polarization angle ratio constraint.

6.1. Diffuse Polarization

Let us consider the image formation for diffuse polarisation and expand the cosine term to get the following.

$$I(\phi_{pol}) = A_d + B_d\Big(cos(2\phi_{pol})\Big(2cos^2(\phi_d) - 1\Big) + 2sin(2\phi_{pol})sin(\phi_d)cos(\phi_d)\Big) \tag{5}$$

The first two components of the non-unit normal vector to the surface $\hat{\mathbf{n}} = (n_x, n_y, n_z)$ are proportional to ∇z up to a factor depending on the focal length f such that $\mathbf{n} = \frac{\hat{\mathbf{n}}}{||\hat{\mathbf{n}}||} = \big[g(f)z_x \quad g(f)z_y \quad -1\big]^T = \big[\sin(\theta)\cos(\phi) \quad \sin(\theta)\sin(\phi) \quad \cos(\theta)\big]^T$. By substituting for $cos(\phi)$ and $sin(\phi)$ at $\phi = \phi_d$, we obtain the following.

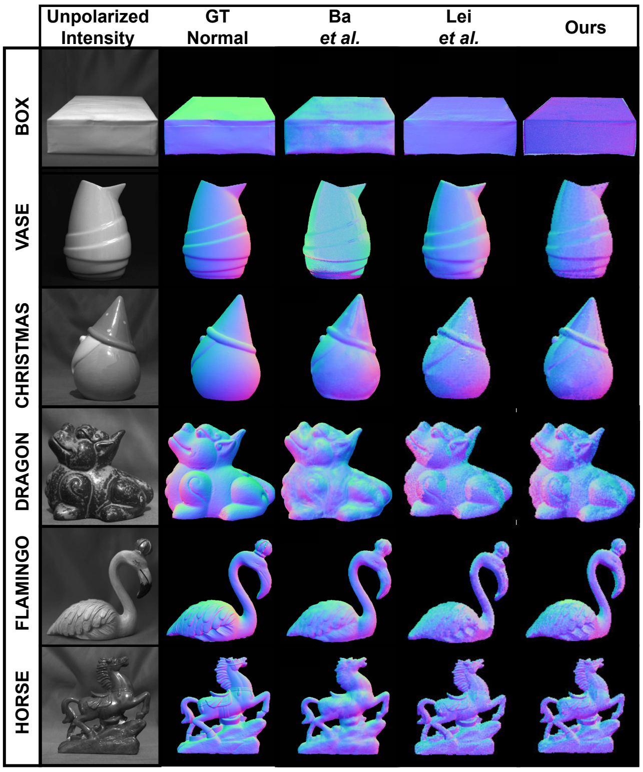

Figure 5: *Additional qualitative results on the test set of the DeepSfP dataset [BGW*20].*

$$I(\phi_{pol}) = A_d + B_d\Big(cos(2\phi_{pol})\Big(2\frac{z_x^2}{||\hat{\mathbf{n}}||^2 sin^2(\theta)} - 1\Big) + 2sin(2\phi_{pol})\frac{z_x z_y}{||\hat{\mathbf{n}}^2 sin^2(\theta)}\Big) \tag{6}$$

Here, for ease of understanding, we consider $g(f) = 1$ (orthographic case). However, the same set of constraints also applies to the perspective case since the factor gets canceled out while taking the ratio. Simplifying the Equation 6, we get,

$$I(\phi_{pol}) - A_d + B_d cos(2\phi_{pol}) = B_d\Big(cos(2\phi_{pol})z_x + sin(2\phi_{pol})z_y\Big)\frac{2z_x}{||\hat{\mathbf{n}}||^2 sin^2(\theta)} \tag{7}$$

Now, we consider the ratio of the Equation 7 evaluated at two polarizer angles ϕ^1_{pol} and ϕ^2_{pol}.

$$\frac{I(\phi^1_{pol}) - A_d + B_d cos(2\phi^1_{pol})}{I(\phi^2_{pol}) - A_d + B_d cos(2\phi^2_{pol})} = \frac{cos(2\phi^1_{pol})z_x + sin(2\phi^1_{pol})z_y}{cos(2\phi^2_{pol})z_x + sin(2\phi^2_{pol})z_y} \tag{8}$$

Cross multiplying and rearranging the Equation 8 gives us the fol-

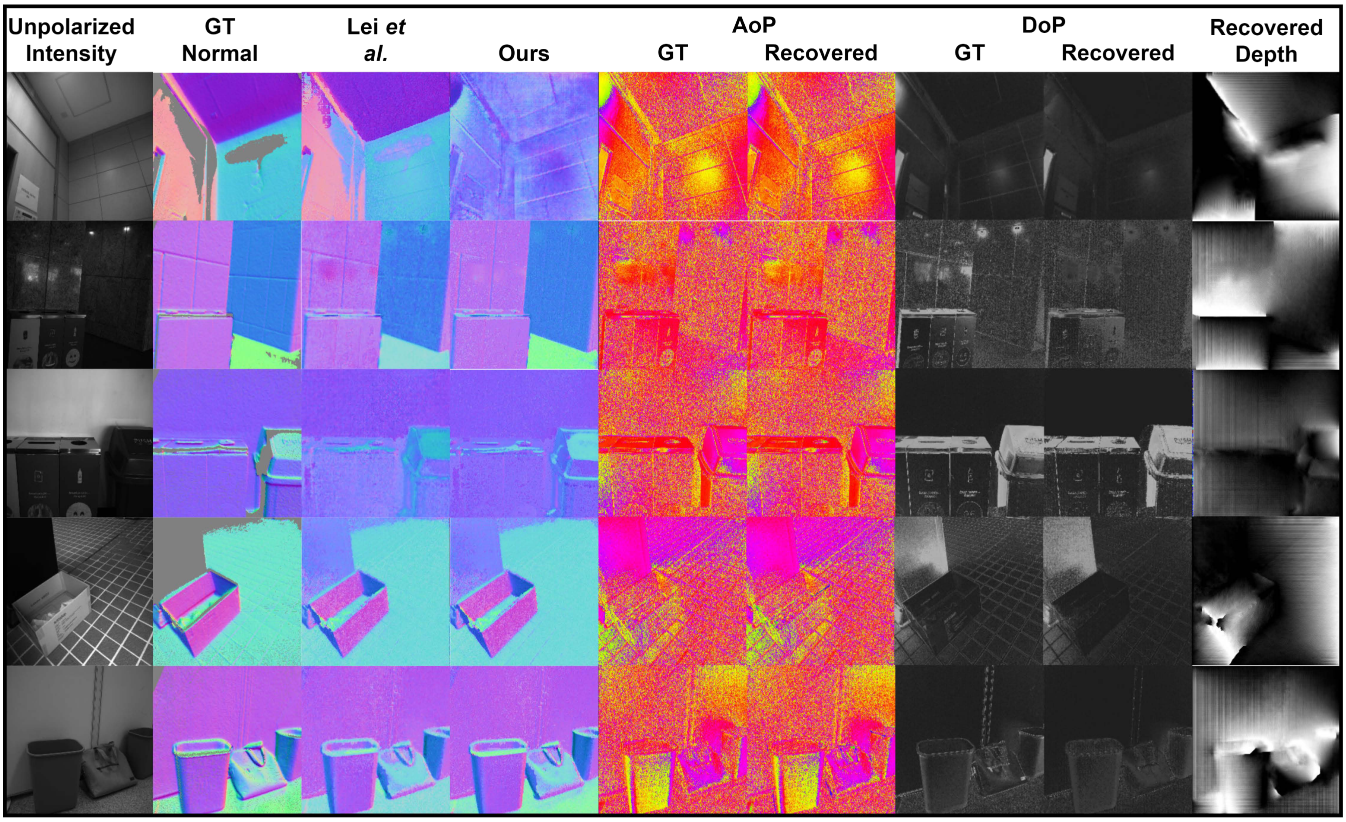

Figure 6: *Additional qualitative results on the test set of the SPW dataset [LQX*22]. We also demonstrate the recovered phase angle (AoP), degree of polarization (DoP), and coarse depth maps.*

lowing.

$$\Big(\big(I(\phi^1_{pol}) - A_d + B_d cos(2\phi^1_{pol})\big)cos(2\phi^2_{pol}) - \big(I(\phi^1_{pol}) - A_d + B_d cos(2\phi^2_{pol})\big)cos(2\phi^1_{pol})\Big)z_x + \Big(\big(I(\phi^1_{pol}) - A_d + B_d cos(2\phi^1_{pol})\big)sin(2\phi^2_{pol}) - \big(I(\phi^1_{pol}) - A_d + B_d cos(2\phi^2_{pol})\big)sin(2\phi^1_{pol})\Big)z_y = 0 \qquad (9)$$

Evaluating Equation 9 at $\phi^1_{pol} = 0$ and $\phi^2_{pol} = \frac{\pi}{4}$, we get the final form as follows.

$$F_d z_x + G_d z_y = 0 \qquad (10)$$

Here, $F_d = \big(-I(\frac{\pi}{4}) + A_d\big)$ and $G_d = \big(I(0) - A_d + B_d\big)$ are the components of the bi-dimensional vector field $v = (F, G)^T$ characterizing the level set in the differential formulation $v^T \nabla z = 0$, as per Equation 9.

6.2. Specular Polarization

We need to account for a $\frac{\pi}{2}$ phase shift for specular polarisation. The bi-dimensional vector field v describing the level-set at a specular pixel has orthogonal direction to those at the diffuse pixel, accounting for the inherent π-periodic ambiguity in the azimuth angle represented by the phase angle ϕ [LMSC19, MLC17] such that the following holds.

$$-\Big(\big(I(\phi^1_{pol}) - A_s + B_s cos(2\phi^1_{pol})\big)sin(2\phi^2_{pol}) + \big(I(\phi^1_{pol}) - A_s + B_s cos(2\phi^2_{pol})\big)sin(2\phi^1_{pol})\Big)z_x + \Big(\big(I(\phi^1_{pol}) - A_s + B_s cos(2\phi^1_{pol})\big)cos(2\phi^2_{pol}) - \big(I(\phi^1_{pol}) - A_s + B_s cos(2\phi^2_{pol})\big)cos(2\phi^1_{pol})\Big)z_y = 0 \qquad (11)$$

Again, evaluating Equation 11 at $\phi^1_{pol} = 0$ and $\phi^2_{pol} = \frac{\pi}{4}$, we get the following constraint.

$$-G_s z_x + F_s z_y = 0 \qquad (12)$$

Here, $F_s = \big(-I(\frac{\pi}{4}) + A_s\big)$ and $G_s = \big(I(0) - A_s + B_s\big)$. We use Equation 10 and 12 as the constraints over diffuse and specular regions, as described in the main paper.

References

[BGW*20] Ba Y., Gilbert A., Wang F., Yang J., Chen R., Wang Y., Yan L., Shi B., Kadambi A.: Deep shape from polarization. In *European Conference on Computer Vision* (2020), Springer, pp. 554–571. 2, 4

[CLY*21] Chen J., Lu Y., Yu Q., Luo X., Adeli E., Wang Y., Lu L., Yuille A. L., Zhou Y.: Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021). 2

[Col05] Collett E.: Field guide to polarization. Spie Bellingham, WA. 2

[KB14] Kingma D. P., Ba J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 2

[LMSC19] Logothetis F., Mecca R., Sgallari F., Cipolla R.: A differential approach to shape from polarisation: A level-set characterisation. *International Journal of Computer Vision 127* (2019), 1680–1693. 4, 5

[LQX*22] Lei C., Qi C., Xie J., Fan N., Koltun V., Chen Q.: Shape from polarization for complex scenes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 12632–12641. 2, 3, 4, 5

[MLC17] Mecca R., Logothetis F., Cipolla R.: A differential approach to shape from polarization. 4, 5

[PGC*17] Paszke A., Gross S., Chintala S., Chanan G., Yang E., DeVito Z., Lin Z., Desmaison A., Antiga L., Lerer A.: Automatic differentiation in pytorch. 2

[SYC*20] Shi B., Yang J., Chen J., Zhang R., Chen R.: Recent progress in shape from polarization. *Advances in Photometric 3D-Reconstruction* (2020), 177–203. 2

[WB93] Wolff L. B., Boult T. E.: Constraining object features using a polarization reflectance model. *Phys. Based Vis. Princ. Pract. Radiom 1* (1993), 167. 1

[YTD*21] Yang G., Tang H., Ding M., Sebe N., Ricci E.: Transformer-based attention networks for continuous pixel-wise prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 16269–16279. 2

[YZS17] Yu Y., Zhu D., Smith W. A.: Shape-from-polarisation: a nonlinear least squares approach. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (2017), pp. 2969–2976. 4