
SfPUEL: Shape from Polarization under Unknown Environment Light *Supplementary Material*

Anonymous Author(s)

Affiliation

Address

email

1 In the supplementary material, we further verify the efficiency of our network initialization strategy
2 and provide details regarding ground truth (GT) normal acquisition in our real-world dataset, more
3 material and normal map estimation results. We compare the model size and inference time of
4 SfPUEL against other methods and then provide the detailed network architecture.

5 A Code for Reproduction

6 We provide the test code, the model checkpoint of SfPUEL, and examples of synthetic data and
7 real-world data at <https://anonymous.4open.science/r/SfPUEL-E243> for reproduction.

8 B Efficiency of Network Initialization with SDM-UniPS Weights

9 We initialize the Polarization Feature Extraction module (PolFEM) and the image-level attention
10 module in Global Context Extractor with the pretrained weights from SDM-UniPS (7). To evaluate
11 the impact of network initialization on framework performance, we initialized these two modules of
12 SfPUEL with Xavier initialization (6) and trained the framework with the same strategy discussed in
13 Sec. 4.3 of the main paper. We find that the network struggled to converge even after being trained
14 for over 80 epochs on the synthetic dataset. It suggests that initializing SfPUEL with the pretrained
15 weights from SDM greatly facilitates the training process.

16 C Ground Truth Normal Acquisition

17 In our real dataset, polarization images and the ground truth normal maps of 6 objects are provided
18 for quantitative evaluation. We acquire the GT normal maps following the guideline of (12). We use
19 EinScan-SP V2 SPECS Desktop 3D Scanner to scan the objects and generate the object meshes. The
20 six objects in our dataset and the scanned meshes are displayed in Fig. 1. We calibrate the polarization
21 camera to get camera intrinsic parameters (14), then conduct the image-mesh alignment to get the
22 camera extrinsic parameters, and finally render the “ground truth” normal in Blender (4).

23 D Material Estimation Results

24 To further validate SfPUEL on material estimation, we provide more results on synthetic data, as
25 shown in Fig. 2. Our method also stably works in scenes containing objects with multiple material
26 types.



Figure 1: Six objects in our dataset and the corresponding scanned meshes.

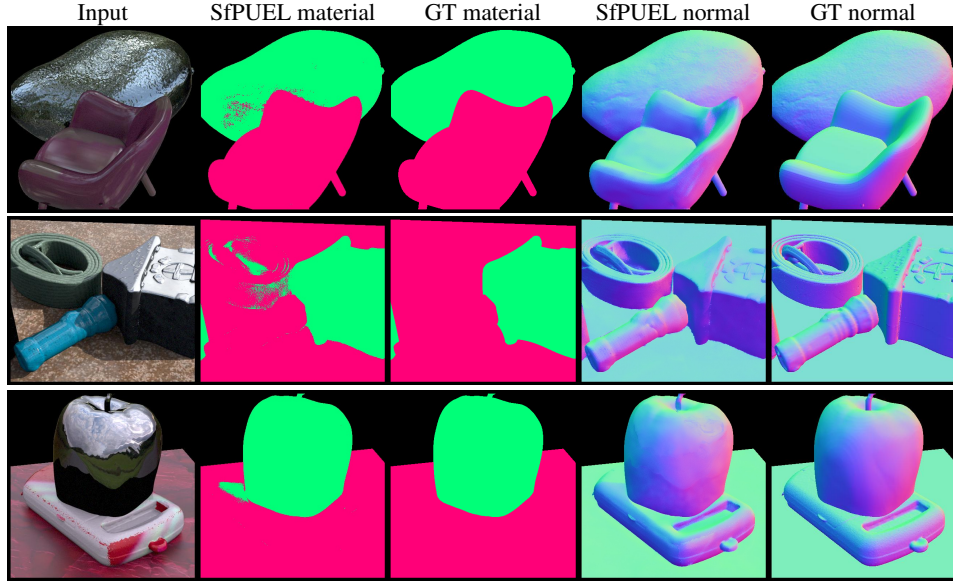


Figure 2: Material estimation of our method on the synthetic data, where red denotes dielectric material and green denotes metallic material.

27 E Model Size and Inference Time

28 We compare the model size (#Param) and test running time of the state-of-the-art methods (*i.e.*,
 29 SfPW (8), DeepSfP (1), UNE (2), DSINE (3), and One-2-3-45 (9)) and our model. The test time
 30 of each method is calculated by processing a single test sample with a resolution of 512, and these
 31 experiments are conducted on the same device (Ubuntu 20.04 LTS with an NVIDIA RTX 3090). The
 32 results are listed in Table 1. One-2-3-45 has the most parameters and takes the longest time in the
 33 inference stage. The test time of our method is slightly longer than other single-shot-based methods
 34 since the current model has not been optimized for computational efficiency. Adopting advanced
 35 lightweight attention mechanisms like efficient additive attention (11) in Global Context Extractor
 may help to reduce our model’s computation complexity.

Table 1: Model size and computational costs comparisons.

Method	SfPW (8)	DeepSfP (1)	One-2-3-45 (9)	UNE (2)	DSINE (3)	SfPUEL
#Param	42.5M	10.8M	1.29G	72.4M	72.6M	141M
Test time	.571s	1.06s	136s	.319s	.423s	1.61s

36

37 F Normal Estimation on Real Data

38 In the main paper, we display the normal predictions of SfPUEL on 4 objects compared to the
 39 state-of-the-art methods. In this section, we provide normal results on the rest two objects in Fig. 3.

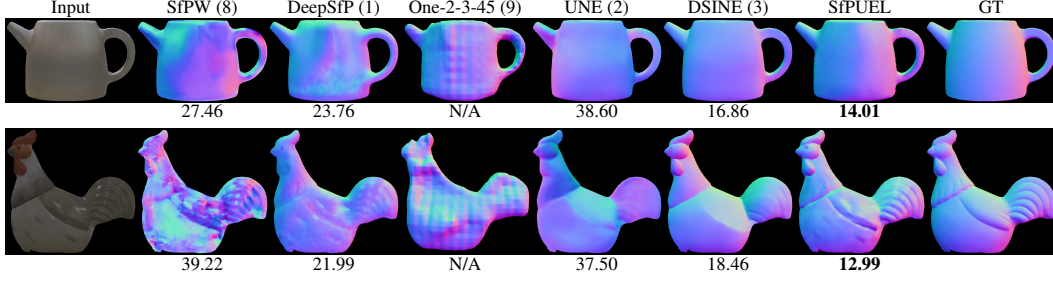


Figure 3: Qualitative results of our method on real data compared to the state-of-the-art approaches. The number below each normal map represents mean angular error.

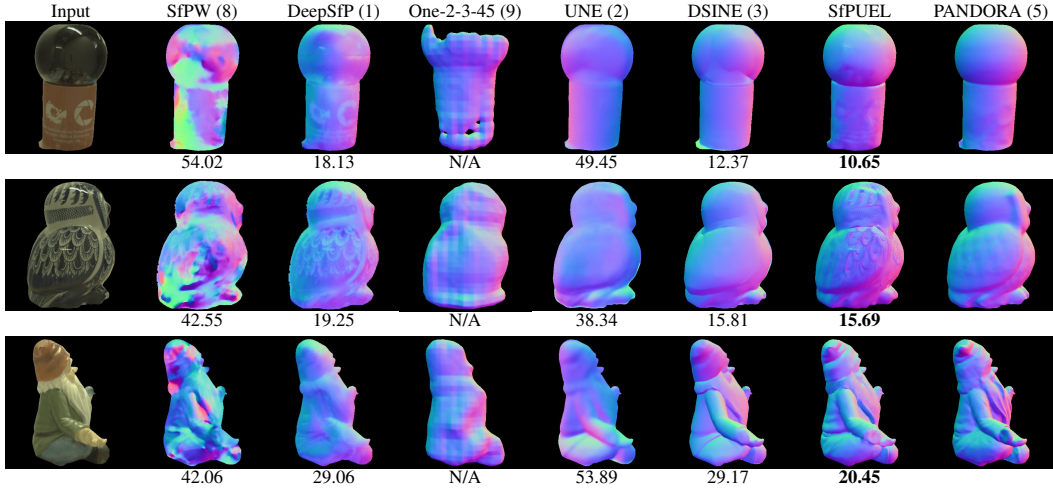


Figure 4: Visual results of our method against previous approaches, including SfPW (8), DeepSfP (1), One-2-3-45 (9), UNE (2), DSINE (3) and PANDORA (5). The number below each result denotes mean angular error.

In addition, we compare SfPUeL to PANDORA (5), the multi-view 3D reconstruction method taking polarization images, as well as SfP (8; 1), 3D generation approach One-2-3-45, and single-image-based approaches (2; 3) on the real data released by (5). The qualitative results are shown in Fig. 4. Our method outperforms previous SfP and single-shot normal estimation approaches. Taking as input single-view polarization images, SfPUeL also produces comparable results against the multi-view method (5).

G Network Details

We tabulate the detailed structures of some SfPUeL modules in Table 2. SfPUeL consists of two main parts: Pol&PS Feature Extractor and Global Context Extractor. Pol&PS Feature Extractor takes as input angle of linear polarization (AoLP) and degree of linear polarization (DoLP) maps, image intensities, polarization images, and the mask, which has two parallel branches: the polarization feature extraction module (PolFEM) and the photometric stereo prior extraction module (PSPEM). PolFEM and PSPEM produce features corresponding to individual input images in a shared-weight manner. The backbone of PolFEM has the same structure as that of PSPEM, and ConvNeXt-T (10) is adopted as the image encoder in the two branches. Pyramid Pooling Module (PPM) of UPerNet (13) is used for fusing hierarchical features from Image Encoder. In PolFEM, we propose to extract features directly from polarization properties using the polarization encoder, whose structure is tabulated in Table 2. For efficient feature fusion between PolFEM and PSPEM, we introduce the DoLP cross-attention block in PSPEM. The polarization features encoded from PolFEM are taken as the query, and the PS features from PSPEM are taken as the key and the value in the cross-attention block. After two-source feature fusion, the extracted features $\mathcal{F}_{\text{PolPS}}$ from Pol&PS Feature Extractor

Table 2: The detailed structures of the polarization encoder in PolFEM, adopted zero convolutional layers, and prediction heads in SfPUEL. H and W denote the height and the width of the input tensor, respectively, and N_{pixel} represents the number of sampled pixels.

Layer Description	Output Tensor Size
Polarization Encoder	
Conv2d(kernel=3, padding=1)	$(H, W, 16)$
SiLU	$(H, W, 16)$
Conv2d(kernel=3, padding=1)	$(H, W, 16)$
SiLU	$(H, W, 16)$
Conv2d(kernel=3, padding=1, stride=2)	$(\frac{1}{2}H, \frac{1}{2}W, 32)$
SiLU	$(\frac{1}{2}H, \frac{1}{2}W, 32)$
Conv2d(kernel=3, padding=1)	$(\frac{1}{2}H, \frac{1}{2}W, 32)$
SiLU	$(\frac{1}{2}H, \frac{1}{2}W, 32)$
Conv2d(kernel=3, padding=1, stride=2)	$(\frac{1}{4}H, \frac{1}{4}W, 96)$
SiLU	$(\frac{1}{4}H, \frac{1}{4}W, 96)$
Conv2d(kernel=3, padding=1)	$(\frac{1}{4}H, \frac{1}{4}W, 96)$
SiLU	$(\frac{1}{4}H, \frac{1}{4}W, 96)$
Zero Convolution Layer	
Conv2d(kernel=1, padding=0)	$(H, W, \text{dim_input})$
Normal-MLP	
Linear(in_dim=384, out_dim=192)	$(N_{\text{pixel}}, 192)$
ReLU	$(N_{\text{pixel}}, 192)$
Linear(in_dim=192, out_dim=3)	$(N_{\text{pixel}}, 3)$
Material-MLP	
Linear(in_dim=384, out_dim=192)	$(N_{\text{pixel}}, 192)$
ReLU	$(N_{\text{pixel}}, 192)$
Linear(in_dim=192, out_dim=2)	$(N_{\text{pixel}}, 2)$

are fed to 5 cascaded image-level self-attention blocks. The image-axis self-attention block has a vanilla transformer structure composed of multi-head self-attention blocks, layer normalization, and feed-forward networks, producing image-level enhanced feature \mathcal{F}_{enh} . Then, \mathcal{F}_{enh} are sampled spatially, and we use cross-attention to query per-pixel features and conduct pixel-level self-attention to generate the global context features. Finally, the global features are fed into two MLPs to predict normal vectors and material logits.

References

- [1] Yunhao Ba, Alex Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. Deep shape from polarization. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.
- [2] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021.
- [3] Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [4] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, 2024.
- [5] Akshat Dave, Yongyi Zhao, and Ashok Veeraraghavan. PANDORA: Polarization-aided neural decomposition of radiance. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022.
- [6] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2010.
- [7] Satoshi Ikehata. Scalable, detailed and mask-free universal photometric stereo. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [8] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

- 89 [9] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su.
 90 One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In
 91 *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, 2023.
- 92 [10] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining
 93 Xie. A convnet for the 2020s. In *Proceedings of Conference on Computer Vision and Pattern
 94 Recognition (CVPR)*, 2022.
- 95 [11] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang,
 96 and Fahad Shahbaz Khan. Swiftformer: Efficient additive attention for transformer-based
 97 real-time mobile vision applications. In *Proceedings of International Conference on Computer
 98 Vision (ICCV)*, 2023.
- 99 [12] Boxin Shi, Zhipeng Mo, Zhe Wu, Dinglong Duan, Sai-Kit Yeung, and Ping Tan. A benchmark
 100 dataset and evaluation for non-lambertian and uncalibrated photometric stereo. *Transactions on
 101 Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.
- 102 [13] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing
 103 for scene understanding. In *Proceedings of European Conference on Computer Vision (ECCV)*,
 104 2018.
- 105 [14] Zhengyou Zhang. A flexible new technique for camera calibration. *Transactions on Pattern
 106 Analysis and Machine Intelligence (TPAMI)*, 2000.