Shape from Polarization for Complex Scenes in the Wild

Chenyang Lei[*1] Chenyang Qi[*1] Jiaxin Xie[*1] Na Fan[1] Vladlen Koltun[2] Qifeng Chen[1]
[1]HKUST [2]Apple

Abstract

We present a new data-driven approach with physics-based priors to scene-level normal estimation from a single polarization image. Existing shape from polarization (SfP) works mainly focus on estimating the normal of a single object rather than complex scenes in the wild. A key barrier to high-quality scene-level SfP is the lack of real-world SfP data in complex scenes. Hence, we contribute the first real-world scene-level SfP dataset with paired input polarization images and ground-truth normal maps. Then we propose a learning-based framework with a multi-head self-attention module and viewing encoding, which is designed to handle increasing polarization ambiguities caused by complex materials and non-orthographic projection in scene-level SfP. Our trained model can be generalized to far-field outdoor scenes as the relationship between polarized light and surface normals is not affected by distance. Experimental results demonstrate that our approach significantly outperforms existing SfP models on two datasets. Our dataset and source code will be publicly available at https://github.com/ChenyangLEI/sfp-wild.

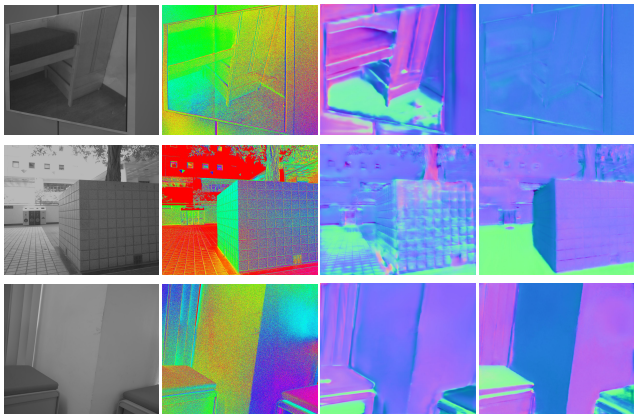


Figure 1. Our method can estimate dense scene-level surface normals from a single polarization image. Polarization can provide effective cues for obtaining more accurate results. In the first row, polarization provides geometry cues for our model so that it is not fooled by objects in the printed image on a wall. In the second and third rows, polarization provides guidance for planes with different surface normals even when their materials are quite similar. I_{un}: unpolarized image; ϕ: angle of polarization.

1. Introduction

Accurate surface normal estimation in the wild can provide valuable information about a scene's geometry and can be used in various computer vision tasks, including segmentation [19], 3D reconstruction [26], and many others [22, 33]. Therefore, normal estimation is an important task studied for a long time. However, estimating high-quality normals in the wild is still an open problem. Various techniques such as photometric stereo [9, 10] can produce high-frequency normals, but most of them only provide short-range object-level normal maps. Active depth sensors can be another approach to obtaining normals from depth maps, but the corresponding depth maps are often sparse (LiDAR) or noisy (time-of-flight, structured light) so they can not estimate normals reliably. Also, the depth range of active sensors is limited.

In this work, we are interested in estimating surface normal from a single polarization image for complex scenes in the wild. Since the polarization of light changes differently when the light interacts with the surfaces of different shapes and materials (governed by the Fresnel equations [12]), the polarization images can provide dense surface orientation cues from the polarized light perceived at each pixel. Also, compared with the active sensors and object-level normal estimation techniques (e.g., photometric stereo), the polarization camera is a passive sensor that is not constrained to a specific depth range. Thus polarization images are promising data sources for accurate normal estimation in the wild.

However, estimating normals from a polarization image for complex scenes (scene-level SfP) is challenging. To the best of our knowledge, no existing SfP work focuses on complex scenes, and several challenges are yet to be solved. Firstly, polarization contains ambiguities from unknown information such as object materials and reflection

[*]Joint first authors

types [12]. Object-level SfP methods approach these ambiguities by utilizing various cues (e.g., shading [44]) or making restrictive assumptions (e.g., known albedo [35]), which are unfeasible for multiple-object scenes because of the variabilities of material properties and complexities of reflections. Secondly, while some works [4, 29] demonstrate the potential of combining convolutional neural networks and polarization cues in estimating normals for unknown materials, there are only object-level [4] or synthetic data [29] for training, which are not sufficient for scene-level SfP. Finally, scene-level SfP brings up another challenge. The viewing direction influences the measured polarization information. Previous object-level SfP approaches ignore the impact of viewing direction since they assume orthographic projection by placing objects at the center of an image, which does not hold for scene-level SfP.

To solve the challenge of lacking real-world scene-level polarization data, we construct the first real-world scene-level SfP dataset that contains diverse complex scenes. Building such a new dataset is necessary because the existing DeepSfP dataset [4] only contains a single object per image and the dataset by Kondo et al. [29] is synthetic and not publicly available.

Due to the challenges of scene-level SfP, the performances of previous learning-based SfP works [4, 29] are not satisfactory when they are trained on our scene-level data. To improve the performance of scene-level SfP in the wild, we adopt three novel designs in our model. First, we introduce multi-head self-attention [45] in a convolutional neural network (CNN) for SfP. Multi-head self-attention utilizes the global context of an image, which helps the CNN resolve the local ambiguities in polarization cues. Second, to handle non-orthographic projection for scene-level SfP, the neural network must be aware of the viewing direction of each pixel since the convolution operation is translation invariant. We thus propose a simple but critical technique that improves the performance of SfP methods on scene-level data: providing per-pixel viewing encoding to the neural network. Finally, as an additional contribution, we design a novel polarization representation, which is effective and considerably more efficient than the representations in prior work [4].

We compare our approach with various state-of-the-art methods. Experimental results show that our model can generate a high-quality normal map from a single polarization image (Fig. 1) and can generalize beyond the depth range of the training data. In summary, our contributions are as follows.

- We construct the first real-world SfP dataset containing paired input polarization images and ground-truth normal maps in complex scenes.
- Our proposed shape-from-polarization approach is the first one trained on complex real-world scene-level data and also the best-performing one for normal estimation from polarization in the wild.
- Technically, we introduce three novel designs to scene-level SfP: viewing encoding that can handle the challenge of non-orthographic projection in scene-level SfP, a dedicated network architecture that adopts multi-head self-attention for SfP, and a practical polarization representation that is effective and efficient.

2. Related Work

Shape from polarization. The polarization of light changes when the light interacts with a surface, which can be described by the Fresnel equations using the geometry and materials of objects [12]. Shape from polarization (SfP) works [2, 38, 42] utilize this effect to estimate the surface normal of objects. Since the polarization state is affected by various factors simultaneously, early SfP methods usually enforce assumptions of reflection types and materials to constrain the problem. For example, Rahmann et al. [42] assume pure specular reflection and some works [1, 38] assume pure diffuse reflection.

Various cues and techniques have been explored to resolve the ambiguities in this problem. Atkinson et al. [3] use shading from two views. Baek et al. [5] perform joint optimization of appearance, normals, and refractive index. A coarse depth map from a depth sensor [25], two-view stereo [21, 56], reciprocal image pairs [16], or multi-view stereo [13, 37], can also be served to disambiguate the problem. For single-view SfP, some methods combine photometric stereo [1] or shading information [35, 44] with SfP. Also, some works try to solve this problem under specific illumination conditions (e.g., front-flash illumination [15] and sunlight under the clear sky [23]). Unlike these works, our approach aims to estimate surface normal in the wild without specific assumptions or additional tools.

Deep learning is proven effective in solving the ambiguities of object-level SfP. Ba et al. [4] collect a real-world object-level dataset and train a CNN to obtain normals from polarization, significantly outperforming physics-based SfP. Instead of collecting real-world data, Kondo et al. [29] create a synthetic dataset of polarization images with a new polarimetric BRDF model. However, these approaches have not studied complex scenes in the wild due to the lack of real-world scene-level data. To address this issue, we propose the first real-world scene-level SfP dataset. Besides, we also notice existing frameworks [4, 29] cannot achieve satisfactory results on our dataset due to the challenges that emerge in scene-level SfP, and we propose effective solutions to these challenges.

Surface normals from an RGB image. Even though RGB data does not directly contain geometry cues for objects

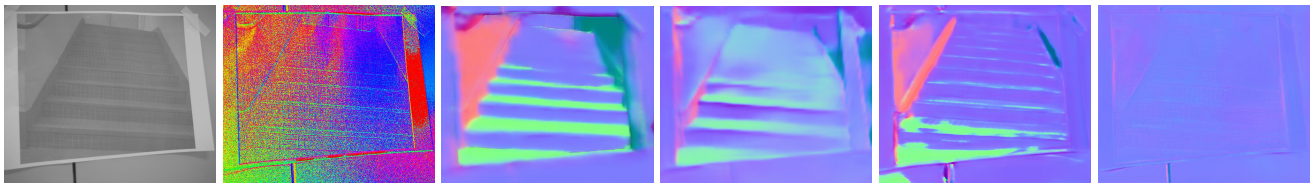


Figure 2. **The importance of polarization.** Note that polarization conveys the underlying physical shape while RGB-based methods [17, 51] are distracted by the semantics in a printed picture attached to a wall.

such as polarization data, estimating the surface normal from a single RGB image is feasible, especially with the advent of deep learning. The related works [6, 7, 31, 48, 52] train a neural network using a large amount of RGB-surface normal paired data, including real-world indoor dataset [43] or synthetic dataset [14]. However, without the guidance of physics-based cues, these learning-based approaches mainly rely on semantic cues in the image, which leads to performance degradation when they are applied to data out of the training distribution (e.g., from indoor to outdoor data [11] and from gravity-aligned to tilted images [17]).

Some approaches attempt to utilize the relevance between surface normals and other information (e.g., depth, semantic information and shading). It has been shown that better surface normal estimation can be achieved by simultaneously estimating geometric information, such as depth [40, 41], local principal axes [22], Manhattan label map [47], planes and edges [46]. Eigen and Fergus [18] and Zhang et al. [53] jointly predict depth, normals, and semantics, exploiting affinity between these three modalities. Zamir et al. [51] consider the consistency of normals and other attributes, such as shading, depth, occlusion, and curvature. Surface normal estimation is also an essential element in inverse rendering, which aims to recover normals, reflectance, and illumination from one image [8, 33, 50] or multiple images [27, 54]. However, according to our experiments, these approaches mainly depend on the semantic information of images for normal estimation. As a comparison, our method can better recover the physical geometry with the polarization cues, as shown in Fig. 2.

3. SPW Dataset

We construct the SPW (Shape from Polarization in the Wild) dataset, the first real-world dataset that contains scene-level ground-truth surface normals for polarization images in the wild. Table 1 provides a comparison between SPW and prior SfP datasets. The only existing real-world SfP dataset is DeepSfP [4]; however, it only provides ground-truth normals on masked objects. Kondo et al. [29] built a big SfP dataset, but it is synthetic and not publicly available.

Dataset	Level	Collection	Size	Resolution	Public
DeepSfP [4]	Object	Real-world	263	1224×1024	✓
Kondo [29]	Scene	Synthetic	44305	256×192	×
Ours	Scene	Real-world	522	1224×1024	✓

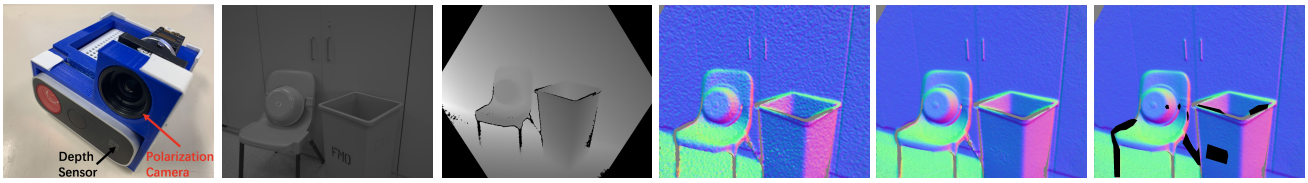
Table 1. **Comparison among different datasets.** DeepSfP [4] is real-world but focuses on object-level, Konda et al. [29] has a big dataset size, but it is synthetic and not publicly available. Ours is the first real-world scene-level SfP dataset.

Our dataset consists of 522 sets of images from 110 different scenes containing diverse object materials and lighting conditions, and each scene includes 3-7 sets of images with different depths and viewing directions. A polarization image and the corresponding normal map are provided in each set. In addition to these image sets with ground truth normals, we also capture a separate set of images in outdoor scenes. Since most depth sensors cannot easily acquire dense depth for faraway scene content, these images are used for perceptual evaluation only, to assess generalization beyond the depth range of the training data.

Fig. 3 shows our data preparation pipeline. The pipeline can be divided into the following four parts.

a) Devices. Since there is no existing polarization-depth camera, we need to choose a depth sensor to capture dense scene-level depth. We notice that most LiDAR cannot produce dense point clouds efficiently, and the 360-degree rotating device used in [4] can only reconstruct small objects. Hence, we use a ToF sensor (Azure Kinect) to capture scene-level depth, and this depth sensor's resolution is 640×576. For polarization, we use a PHX050S-P polarization camera that can capture four polarization images with polarizer angles of $0°, 45°, 90°, 135°$ in a single shot, and the resolution of this polarization camera is 1224×1024. We fix the two sensors with a custom mount to make sure the camera pose between the depth sensor and polarization camera is the same in each capture, as shown in Fig. 3(a).

b) Depth-polarization alignment. We obtain the intrinsic and the initial extrinsic parameters between the polarization and depth sensor from stereo calibration. We then use



(a) Capture setup (b) Polarization image (c) Depth map (d) Coarse normal (e) Denoised normal (f) Final normal

Figure 3. **Our capture setup (a)** fixes a polarization camera (red arrow) and a depth sensor (black arrow) with a custom mount. The polarization camera captures polarization images (b), and the depth sensor collects scene-level depth (c). We use PCA-based normal estimation to obtain normal maps from the depth data, as shown in (d) and (e). (d) uses a single depth map, and (e) uses a median-denoised depth map. Post-processing the denoised normal map in (e) yields the final normal map shown in (f), there we exclude normals in areas where the depth sensor returns the inaccurate or sparse depth (e.g., dark region).

coordinate descent to improve depth-polarization alignment further. Specifically, we optimize the extrinsic parameters with fixed intrinsic parameters to minimize the projection error between the reprojected RGB image and the polarization image. The optimized extrinsic parameters are used to produce polarization-aligned depth.

c) Depth denoising. Since surface normals computed from a single polarization-aligned depth map are noisy, we capture 50 depth images with a stationary setup and compute the median at each pixel to reduce noise in the depth map. Finally, we generate a point cloud from the denoised depth map and calculate the surface normals from the aligned point cloud using Principal Component Analysis (PCA) from the Open3D library [55]. As shown in Fig. 3(d,e), denoising the depth map yields much cleaner normals.

d) Post-processing. Even though we get high-quality normals by the above steps, we further improve the quality by excluding normals in areas where the depth sensor returns inaccurate values, such as dark and occluded regions. We also exclude normals on thin structures where the depth sensor only captures very sparse point clouds, such as chair legs or wires. The final normals are shown in Fig. 3(f).

4. Method

In this paper, we consider linear polarization. A polarization camera can measure the intensity of light $\mathbf{I}^{\phi_{pol}}$ passing a polarizer [12, 56], which is determined by the polarizer angle ϕ_{pol} and the polarization of the light:

$$\mathbf{I}^{\phi_{pol}} = \mathbf{I}_{un}\{1 + \rho\cos(2\phi - 2\phi_{pol})\}, \tag{1}$$

where ϕ is the angle of polarization (AoP), ρ is the degree of polarization (DoP), $\mathbf{I}_{un}$ is the unpolarized intensity of light. ϕ, ρ, and $\mathbf{I}_{un}$ can be computed from images with different polarizer angles by [30, 32].

Degree of polarization (DoP) ρ contains cues for the viewing angle $\theta_{\mathbf{v}}$ between surface normal $\mathbf{n}$ and viewing direction $\mathbf{v}$. Specifically, the DoP ρ is decided by the viewing angle $\theta_{\mathbf{v}}$, the refractive index η of the object and reflection type r (specular or diffuse reflection). More details are provided in the supplement.

Angle of polarization (AoP) ϕ is the projection of the polarization direction $\mathbf{d}$ on the image plane. In terms of physical properties, $\mathbf{d}$ is always parallel or perpendicular to the incidence plane, which is defined by surface normal $\mathbf{n}$ and viewing direction $\mathbf{v}$. There are two ambiguities for the polarization angle: π-ambiguity and diffuse/specular-ambiguity. The π-ambiguity is because ϕ is from 0 to π and there is no difference between ϕ and $\phi + \pi$ (Eq. 1). The reflection type causes the diffuse/specular-ambiguity: the polarization direction is parallel or perpendicular to the incidence plane respectively for diffuse/specular dominant reflection.

4.1. Overview

Fig. 4 provides an overview of our approach. The raw polarization image $\mathbf{I} \in \mathbb{R}^{H\times W\times 4}$ consists of four polarization images $\mathbf{I}^{\phi_{pol}} \in \mathbb{R}^{H\times W\times 1}$ under four polarizer angles $\phi_{pol} \in \{0, \pi/4, \pi/2, 3\pi/4\}$. We firstly compute a polarization representation $\mathbf{P}$ for normal estimation (Sec. 4.2). Then, to handle the perspective projection for scene-level SfP, we provide the viewing encoding $\mathbf{V}$ as an extra input (Sec. 4.3). At last, we predict the normal $\hat{\mathbf{n}}$ from all the provided information with our designed architecture (Sec. 4.4).

4.2. Polarization representation

Having a proper polarization representation $\mathbf{P}$ as the input to a neural network can effectively improve the performance of SfP [4, 29]. Kondo et al. [29] directly compute AoP ϕ, DoP ρ and I_{un} as the polarization representation. DeepSfP [4] calculates possible SfP solutions under the assumption of orthographic projection:

$$\mathbf{n} = (\sin\theta\cos\alpha, \sin\theta\sin\alpha, \cos\theta)^{\intercal}, \tag{2}$$

where θ and α are the zenith angle and azimuth angle computed from DoP ρ and AoP ϕ, respectively. While it is effective, computing their polarization representation is quite time-consuming, as reported in their paper [4].

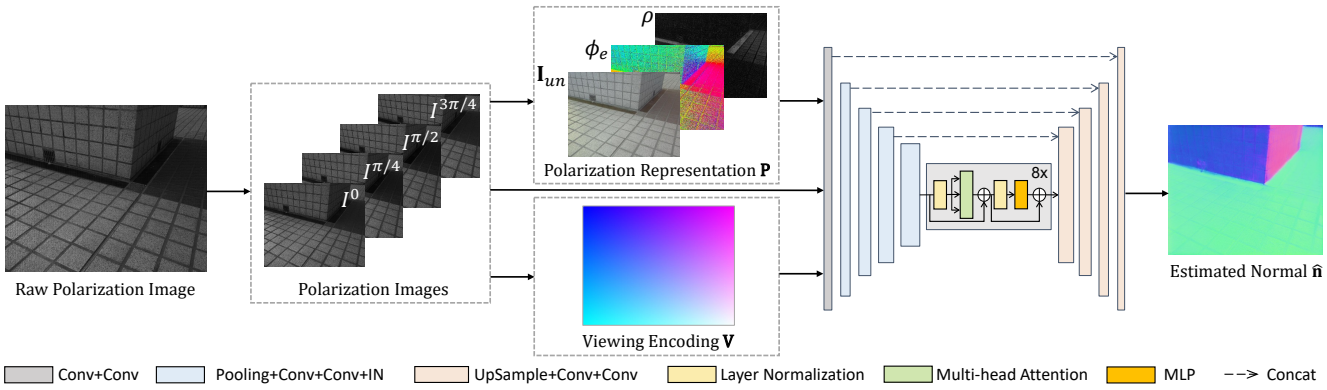


Figure 4. **An overview of our approach.** The input to our network includes three parts: (1) Polarization images $\mathbf{I}^0, \mathbf{I}^{\pi/4}, \mathbf{I}^{\pi/2}, \mathbf{I}^{3\pi/4}$. (2) The polarization representation $\mathbf{I}_{un}$, ϕ_e, and ρ computed from polarization images. (3) The viewing encoding $\mathbf{V}$ is vital to handle the perspective projection for scene-level SfP. The concatenated inputs are fed into the neural network to output an estimated normal $\hat{\mathbf{n}}$.

We propose a new polarization representation $\mathbf{P}$ that is efficient and more effective compared with existing polarization representations [4, 29], as shown in our experiments (Table 4). $\mathbf{P} \in \mathbb{R}^{H\times W\times 4}$ consists of $\mathbf{I}_{un}, \phi_e, \rho$, where ϕ_e is the encoded AoP:

$$\phi_e = (cos2\phi, sin2\phi). \tag{3}$$

The encoded AoP ϕ_e is designed to address a weakness of raw AoP ϕ. For example, given two polarization angles $0°$ and $179°$, the distance between them should be $1°$ in physics for polarization. However, in the calculated ϕ space, the difference is $179°$. Encoding helps solve the weakness of raw AoP representation since there is no difference between ϕ and $\phi + \pi$ in the encoding space.

We input the DoP ρ as cues for solving specular/diffuse ambiguity since the DoP ρ is usually large when specular reflection dominates. This strategy improves the performance but does not fully resolve the specular/diffuse ambiguity.

Importance of polarization. Polarization contains useful cues about physical 3D information of objects based on real-world reflection. Thus, utilizing polarization can improve the fidelity of estimated normals, especially for areas with rare or wrong semantic information. Fig. 2 shows an example about the advantage of polarization: given an image printed on a flat sheet of paper, the RGB-based baselines are distracted by the content of the image and fail to predict correct normals for the physical content of the scene (i.e., the flat paper). Polarization provides an alternative modality that can convey the true shape of objects in the scene. Hence, the polarization can give a robust cue to distinguish that the wall (paper) is exactly flat.

4.3. Viewing encoding

We introduce the viewing encoding $\mathbf{V}$ to account for non-orthographic projection in scene-level SfP, which contains the viewing direction cues for every pixel of the polarization representation. Previous object-level SfP approaches [3, 4] assume viewing directions are $(0, 0, 1)^\intercal$ for all pixels (i.e., orthographic projection) since an object is always put at the image center. However, the viewing direction is spatially varying in scene-level SfP, and the polarization representation is heavily influenced by the viewing direction. As shown in Fig. 5, for pixels with the same material and surface normal, their polarization representations are quite different. Also, since the CNN is translation-invariant, it is hard for CNN to know the viewing direction without explicitly providing it to the CNN. We thus propose to input the viewing encoding to the CNN.

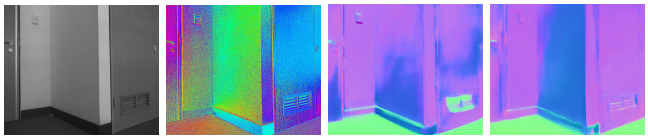


Figure 5. **An analysis of our proposed viewing encoding $\mathbf{V}$.** The polarization representation is affected by spatial-varying viewing directions in scene-level SfP. Enforcing the viewing encoding $\mathbf{V}$ can effectively calibrate the impact of viewing direction on the polarization representation.

A direct representation of viewing encoding is the viewing direction, which is computed from the intrinsic parameters of the polarization camera. If the intrinsic parameters of the camera are not available, we can also use the 2D coordinate (u, v) of each pixel and normalize it to $[-1, 1]$ as input [34]. Fig. 5 presents an example for the effectiveness of our viewing encoding.

Discussion. Note that our viewing encoding is different from the positional encoding used in NeRF [36] or transformers [45]. Our design is inspired by the fact that per-

Method	Task	Angular Error ↓			Accuracy ↑		
		Mean	Median	RMSE	11.25°	22.5°	30.0°
Miyazaki et al. [38]	Physics-based SfP	55.34	55.19	60.35	2.6	10.4	18.8
Mahmoud et al. [35]	Physics-based SfP	52.14	51.93	56.97	2.7	11.6	21.0
Smith et al. [44]	Physics-based SfP	50.42	47.17	55.53	11.0	24.7	33.2
DeepSfP† [4]	Learning-based SfP	28.43	24.90	33.17	18.8	48.3	62.3
Kondo et al.† [29]	Learning-based SfP	28.59	25.41	33.54	17.5	47.1	62.6
Ours	Learning-based SfP	**17.86**	**14.20**	**22.72**	**44.6**	**76.3**	**85.2**

Table 2. **Quantitative evaluation on the SPW dataset.** Our approach outperforms all baselines by a large margin on all evaluation metrics. †: our implementation.

Method	Mean Angular Error↓
Miyazaki et al. [38]	43.94
Mahmoud et al. [35]	51.79
Smith et al. [44]	45.39
DeepSfP [4]	18.52
Ours	**14.68**

Table 3. **Quantitative evaluation on the DeepSfP dataset [4].** Our approach obtains the best score. The results of other baselines are collected from the official results in DeepSfP [4].

pixel viewing directions influence polarization. Besides, viewing encoding yields better performance than the positional encoding in our experiments.

4.4. Network architecture and training

To handle the ambiguities that exist in polarization cues, we introduce multi-head self-attention [45] to SfP for utilizing the global context information. As shown in Fig. 4, the self-attention block is added in the bottleneck of an Encoder-Decoder architecture [39]. Different from similar architectures that combine CNN encoder with transformer [24, 49], we remove the linear projection layer since the CNN encoder already extracts the embeddings. Besides, similar to position embedding of transformer [24, 49], our viewing encoding can provide the position information to self-attention, and we thus remove the position embedding. Finally, we add instance normalization to the encoder since we notice it helps convergence.

We adopt a cosine similarity loss [4] for training. We implement our model in PyTorch. The model is trained for 1000 epochs with batch size 16 on four Nvidia Tesla V100 GPUs, each with 16 GB memory. We use the Adam optimizer [28] with initial learning rate 1e-4 and we adopt a cosine decay scheduler for the learning rate. The learning rate is linearly scaled with the batch size. We crop images to 512×512 patches in each iteration for data augmentation.

5. Experiments

5.1. Experimental setup

Evaluation metrics. Following previous surface normal estimation works [6, 48], we adopt six widely used metrics. The first three are *Mean*, *Median*, and *RMSE* (lower is better ↓), which are the mean, median, and RMSE of angular errors. The last three are $11.25°$, $22.5°$, and $30.0°$ (higher is better ↑), and each shows the percentage of pixels within a specific angular error.

Datasets. We use two datasets in the experiments.

- **DeepSfP [4]**. DeepSfP is the only publicly available SfP dataset that contains real-world ground-truth surface normals. There is only one object in each image, but the surface normal is high-quality. We use the train/test split provided in the original paper [4].
- **SPW**. We use our SPW dataset, presented in Sec. 3. We use 403 and 119 images for training and evaluation, respectively. Train and test sets do not include the images from the same scene to avoid overfitting. We also use the far-field data for perceptual evaluation. randomly divided based on scenes (instead of images).

5.2. Comparison to SfP baselines

Our approach is compared with three physics-based SfP methods (Miyazaki et al. [38], Mahmoud et al. [35], and Smith et al. [44]) and two learning-based SfP methods (DeepSfP [4] and Kondo et al. [29]). The source code and results of DeepSfP [4] and Konda et al. [29] are not available. We reimplement these two approaches and retrain their models on our SPW dataset.

Table 2 presents the quantitative results of all the methods on our SPW dataset. Our approach outperforms all baselines by a large margin on all metrics.

Fig. 6 provides a qualitative comparison on images from the SPW dataset. Our estimated surface normal maps are more accurate. Besides, Our approach can produce high-quality normals while other methods do not.

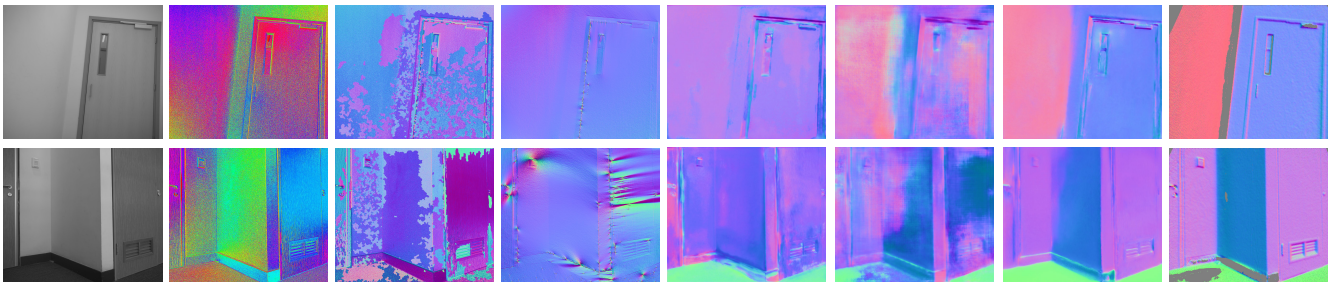


Figure 6. Qualitative comparison between our approach and other shape from polarization methods baselines [4, 29, 38, 44] on the SPW dataset [4].

Table 3 presents the quantitative results on the public DeepSfP dataset [4], in which we also achieve the best performance. Our approach reduces the mean angular error by 20% compared to the second-best result reported by DeepSfP [4].

For physics-based SfP [35,38,44], since the assumptions of these methods do not hold in the wild, the quantitative accuracy of these approaches is low on the SPW dataset. They cannot obtain satisfying performance on the DeepSfP dataset, either. For example, Mahmoud et al. [35] assume a distant light source, which is not common in the real world (e.g., multiple light sources can exist in a room). As for the learning-based SfP methods [4, 29], our approach is the best-performing one and we analyze the designs (i.e., polarization representation, viewing encoding and architectures) that contribute to our model in Sec. 5.4.

5.3. Generalization to outdoor scenes

Although our model is trained on near-field depth estimated by a Kinect camera, it can generalize to outdoor scenes with distances far beyond the Kinect depth range. This is illustrated qualitatively in Fig. 7. Quantitative results are not provided due to the lack of ground-truth normals in this regime. This generalization is possible because the relationship between polarized light and surface normals is not affected by distance. Thus our model that learns to estimate normals from near-field polarization data can generalize to outdoor scenes.

5.4. Controlled experiments

5.4.1 Polarization representation

The experiments are conducted on both DeepSfP dataset [4] and SPW dataset. We remove the polarization information or replace our proposed polarization representation with other representations as input to our model. Table 4 provides the quantitative results of various polarization representations. Utilizing our polarization representation reduces the mean angular error by 9 °. Besides, we obtain the lowest MAE on both datasets and the running time of our representation is much shorter than DeepSfP [4].

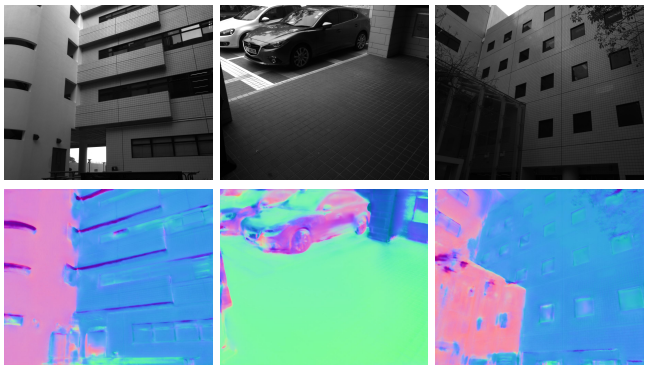


Figure 7. **Our results on outdoor scenes.** Although our model is trained on near-field content, it appears to successfully generalize to large-scale outdoor scenes.

Polarization representation $\mathbf{P}$	Mean Angular Error↓ SPW	DeepSfP [4]	Time(s)
Without polarization	27.52	19.14	0.000
Raw polarization	21.77	14.89	0.000
$\mathbf{P}$ from Kondo et al. [29]	18.26	15.44	0.203
$\mathbf{P}$ from DeepSfP [4]	18.05	14.82	1.514
$\mathbf{P}$ from our approach	**17.86**	**14.68**	0.281

Table 4. **Controlled experiments for polarization representations on SPW and DeepSfP [4] datasets.** Please check Sec. 4.2 for the details of various representations. We test the preprocessing time of a raw image with resolution 1024×1224 using a single thread on Intel Xeon Gold CPU with 2.30GHz frequency.

5.4.2 Viewing encoding

Since the DeepSfP dataset is not designed for scene-level SfP, we only conduct the controlled experiments on the

Viewing encoding	Angular Error ↓			Accuracy ↑		
	Mean	Median	RMSE	11.25°	22.5°	30.0°
Ours without $\mathbf{V}$	22.12	18.00	27.03	32.2	66.9	77.8
Ours with $\mathbf{V}_p$	20.31	16.02	25.68	40.4	71.0	80.5
Ours with $\mathbf{V}_c$	18.44	14.62	23.46	43.7	76.1	84.8
Ours with $\mathbf{V}$	**17.86**	**14.20**	**22.72**	**44.6**	**76.3**	**85.2**

Table 5. **Controlled experiments for the viewing encoding on the SPW dataset.** Previous learning-based SfP methods [4,29] do not input any viewing encoding. In addition to our viewing encoding $\mathbf{V}$, we also try to use the positional encoding of NeRF [36] $\mathbf{V}_p$ and normalized coordinates $\mathbf{V}_c$ as the viewing encoding.

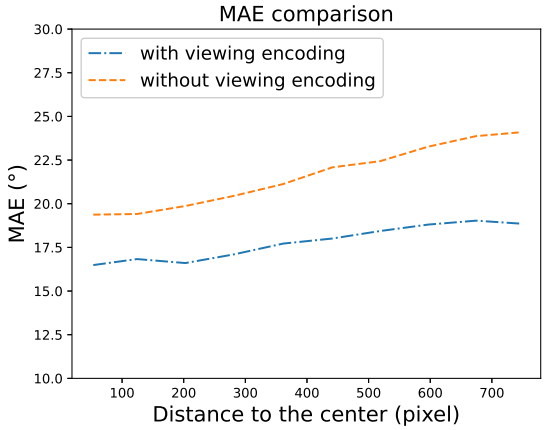


Figure 8. **An analysis of the viewing encoding.** We calculate the mean angular error (MAE) for each pixel on the test set. We notice that the improvement brought by viewing encoding increases with the distance to the image center. We believe this is because the impact of non-orthographic projection is more severe in the corners of images.

SPW dataset. For each model, we remove the viewing encoding from the input or use different types of viewing encoding. In Table 5, using viewing encoding improves our model in all the metrics effectively by a large margin. The model that uses raw viewing directions achieves the best performance. The model that uses positional encoding of NeRF [36] also improves the performance but is not as good as ours. When the images in a dataset are captured with the same intrinsic parameters, using normalized coordinates as viewing encoding also obtains satisfying performance. We further analyze the impact of viewing encoding in Fig. 5 and Fig. 8.

5.4.3 Network architectures

We study different network architectures in this section. In addition to networks of previous SfP methods (DeepSfP [4] and Kondo et al. [29]), we also compare with two RGB-based normal estimation methods: TransDepth [21] and DORN [20]. For all the experiments, we provide the same polarization representation and viewing encoding to these compared network architectures. Table 6 presents the quantitative results of different architectures. Our architecture obtains the best performance. Besides, removing self-attention or replacing instance normalization with batch normalization leads to performance degradation.

Network	Angular Error ↓			Accuracy ↑		
	Mean	Median	RMSE	11.25°	22.5°	30.0°
Kondo et al.† [29]	26.43	22.69	31.80	23.8	54.1	67.6
DeepSfP† [4]	24.97	20.83	30.13	25.6	58.4	70.9
U-Net [39]	26.35	22.45	31.97	25.4	54.5	67.6
DORN [20]	20.16	15.60	25.47	39.8	71.3	81.1
TransDepth [49]	22.05	17.46	27.77	33.0	66.6	77.9
Ours without IN	20.74	16.63	25.98	38.5	69.3	79.0
Ours without SA	21.08	16.54	26.62	36.1	68.5	79.3
Ours	**17.86**	**14.20**	**22.72**	**44.6**	**76.3**	**85.2**

Table 6. **Controlled experiments for network architectures on the SPW dataset.** We retrain Kondo et al. [29], DeepSfP [4] and other networks with the same representation as ours (e.g. viewing encoding and our novel polarization representation) for fair comparison. SA: self-attention. IN: instance normalization. †: our implementation.

6. Conclusion

We present the first approach dedicated to scene-level surface normal estimation from a single polarization image in the wild. The accuracy of our model is demonstrated on SPW, the first scene-level dataset for real-world SfP. By introducing the viewing encoding, a self-attention module and a novel polarization representation to SfP, our model substantially outperforms prior work on both SPW and the object-level DeepSfP dataset. In addition, our model can generalize from near-field scenes (used during training) to far-field outdoor scenes. This is possible because the polarization sensor is based on passive sensing, so our trained model is expected to generalize to distant scenes. We hope our work including the proposed SPW dataset and our technical designs can contribute to high-quality normal estimation, especially shape from polarization.

Limitations. One of the limitations of our work is the lack of quantitative evaluation in outdoor scenes. Note that the quantitative experiments in outdoor scenes will require long-range high-resolution depth and normal estimation with high-end depth sensors.

References

[1] Gary A. Atkinson. Polarisation photometric stereo. *Comput. Vis. Image Underst.*, 160:158–167, 2017. 2

[2] Gary A. Atkinson and Edwin R. Hancock. Recovery of surface orientation from diffuse polarization. *IEEE Trans. Image Process.*, 15(6):1653–1664, 2006. 2

[3] Gary A. Atkinson and Edwin R. Hancock. Shape estimation using polarization and shading from two views. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(11):2001–2017, 2007. 2, 5

[4] 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. 2, 3, 4, 5, 6, 7, 8

[5] Seung-Hwan Baek, Daniel S. Jeon, Xin Tong, and Min H. Kim. Simultaneous acquisition of polarimetric SVBRDF and normals. *ACM Trans. Graph.*, 37(6):268:1–268:15, 2018. 2

[6] Aayush Bansal, Xinlei Chen, Bryan C. Russell, Abhinav Gupta, and Deva Ramanan. Pixelnet: Towards a general pixel-level architecture. *CoRR*, abs/1609.06694, 2016. 3, 6

[7] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr Revisited: 2D-3D model alignment via surface normal prediction. In *CVPR*, 2016. 3

[8] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(8):1670–1687, 2015. 3

[9] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K. Wong. Deep photometric stereo for non-lambertian surfaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–1, 2020. 1

[10] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K Wong. Self-calibrating deep photometric stereo networks. In *CVPR*, 2019. 1

[11] Weifeng Chen, Donglai Xiang, and Jia Deng. Surface normals in the wild. In *ICCV*, 2017. 3

[12] Edward Collett. Field guide to polarization. Spie Bellingham, WA, 2005. 1, 2, 4

[13] Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, and Jan Kautz. Polarimetric multi-view stereo. In *CVPR*, 2017. 2

[14] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 3

[15] Valentin Deschaintre, Yiming Lin, and Abhijeet Ghosh. Deep polarization imaging for 3d shape and svbrdf acquisition. In *CVPR*, 2021. 2

[16] Yuqi Ding, Yu Ji, Mingyuan Zhou, Sing Bing Kang, and Jinwei Ye. Polarimetric helmholtz stereopsis. In *CVPR*, 2021. 2

[17] Tien Do, Khiem Vuong, Stergios I. Roumeliotis, and Hyun Soo Park. Surface normal estimation of tilted images via spatial rectifier. In *ECCV*, 2020. 3

[18] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015. 3

[19] Rui Fan, Hengli Wang, Peide Cai, and Ming Liu. Sne-roadseg: Incorporating surface normal information into semantic segmentation for accurate freespace detection. In *ECCV*. Springer, 2020. 1

[20] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 8

[21] Yoshiki Fukao, Ryo Kawahara, Shohei Nobuhara, and Ko Nishino. Polarimetric normal stereo. In *CVPR*, 2021. 2, 8

[22] Jingwei Huang, Yichao Zhou, Thomas Funkhouser, and Leonidas J. Guibas. Framenet: Learning local canonical frames of 3d surfaces from a single rgb image. In *ICCV*, 2019. 1, 3

[23] Tomoki Ichikawa, Matthew Purri, Ryo Kawahara, Shohei Nobuhara, Kristin Dana, and Ko Nishino. Shape from sky: Polarimetric normal recovery under the sky. In *CVPR*, 2021. 2

[24] Chen Jieneng, Lu Yongyi, Yu Qihang, Luo Xiangde, Adeli Ehsan, Wang Yan, Lu Le, Yuille Alan L., and Zhou Yuyin. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 6

[25] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Polarized 3d: High-quality depth sensing with polarization cues. In *ICCV*, 2015. 2

[26] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 2006. 1

[27] Kichang Kim, Akihiko Torii, and Masatoshi Okutomi. Multi-view inverse rendering under arbitrary illumination and albedo. In *ECCV*, 2016. 3

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[29] Yuhi Kondo, Taishi Ono, Legong Sun, Yasutaka Hirasawa, and Jun Murayama. Accurate polarimetric BRDF for real polarization scene rendering. In *ECCV*, 2020. 2, 3, 4, 5, 6, 7, 8

[30] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *CVPR*, 2020. 4

[31] Bo Li, Chunhua Shen, Yuchao Dai, A. van den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *CVPR*, 2015. 3

[32] L. Li, Z. Li, K. Li, L. Blarel, and M. Wendisch. A method to calculate stokes parameters and angle of polarization of skylight from polarized cimel sun/sky radiometers. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 149:334–346, 2014. 4

[33] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image. In *CVPR*, 2020. 1, 3

[34] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An

intriguing failing of convolutional neural networks and the coordconv solution. In *NeurIPS*, 2018. 5

[35] Ali H. Mahmoud, Moumen T. El-Melegy, and Aly A. Farag. Direct method for shape recovery from polarization and shading. In *ICIP*, 2012. 2, 6, 7

[36] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 5, 8

[37] Daisuke Miyazaki, Takuya Shigetomi, Masashi Baba, Ryo Furukawa, Shinsaku Hiura, and Naoki Asada. Surface normal estimation of black specular objects from multiview polarization images. *Optical Engineering*, 56(4):041303, 2016. 2

[38] Daisuke Miyazaki, Robby T. Tan, Kenji Hara, and Katsushi Ikeuchi. Polarization-based inverse rendering from a single view. In *ICCV*, 2003. 2, 6, 7

[39] Ronneberger Olaf, Fischer Philipp, and Brox Thomas. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 6, 8

[40] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *CVPR*, 2018. 3

[41] Xiaojuan Qi, Zhengzhe Liu, Renjie Liao, Philip H. S. Torr, Raquel Urtasun, and Jiaya Jia. Geonet++: Iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020. 3

[42] Stefan Rahmann and Nikos Canterakis. Reconstruction of specular surfaces using polarization imaging. In *CVPR*, 2001. 2

[43] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012. 3

[44] 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. 2, 6, 7

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 5, 6

[46] Peng Wang, Xiaohui Shen, Bryan C. Russell, Scott Cohen, Brian L. Price, and Alan L. Yuille. SURGE: surface regularized geometry estimation from a single image. In *NeurIPS*, 2016. 3

[47] Rui Wang, David Geraghty, Kevin Matzen, Richard Szeliski, and Jan-Michael Frahm. Vplnet: Deep single view normal estimation with vanishing points and lines. In *CVPR*, 2020. 3

[48] Xiaolong Wang, David F. Fouhey, and Abhinav Gupta. Designing deep networks for surface normal estimation. In *CVPR*, 2015. 3, 6

[49] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *ICCV*, 2021. 6, 8

[50] Ye Yu and William A. P. Smith. Inverserendernet: Learning single image inverse rendering. In *CVPR*, 2019. 3

[51] Amir Roshan Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J. Guibas. Robust learning through cross-task consistency. In *CVPR*, 2020. 3

[52] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. *CVPR*, 2017. 3

[53] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*, 2019. 3

[54] Jinyu Zhao, Yusuke Monno, and Masatoshi Okutomi. Polarimetric multi-view inverse rendering. In *ECCV*, 2020. 3

[55] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *CoRR*, abs/1801.09847, 2018. 4

[56] Dizhong Zhu and William A. P. Smith. Depth from a polarisation + RGB stereo pair. In *CVPR*, 2019. 2, 4