Learning Accurate 3D Shape Based on Stereo Polarimetric Imaging

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

Abstract

Shape from Polarization (SfP) aims to recover surface normal using the polarization cues of light. The accuracy of existing SfP methods is affected by two main problems. First, the ambiguity of polarization cues partially results in false normal estimation. Second, the widely-used assumption about orthographic projection is too ideal. To solve these problems, we propose the first approach that combines deep learning and stereo polarization information to recover not only normal but also disparity. Specifically, for the ambiguity problem, we design a Shape Consistency-based Mask Prediction (SCMP) module. It exploits the inherent consistency between normal and disparity to identify the areas with false normal estimation. We replace the unreliable features enclosed by these areas with new features extracted by global attention mechanism. As to the orthographic projection problem, we propose a novel Viewing Direction-aided Positional Encoding (VDPE) strategy. This strategy is based on the unique pixel-viewing direction encoding, and thus enables our neural network to handle the non-orthographic projection. In addition, we establish a real-world stereo SfP dataset that contains various object categories and illumination conditions. Experiments showed that compared with existing SfP methods, our approach is more accurate. Moreover, our approach shows higher robustness to light variation.

1. Introduction

3D shape recovery is a fundamental problem in computer vision and has been extensively studied [15, 26, 31]. However, existing shape recovery methods have some limitations. For example, the geometry-based methods, e.g., structure from motion [33, 37] have difficulty in dealing with texture-less regions and can only recover sparse point cloud. While the photometric stereo methods [15, 18] can recover dense surface, they need cumbersome photometric calibration. By contrast, shape from polarization (SfP) [11, 20, 40] can avoid the above problems by using the polarization cues of light. Specifically, polarization cues can detect rich geometric details even for white wall [11]. Moreover, such cues can be easily obtained in a single shot with the quad-Bayer polarization camera [46].

Despite the above advantages of SfP, there remain two main problems that affect the recovery accuracy. First, the *ambiguous polarization cues* are inevitable due to unidirectional measurement [10]. These cues partially result in false normal estimation. To solve this problem, early SfP methods rely on specific assumptions about shape prior [2, 40] and lead to unsatisfactory recovery accuracy. Recently, some approaches use stereo [16] or multi-view [49] polarization information for disambiguation. However, the accuracy of these methods is limited by the low quality of stereo matching for polarization cues. Second, most existing SfP approaches assume *orthographic projection* for modelling simplification [35, 40]. Such assumption ignores the influence of viewing directions, which affects the accuracy of polarimetric measurements. A representative work for this problem is based on a perspective phase angle constraint [8], but this constraint is still insufficient.

In addition to the above attempts to solve the ambiguity and orthographic projection problems, some methods are proposed based on deep learning [3, 12, 23, 25]. They improve the shape recovery accuracy to some extent. However, they still partly suffer from the ambiguity problem due to monocular imaging. By contrast, our method is based on *stereo* polarimetric imaging. To the best of our knowledge, our approach is the first one that combines stereo polarization information and deep learning to estimate both normal and disparity.

As shown in Fig. 1, we integrate convolutional neural network (CNN) with Vision Transformer [13, 14] to design the feature extraction module. This module considers both local and global contexts [34] to extract stereo feature maps. For one thing, we use the feature map of the left view to estimate the *normal map*. For another thing, we exploit the stereo feature maps to generate a polarimetric cost volume. This cost volume aligns stereo features to estimate the *disparity map*. Our joint estimation of normal and disparity

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

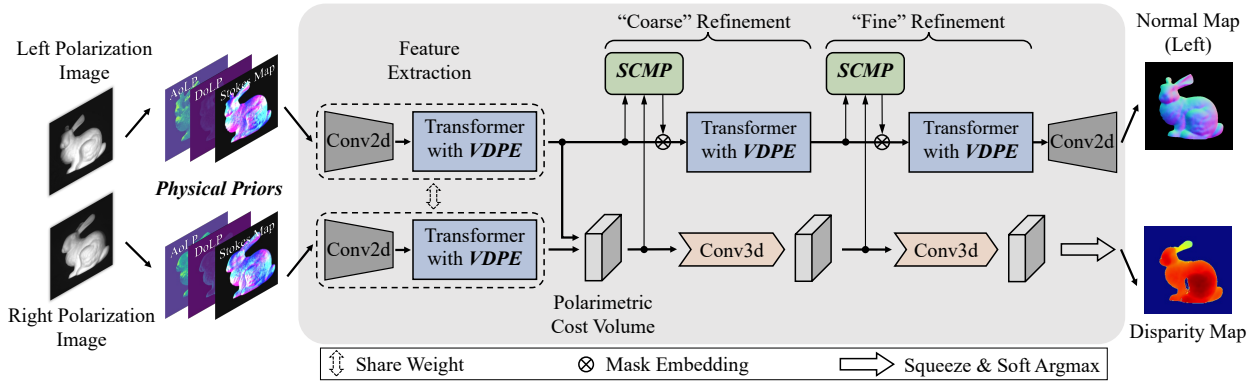


Figure 1. Overview of the proposed approach. Given a pair of stereo polarization images, our method can simultaneously recover normal map and disparity map with high quality.

contributes to solving the above-mentioned problems in SfP, which is introduced in the following.

For the ambiguity problem, we design a Shape Consistency-based Mask Prediction (SCMP) module. This module predicts a mask to identify the areas with inaccurate normal estimation caused by unreliable features in the feature map. We use these areas to achieve a coarse-to-fine refinement for the feature map. At each step, we replace the features enclosed by such areas with new features extracted by the global attention mechanism in Transformer. As to the orthographic projection problem, we introduce a novel Viewing Direction-aided Positional Encoding (VDPE) strategy to Transformer. Based on the pixel-viewing direction encoding, this strategy enables our neural network to handle non-orthographic projection. Moreover, we establish a large real-world stereo SfP dataset.

To summarize, we propose the first approach [1] that combines stereo polarimetric imaging and deep learning to recover accurate normal and disparity maps simultaneously. Our main contributions are as follows:

- We design a mask prediction module to reduce the effect of ambiguous polarization cues based on the consistency between normal and disparity.
- We propose a novel positional encoding design that enables our network to handle the non-orthographic projection in polarimetric measurement.
- We establish a large real-world dataset for stereo SfP problem. Our dataset contains various object categories and illumination conditions.

Extensive experiments showed that compared with existing methods, our approach is more accurate. Moreover, our approach shows higher robustness to light variation.

2. Related Work

We first review SfP methods that typically use pure polarization cues with monocular setup. We then introduce the approaches combining polarization and the other cues. Finally, we present works adopting stereo or multi-view setup.

Pure Polarization. Early SfP methods divide the polarization reflection into two cases, i.e., specular reflection [35] and diffuse reflection [2, 19, 29]. These methods model such two cases individually according to object materials. This strategy is impractical since light reflection involves both cases in the real world. Moreover, their assumptions on shape priors lead to over-smooth shape recovery results. Baek et al. [4] introduced a model that can consider two reflection cases simultaneously. However, their model requires the assumption about ideal surface reflection and thus is relatively unreliable. Considering the complexity of light reflection, some methods [3, 12, 23, 25] based on deep learning are proposed. Although these data-driven methods improve the recovery accuracy, they lead to low generalization. The reason is that their networks are trained on relatively small or synthetic datasets.

Polarization + X. The above pure polarization-based methods suffer from the ambiguity problem of polarization cues. To solve this problem, recent works attempt to introduce additional shape-from-X cues (X represents shading, photometry, depth, etc.). With the *cues of shading*, Smith et al. [39, 40] proposed to generate a large linear system to directly calculate the height of each pixel. However, their assumptions about the known refractive index and orthographic camera projection limit the recovery accuracy. Ngo et al. [32] combined the polarization with shading cues to additionally estimate the refractive index. Nevertheless, their method needs at least two light directions and is time-consuming. In addition, given the *cues of photometry*, Atkinson et al. [1] used two calibrated photometric systems to reduce the effects of polarization ambiguity. Tozza et al. [41] improved their method by solving the uncalibrated case. Moreover, based on the *cues of depth*, Kadambi et al. [21] used the geometric constraints of normals to refine the depth map. This method relies on redundant depth alignment, which leads to low efficiency.

Stereo and Multi-view Polarization. Stereo or multi-view

[1] https://tyhuang98.github.io/learn_stereo_sfp/

imaging setup contributes to the disambiguation of polarization cues. For example, Cui et al. [11] proposed an iso-depth contour tracing mechanism in their multi-view setup to fuse the estimated depth and normal maps. Chen et al. [9] proposed a theoretical polarimetric transport model to constrain the recovery solutions from three views. Zhu and Smith [50] adopted a pair of polarization and RGB images. They introduced a high-order graphical model to label the diffuse-dominant and specular-dominant pixels separately, followed by shape recovery. Recently, Fukao et al. [16] proposed an approach to recover per-pixel normals using stereo polarimetric images. They treated the Stokes vector as polarization cues and constructed a cost volume to regress normals. Compared with the above methods, our approach is more robust to noise thanks to our proposed modules, as will be shown in the experiments.

3. Proposed Method

In Section 3.1, we introduce the basic knowledge of polarization. In Section 3.2, we give an overview of our pipeline. In Section 3.3, we introduce the SCMP module that enables our method to solve the ambiguity problem. In Section 3.4, we introduce our VDPE design that aims at handling the non-orthographic projection. In Section 3.5, we introduce our loss function.

3.1. Basic Knowledge of Polarization

Let us consider a partially polarized light that orthographically passes through a linear polarizer with an angle of ϕ_{pol}. The measured intensity of this light follows a sinusoidal variation [10] as

$$\begin{aligned} I_{\phi_{\text{pol}}} &= \frac{I_{\text{max}} + I_{\text{min}}}{2} + \frac{I_{\text{max}} - I_{\text{min}}}{2} \cos(2\phi_{\text{pol}} - 2\phi) \\ &= \bar{I} + \rho \bar{I} \cos(2\phi_{\text{pol}} - 2\phi), \end{aligned} \tag{1}$$

where I_{max} and I_{min} denote the upper and lower bounds of the detected light intensity, ϕ denotes the angle of linear polarization (AoLP), ρ denotes the degree of linear polarization (DoLP), $\bar{I}$ denotes the average intensity. Based on Eq. (1), two polarization angles ϕ and ϕ' with a π-shift (ϕ' = $\phi \pm \pi$) result in the same observed light intensity, which is called π-ambiguity problem. With the linear quad-Bayer polarization camera [46], we can measure intensities in different angles of polarizer. By combining these intensities, we can solve the above unknown polarization parameters: ϕ with π-ambiguity, ρ, and $\bar{I}$. These parameters have been proved to be highly related to the surface normal [2, 35]. Details are available in the supplementary material.

In addition, there is another representation to describe the polarization state of light [10], i.e., the Stokes vector:

$$\mathbf{s} = \begin{bmatrix} s_0 \\ s_1 \\ s_2 \\ s_3 \end{bmatrix} = \begin{bmatrix} I_0 + I_{90} \\ I_0 - I_{90} \\ I_{45} - I_{135} \\ 0 \end{bmatrix}, \tag{2}$$

where s_0 represents the light intensity; s_1 and s_2 represent the linear polarization components in 0° and 45°, respectively; s_3 represents the right circular polarization component. The Stokes vector $\mathbf{s}$ in Eq. (2) can be expressed by the measured intensities I_0, I_{45}, I_{90} and I_{135}. In our context, we follow [12] to call the normalized Stokes vector image "Stokes map".

3.2. Pipeline Overview

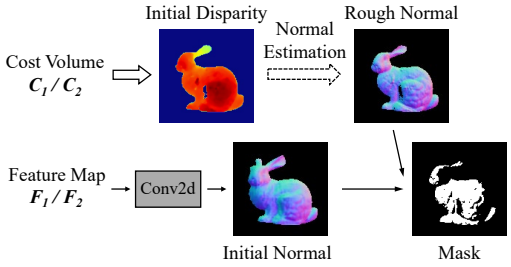
As shown in Fig. 1, given the measured stereo polarization images, we compute AoLP&DoLP and Stokes maps, and treat them as the physical priors. By taking these priors as inputs, our network uses a weight-shared feature extraction module to extract stereo feature maps. Our feature extraction is based on the combination of CNN and Transformer to consider both local and global contexts. Note that our Transformer is adapted by the proposed VDPE design (see Section 3.4) to handle the non-orthographic projection. Based on the extracted stereo feature maps, our network predicts both normal and disparity maps.

Disparity Map. Given the stereo feature maps, we generate a polarimetric cost volume. This volume aligns stereo features and encapsulates the disparity information [22]. We process such volume by a series of Conv3d blocks [42] and the Soft ArgMin operation [22] to infer the disparity map.

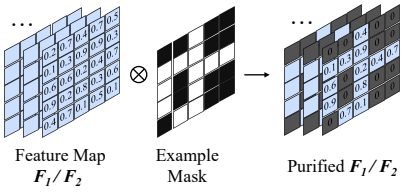
Normal Map. The extracted feature map of the left view is associated with the unknown-but-sought normal map. However, due to the ambiguity problem of polarization cues, this feature map needs to be refined. To achieve this goal, we propose a coarse-to-fine feature refinement strategy. This refinement is based on the proposed SCMP module (see Section 3.3) and Transformer. Our SCMP module incorporates the above cost volume since there exists inherent consistency between disparity and normal. After refinement, the feature map is used to infer the normal map.

3.3. Shape Consistency-based Mask Prediction

Leveraging the shape consistency between normal and disparity can improve the shape recovery accuracy. Let us first intuitively explain the reason. For one thing, the normal map estimated by polarization cues are per-pixel. However, the normals in some areas are false due to the ambiguity problem, and we cannot obtain a mask to exclude such areas a priori. For another thing, given a disparity map, we can compute a normal map that is globally reliable but coarse. The correct normals labelled by the unknown-but-sought mask are relatively consistent with the normals computed by the disparity map. We leverage such consistency



(a) Proposed SCMP module.



(b) The mask embedding operation in our network.

Figure 2. Illustration of proposed SCMP module and mask embedding operation. (a) Based on shape consistency, the proposed SCMP module predicts a mask. (b) We employ the predicted mask to purify the feature map F_1/F_2 by a mask embedding operation.

to identify this mask by introducing the SCMP module. We will introduce our module details in the following.

As shown in Fig. 1, we embed two SCMP modules into our network. Fig. 2(a) shows how our first and second SCMP modules are used in our coarse-to-fine feature refinement. Specifically, at the "coarse" stage, we take the original feature map F_1 of the left view and the polarimetric cost volume C_1 as inputs of the first SCMP module. At the "fine" stage, we feed the intermediate feature map F_2 and cost volume C_2 to the second SCMP module. At each stage, an initial normal map can be inferred from the feature map F_1 or F_2 under supervision. Meanwhile, an initial disparity map can be regressed from the cost volume C_1 or C_2 under supervision. Based on this disparity map, we compute a rough normal map using an adaptive normal estimation method [27]. We then exploit the consistency between the normal inferred by our network and the normal computed based on disparity to predict a mask. Specifically, the value v_i of the i-th pixel in this mask is computed by

$$v_i = \begin{cases} 0, & \text{if } \mathbf{n}_i^N \cdot \mathbf{n}_i^D < \tau \\ 1, & \text{otherwise} \end{cases}, \tag{3}$$

where "$\cdot$" denotes the dot product, $\mathbf{n}_i^N$ and $\mathbf{n}_i^D$ are the normals inferred by our network and computed based on disparity, respectively, τ is a threshold. When $\mathbf{n}_i^N \cdot \mathbf{n}_i^D$ is smaller than the threshold τ, the consistency is violated. In our paper, τ is 0.7 and 0.9 in the first and second SCMP modules, respectively. We adopt such threshold setup to achieve coarse-to-fine feature refinement.

We then employ the above mask to purify the feature map F_1 or F_2 by a mask embedding operation (see Fig. 2(b)). Each unreliable feature in the areas with false normal estimation is set to zero vector. Then the purified feature map is fed to Transformer for global attention. Thanks to the mask embedding, the effect of the unreliable features resulting in false normal estimation is reduced. In addition, the global attention mechanism in Transformer extracts new features for the masked areas by combing global reliable features. Based on such features, the false normals can be re-estimated in an accurate way.

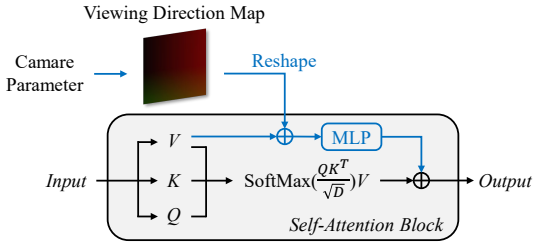


Figure 3. Illustration of proposed VDPE design. For ease of understanding, we only show the self-attention part in the Transformer.

3.4. Viewing Direction-aided Positional Encoding

Most existing SfP works [35, 39] assume orthographic projection (i.e., viewing directions of all pixels are $[0, 0, 1]^\top$). This assumption affects the accuracy of polarimetric measurement and further the accuracy of shape recovery. Intuitively, additionally using a viewing direction map can alleviate this problem. We generate this map based on the pixel-viewing direction encoding, i.e., each pixel in this map is associated with a unique viewing direction. To introduce such map into neural network, a straightforward strategy is to feed it to CNN [25]. However, it is well-known that the feature extracted by CNN lacks global information. By contrast, we propose to integrate this map into Transformer. This integration is based on the proposed VDPE design (see the next paragraph). As will be shown in the experiments, such design is more effective than the strategy based on CNN.

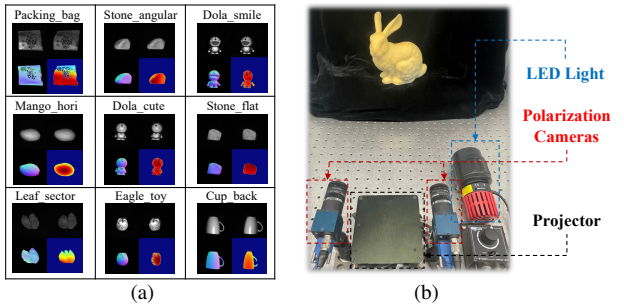
As shown in Fig. 3, we integrate the viewing direction map into the self-attention part of Transformer. Specifically, we reshape the viewing direction map and concatenate it with the linearly projected $value$ in the self-attention block. Then the concatenated volume is passed to a Multi-Layer Perception (MLP) layer to embed with the attention calculation. After that, the self-attention computation result can be formulated as:

$$\text{Output} = (\text{SoftMax}(\frac{QK^\top}{\sqrt{D}})V) \oplus (\text{MLP}(V \oplus \boldsymbol{v})), \tag{4}$$

where V, K, Q represent the linearly projected $value$, key and $query$ respectively, $\boldsymbol{v}$ is the viewing direction map, D is the feature dimension, "$\oplus$" represents the concatenation operation. Based on Eq. (4), our method can handle the non-orthographic projection reliably.

Table 1. Comparison among existing SfP datasets. Our dataset is the first stereo SfP dataset at the object level. Meanwhile, our acquisition is under controlled illumination and does not need reprojection.

Dataset	Level	Type	Setup	Controlled Illumination	Reprojection Operation
DeepSfP [3]	Object	Real-world	Monocular	No	Yes
Kondo et al. [23]	Scene	Synthetic	Monocular	Yes	No
Deschaintre et al. [12]	Object	Synthetic	Monocular	Yes	No
SPW [25]	Scene	Real-world	Monocular	No	Yes
CroMo [44]	Scene	Real-world	Stereo	No	Yes
Ours	Object	Real-world	Stereo	Yes	No



(a) (b)

Figure 4. We establish a real-world stereo SfP dataset, which contains rich shapes and material categories. (a) Representative image pairs. (b) Data capture setup.

3.5. Loss Function

For network training, we use the normal loss and the disparity loss. For the normal loss, we adopt the commonly used cosine similarity loss [6, 7]. Our normal loss consists of three sub-losses $\mathcal{N}_1$, $\mathcal{N}_2$, and $\mathcal{N}_{\text{final}}$. $\mathcal{N}_1$ and $\mathcal{N}_2$ correspond to the initial normal maps in the first and second SCMP modules, respectively. $\mathcal{N}_{\text{final}}$ is related to the final predicted normal map of our network. As to the disparity loss, we adopt the widely-used smooth L_1 loss [5, 17]. Our disparity loss consists of three sub-losses $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_{\text{final}}$. $\mathcal{D}_1$ and $\mathcal{D}_2$ correspond to the initial disparity maps in the first and second SCMP modules, respectively. $\mathcal{D}_{\text{final}}$ is related to the final predicted disparity map of our network.

By combing the above sub-losses, we define our final loss $\mathcal{L}$ as:

$$\begin{aligned}\mathcal{L} &= \alpha \cdot (\mathcal{N}_1 + \mathcal{N}_2) + \beta \cdot (\mathcal{D}_1 + \mathcal{D}_2) \\ &+ \alpha' \cdot \mathcal{N}_{\text{final}} + \beta' \cdot \mathcal{D}_{\text{final}},\end{aligned} \tag{5}$$

where α, β, α', and β' are the trade-off parameters. During training, we set α and β as 0.1 for weak supervision since they correspond to the intermediate results. We set α' and β' as 0.3 for major supervision since they correspond to our final predictions. Further implementation details of our method are available in the supplementary material.

4. Dataset

Most existing SfP datasets are either small [3, 25] or only provide synthetic images [12, 23]. In addition, their images are obtained by monocular cameras and cannot be used for research on stereo polarimetric imaging. While a recent dataset contains abundant stereo polarization images, it only provides the ground truth depth and ignores the normal [44]. To solve this problem, we propose a large real-world SfP dataset with stereo images associated with the ground truth normal and disparity. Considering that arbitrary illumination has been proved to limit the quality of polarization signal [12], we use the controlled illumination (i.e., unpolarized illumination) for data collection. Moreover, the ground truth shapes of existing real-world SfP datasets [3, 25] are obtained by an extra 3D scanner, which introduces the reprojetion operation. By contrast, we propose a more concise acquisition design that can avoid this operation. Table 1 shows the comparison among existing SfP datasets.

Data Composition. Our dataset consists of 2450 pairs of polarization images. We collect data from more than 50 different objects with various shapes and material categories. For data diversity, we collect images from different viewpoints and different illumination conditions for most objects. Each image pair is associated with the ground truth disparity and normal. Moreover, our dataset involves different stereo baselines between stereo images. Fig. 4(a) shows representative image pairs.

System Setup for Image Collection. When capturing stereo images, we fix a pair of polarization cameras (see Fig. 4(b)) and calibrate them using a standard calibration method [48]. To achieve different illumination conditions, we use an unpolarized light source to illuminate each object from a set of random directions. Moreover, to improve the diversity of illuminations, we use two unpolarized light sources, i.e., the digital projector and a LED lighting.

Acquisition of Ground Truth Maps. To acquire the ground-truth normal and disparity maps, we use a stereo structured-light system. We replace the original grayscale cameras of this system with our stereo polarization camera rig (the grayscale information can be extracted from the polarization information). Based on this adapted system, we employ a Gray code-based reconstruction algorithm [30] to reconstruct the disparity map and point cloud of the object. Finally, we use the least-squares normal estimation algorithm [36] to calculate the normal map of the reconstructed point cloud. Each step of our acquisition does not need reprojection operation mentioned above. Our design ensures perfect alignment of the input polarization image and the ground truth normal and disparity.

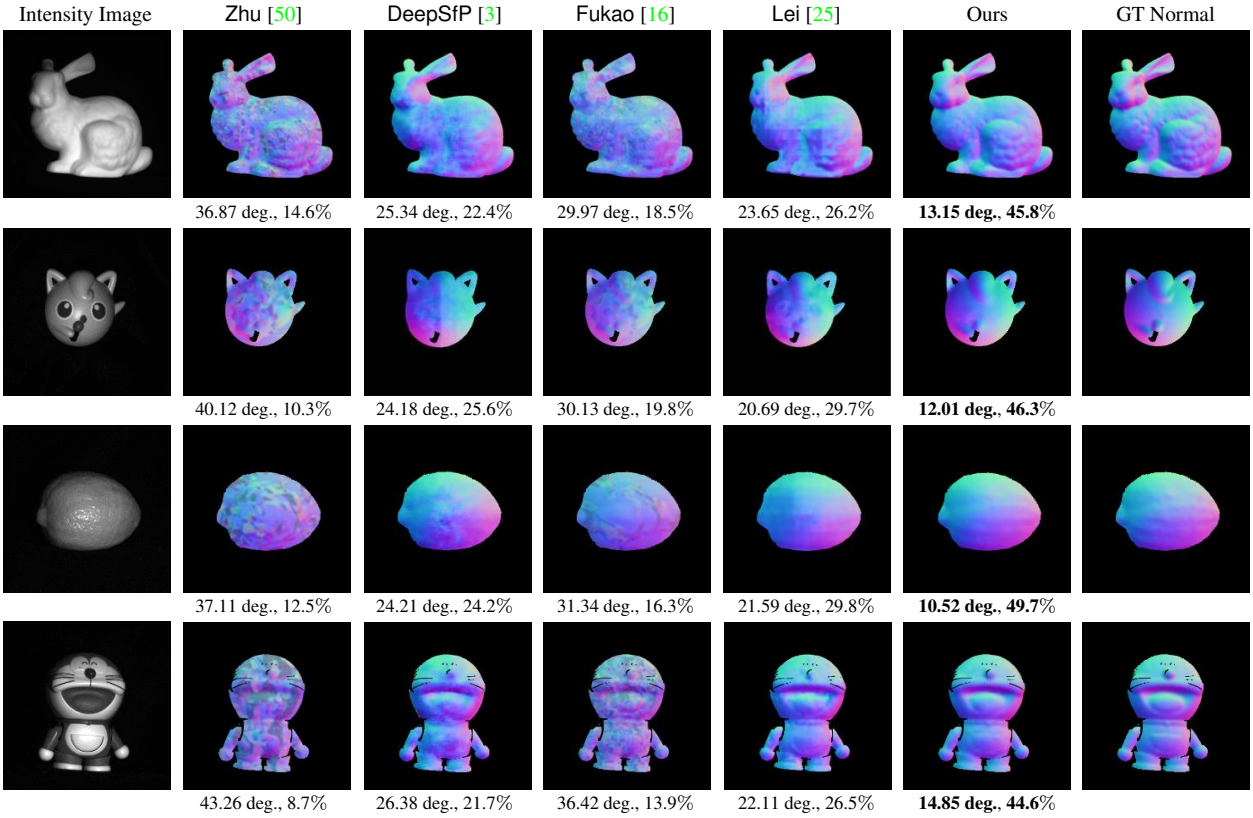


Figure 5. Representative accuracy comparisons between various SfP methods for normal recovery. A pair of numbers below each estimated normal map represents the mean of angular errors and the percentage of pixels with angular errors smaller than 11.25°.

5. Experiments

We first compare our approach with state-of-the-art SfP methods that aim at normal recovery in Section 5.1. Considering that our approach can predict both normal and disparity, we also compare our method with two baselines that jointly estimate normal and disparity in Section 5.2. In addition, we conduct ablation studies in Section 5.3. All the experiments are conducted on our dataset introduced above. Additional experimental results are available in the supplementary material.

In terms of evaluation criteria, we follow [25] to evaluate the normal accuracy by the angular error between the estimated and ground truth normals. In addition, we follow [45] to report the percentage of pixels with angular errors smaller than specific thresholds (i.e., 11.25°, 22.5°, and 30.0°). To evaluate the accuracy of disparity map, we follow [47] to use the absolute errors between the estimated and ground truth disparities.

5.1. Comparison to SfP Methods

Methods for Comparison. We compare our approach with four state-of-the-art SfP methods (i.e., Zhu [50], DeepSfP [3], Fukao [16] and Lei [25]) in terms of normal recovery accuracy and robustness to light variation. As introduced in Section 2, Zhu and Fukao use stereo setup and numerical optimization strategies. DeepSfP and Lei are based on monocular setup and deep learning.

Table 2. Accuracy comparisons between various SfP methods for normal recovery on all the data of our testing set.

Method	Angular Error (deg.)			Pixel Percentage (%)		
	Mean	Median	RMSE	11.25°	22.5°	30.0°
Zhu [50]	37.69	34.27	43.01	12.4	31.9	45.7
DeepSfP [3]	24.71	21.43	30.18	23.6	47.4	67.8
Fukao [16]	31.28	27.52	37.15	17.4	39.5	54.3
Lei [25]	22.27	19.32	28.67	28.8	55.9	72.1
Ours	**13.22**	**11.14**	**17.22**	**46.2**	**77.5**	**90.1**

Normal Recovery Accuracy. As shown in Table 2 and Fig. 5, Zhu and Fukao are affected by the stereo matching quality and thus recover noisy normal maps. DeepSfP shows mistakes in many local areas. While Lei partly alleviates the ambiguity problem, it still loses several details. By contrast, our method achieves the highest accuracy on all the evaluation metrics and also leads to rich object details. This result demonstrates that our integration of deep learning and stereo polarization is effective.

Robustness to Light Variation. To evaluate the robustness to light variation, we use different illumination conditions

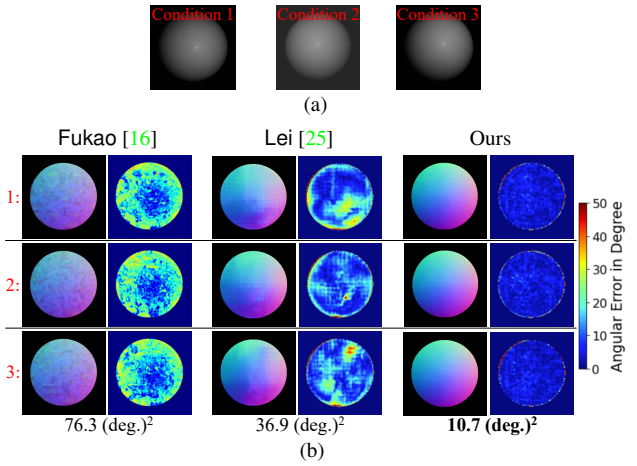


Figure 6. Representative comparisons regarding robustness to light variation between various SfP methods. (a) Three different illumination conditions. (b) Due to limited space, we only report the results of Fukao, Lei, and our method. Each method corresponds to two columns. The first column shows the 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.

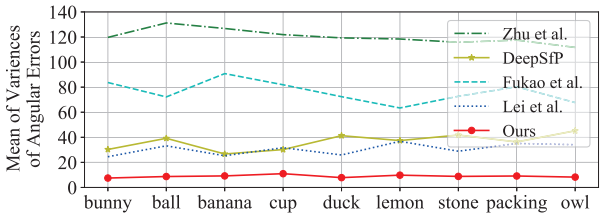


Figure 7. Comparisons regarding robustness to light variation between various SfP methods on our testing set. The horizontal axis represents different objects.

while keeping the viewing direction unchanged. Accordingly, each pixel is associated with a set of angular errors. We compute the variance of the error set of each pixel, followed by obtaining the mean of these variances. Figs. 6 and 7 show that under different illumination conditions, existing SfP approaches lead to unstable recovery results (i.e., the variance is large) and are sensitive to light variation. By contrast, the recovery results of our method are stable and the accuracy remains high despite light variation. The reason is that our network leverages redundant stereo information, and also extracts both local and global contexts for stable estimation.

5.2. Joint Estimation of Disparity and Normal

Methods for Comparison. Recall that our method can recover both normal and disparity maps. To evaluate performance, we compare our method with two well-known baselines, i.e., Long [28] and Kusupati [24] for joint estimation of normal and depth (depth and disparity are basically equivalent). These methods originally take stereo

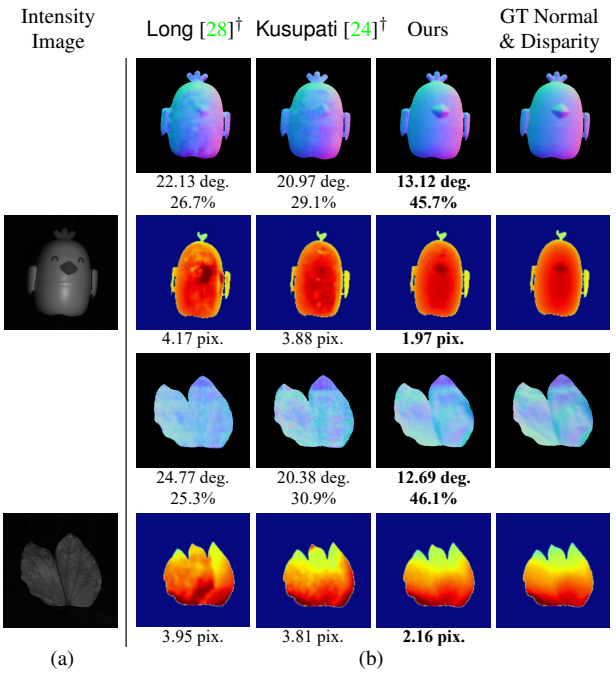


Figure 8. Representative comparisons between various methods for joint estimation of normal and disparity. (a) Intensity images of recovered objects. (b) The first and third rows show the normal maps. The second and fourth rows show the disparity maps. The meanings of numbers below each normal map are the same as those in Fig. 5. The number below each disparity map represents the mean of disparity errors. †: the same network inputs as ours.

Table 3. Accuracy comparisons between various methods for joint estimation of normal and disparity on all the data of our testing set. †: the same network inputs as ours.

Method	Angular Error (deg.)		Disparity Error (pix.)	
	Mean	Median	Mean	Median
Long [28]†	22.34	18.59	3.91	3.02
Kusupati [24]†	19.69	16.11	3.26	2.79
Ours	**13.22**	**11.14**	**2.11**	**1.98**

intensity images as inputs. For an unbiased comparison, we replace their inputs with the physical priors we use (see Section 3.2). Accordingly, we adjust and retrain their networks following their original setups.

Accuracy of Normal and Disparity. As shown in Table 3 and Fig. 8, Long and Kusupati lead to relatively unsatisfactory results. By contrast, our method can achieve the highest accuracy for both normal and disparity estimation. The reason is that our network architecture can leverage the polarization information, and also alleviate the ambiguity and orthographic projection problems in SfP.

5.3. Ablation Study

In this section, we conduct ablation studies regarding our network inputs, SCMP module, and VDPE design.

Network Inputs. Recall that we take AoLP&DoLP and

Table 4. Ablation study regarding different network inputs on all the data of our testing set.

Network Inputs	Angular Error (deg.)		Disparity Error (pix.)	
	Mean	Median	Mean	Median
Intensity images	33.25	30.61	7.23	6.15
Raw polarization	19.65	16.94	5.19	4.78
AoLP&DoLP	16.51	12.86	4.23	3.97
Stokes maps	17.13	15.25	3.69	3.12
Inputs as DeepSfP [3]	16.12	14.37	4.05	3.68
Original (AoLP&DoLP and Stokes maps)	**13.22**	**11.14**	**2.11**	**1.98**

Table 5. Ablation study regarding SCMP module on all the data of our testing set.

Network Design	Angular Error (deg.)		Disparity Error (pix.)	
	Mean	Median	Mean	Median
Without any SCMP	22.13	18.93	3.62	3.11
Without first SCMP	16.21	14.77	2.91	2.49
Without second SCMP	15.23	14.16	2.58	2.17
Original (with both SCMPs)	**13.22**	**11.14**	**2.11**	**1.98**

Stokes maps as the network inputs (see Section 3.2). We replace these inputs with 1) intensity images, 2) raw polarization images, 3) AoLP&DoLP, 4) Stokes maps, and 5) inputs as DeepSfP, respectively. Table 4 shows that our network with original inputs leads to the highest recovery accuracy on both normal and disparity. The reason is that AoLP&DoLP and Stokes maps complement each other and facilitate the usage of polarization cues for our network.

SCMP Module. Recall that we embed two SCMP modules in our network for coarse-to-fine feature refinement to solve the ambiguity problem (see Section 3.3). In this study, we compare our strategy with the simplified designs: 1) without any SCMP module, 2) without the first SCMP module, 3) without the second SCMP module. Table 5 shows that with only one SCMP module (regardless of the first or second one), the network can still achieve improvement in normal recovery accuracy, which proves the effectiveness of SCMP. In addition, the effect of the first module is more obvious. By contrast, our original design achieves the highest accuracy, demonstrating the effectiveness of our feature refinement strategy.

VDPE Design. Recall that we use the VDPE design in our network to handle the non-orthographic projection (see Section 3.4). In this study, we compare our design with the other encoding strategies: 1) classical position encoding methods based on Absolute Positional Embedding (APE) [43] or Relative Positional Embedding (RPE) [38], 2) viewing encoding (VE) of Lei [25]. Table 6 shows that compared with APE and RPE, both VE and our VDPE show improvement since they introduce the viewing directions and enable the networks to handle the non-orthographic projection. Moreover, the proposed VDPE design is more accurate than VE. Fig. 9 shows that RPE and VE both suffer from the non-orthographic projection problem and show higher errors on the area far from the image center. By contrast, our VDPE design avoids this problem and achieves better recovery results. The reason is that our design treats the viewing directions as part of the positional encoding in Transformer, which enhances the usage of viewing directions by global attention.

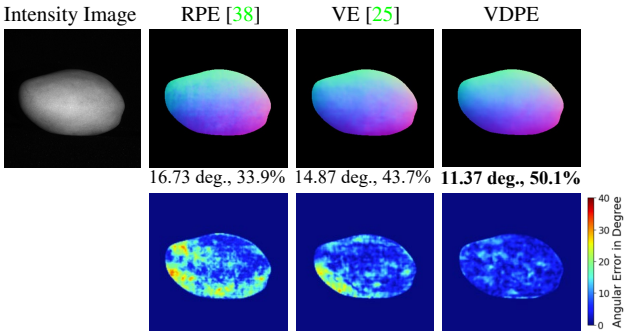


Figure 9. Representative ablation study regarding VDPE design. The meanings of numbers below each normal map are the same as those in Fig. 5. The second row shows the error maps corresponding to the normal maps.

Table 6. Ablation study regarding VDPE design on all the data of our testing set.

Encoding Strategy	Angular Error (deg.)		Disparity Error (pix.)	
	Mean	Median	Mean	Median
APE [43]	17.68	15.63	3.15	2.81
RPE [38]	16.43	14.79	3.11	2.56
VE [25]	15.89	14.17	2.93	2.42
Original (VDPE)	**13.22**	**11.14**	**2.11**	**1.98**

6. Conclusions

In this paper, we propose the first approach that combines stereo polarimetric imaging and deep learning to recover accurate normal and disparity. We design a mask prediction module to solve the ambiguity problem based on the shape consistency between normal and disparity. In addition, we propose a novel positional encoding method to solve the orthographic projection problem. Moreover, we establish a large real-world dataset for stereo SfP problem. Experiments on the proposed dataset demonstrate that our method outperforms existing SfP methods in terms of accuracy and robustness to light variation.

Acknowledgement. This work is supported in part by the CUHK T Sone Robotics Institute, and in part by the InnoHK of the Government of Hong Kong via the Hong Kong Centre for Logistics Robotics.

References

[1] Gary A Atkinson. Polarisation photometric stereo. *Computer Vision and Image Understanding*, 160:158–167, 2017. 2

[2] 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, 3

[3] 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, 2, 5, 6, 8

[4] 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. 2

[5] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *IEEE conference on computer vision and pattern recognition*, pages 5410–5418, 2018. 5

[6] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K Wong. Deep photometric stereo for non-lambertian surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):129–142, 2020. 5

[7] Guanying Chen, Kai Han, and Kwan-Yee K Wong. Ps-fcn: A flexible learning framework for photometric stereo. In *European conference on computer vision (ECCV)*, pages 3–18, 2018. 5

[8] Guangcheng Chen, Li He, Yisheng Guan, and Hong Zhang. Perspective phase angle model for polarimetric 3d reconstruction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 398–414. Springer, 2022. 1

[9] Lixiong Chen, Yinqiang Zheng, Art Subpa-Asa, and Imari Sato. Polarimetric three-view geometry. In *European Conference on Computer Vision (ECCV)*, pages 20–36, 2018. 3

[10] Edward Collett. Field guide to polarization. Spie Bellingham, WA, 2005. 1, 3

[11] 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, 3

[12] Valentin Deschaintre, Yiming Lin, and Abhijeet Ghosh. Deep polarization imaging for 3d shape and svbrdf acquisition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15567–15576, 2021. 1, 2, 3, 5

[13] 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. 1

[14] 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. 1

[15] Jean-Denis Durou, Maurizio Falcone, Yvain Quéau, and Silvia Tozza. *Advances in photometric 3d-reconstruction*. Springer, 2020. 1

[16] 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. 1, 3, 6, 7

[17] Ross Girshick. Fast r-cnn. In *IEEE international conference on computer vision*, pages 1440–1448, 2015. 5

[18] Aaron Hertzmann and Steven M Seitz. Example-based photometric stereo: Shape reconstruction with general, varying brdfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1254–1264, 2005. 1

[19] Cong Phuoc Huynh, Antonio Robles-Kelly, and Edwin Hancock. Shape and refractive index recovery from single-view polarisation images. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1229–1236, 2010. 2

[20] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Polarized 3d: High-quality depth sensing with polarization cues. In *IEEE International Conference on Computer Vision*, pages 3370–3378, 2015. 1

[21] Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Depth sensing using geometrically constrained polarization normals. *International Journal of Computer Vision*, 125(1):34–51, 2017. 2

[22] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *IEEE international conference on computer vision*, pages 66–75, 2017. 3

[23] Yuhi Kondo, Taishi Ono, Legong Sun, Yasutaka Hirasawa, and Jun Murayama. Accurate polarimetric brdf for real polarization scene rendering. In *European Conference on Computer Vision*, pages 220–236. Springer, 2020. 1, 2, 5

[24] 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. 7

[25] 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. 1, 2, 4, 5, 6, 7, 8

[26] Chunyu Li, Yusuke Monno, Hironori Hidaka, and Masatoshi Okutomi. Pro-cam ssfm: Projector-camera system for structure and spectral reflectance from motion. In *IEEE/CVF International Conference on Computer Vision*, pages 2414–2423, 2019. 1

[27] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Wei Li, Christian Theobalt, Ruigang Yang, and Wenping Wang. Adaptive surface normal constraint for depth estimation. In *IEEE/CVF International Conference on Computer Vision*, pages 12849–12858, 2021. 4

[28] 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. 7

[29] Daisuke Miyazaki, Robby T Tan, Kenji Hara, and Katsushi Ikeuchi. Polarization-based inverse rendering from a single view. In *IEEE International Conference on Computer Vision*, volume 3, pages 982–982, 2003. 2

[30] 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. 5

[31] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pages 127–136, 2011. 1

[32] Trung Thanh Ngo, Hajime Nagahara, and Rin-ichiro Taniguchi. Surface normals and light directions from shading and polarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[33] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32(6):1–11, 2013. 1

[34] Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations*, 2022. 1

[35] Stefan Rahmann and Nikos Canterakis. Reconstruction of specular surfaces using polarization imaging. In *IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001. 1, 2, 3, 4

[36] 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. 5

[37] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1

[38] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. 8

[39] 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. 2, 4

[40] 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, 2

[41] Silvia Tozza, Dizhong Zhu, William Smith, Ravi Ramamoorthi, and Edwin Hancock. Uncalibrated, two source photo-polarimetric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[42] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE international conference on computer vision*, pages 4489–4497, 2015. 3

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017. 8

[44] Yannick Verdié, Jifei Song, Barnabé Mas, Benjamin Busam, Aleš Leonardis, and Steven McDonagh. Cromo: Cross-modal learning for monocular depth estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3927–3937, 2022. 5

[45] Xiaolong Wang, David Fouhey, and Abhinav Gupta. Designing deep networks for surface normal estimation. In *IEEE conference on computer vision and pattern recognition*, pages 539–547, 2015. 6

[46] Tomohiro Yamazaki, Yasushi Maruyama, Yusuke Uesaka, Motoaki Nakamura, Yoshihisa Matoba, Takashi Terada, Kenta Komori, Yoshiyuki Ohba, Shinichi Arakawa, Yasutaka Hirasawa, et al. Four-directional pixel-wise polarization cmos image sensor using air-gap wire grid on 2.5-μm back-illuminated pixels. In *2016 IEEE International Electron Devices Meeting (IEDM)*, pages 8–7, 2016. 1, 3

[47] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2019. 6

[48] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000. 5

[49] Jinyu Zhao, Yusuke Monno, and Masatoshi Okutomi. Polarimetric multi-view inverse rendering. In *European Conference on Computer Vision*, pages 85–102. Springer, 2020. 1

[50] 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, 6