DPS-Net: Deep Polarimetric Stereo Depth Estimation

Chaoran Tian[1] Weihong Pan[1] Zimo Wang[1] Mao Mao[1] Guofeng Zhang[1]
Hujun Bao[1] Ping Tan[2] Zhaopeng Cui[1*]
[1]State Key Lab of CAD&CG, Zhejiang University [2]Hong Kong University of Science and Technology

Abstract

Stereo depth estimation usually struggles to deal with textureless scenes for both traditional and learning-based methods due to the inherent dependence on image correspondence matching. In this paper, we propose a novel neural network, i.e., DPS-Net, to exploit both the prior geometric knowledge and polarimetric information for depth estimation with two polarimetric stereo images. Specifically, we construct both RGB and polarization correlation volumes to fully leverage the multi-domain similarity between polarimetric stereo images. Since inherent ambiguities exist in the polarization images, we introduce the iso-depth cost explicitly into the network to solve these ambiguities. Moreover, we design a cascaded dual-GRU architecture to recurrently update the disparity and effectively fuse both the multi-domain correlation features and the iso-depth cost. Besides, we present new synthetic and real polarimetric stereo datasets for evaluation. Experimental results demonstrate that our method outperforms the state-of-the-art stereo depth estimation methods.

1. Introduction

Stereo depth estimation is a classical computer vision task, playing an important role in many applications, *e.g.*, autonomous driving, robotics, and VR/AR. Traditional methods [16, 21, 27, 33, 36, 44, 47] often solve a per-pixel depth by finding pixelwise correspondences across input images, which thus normally struggle to handle textureless areas. Recently, many methods [5, 6, 15, 20, 46, 14] have been proposed to exploit deep learning techniques and achieved promising results. Thanks to their learned shape priors, these methods can generate smooth results in textureless regions. But the recovered depth at textureless regions is unreliable since the correspondence is intrinsically ambiguous.

To reconstruct textureless objects, some previous methods exploit additional shape cues, such as photometric stereo [12, 30, 32, 50], shape-from-shading [35, 31, 40], and shape-from-polarization [1, 18, 28, 29, 31, 2]. Among them, the polarization constraint is passive and provides strong information on local surface orientations. Thus, it attracts more and more attention for 3D reconstruction and has been integrated with multi-view stereo [8, 49], SLAM [43], stereo normal estimation [13] and stereo depth estimation [51, 25]. These methods combine multi-view triangulation with the surface normal estimated from the polarization, and therefore improve the 3D reconstruction at textureless regions. However, these methods usually assume a smooth surface or controlled illumination, which is normally limited to object-level scenarios or specific capturing environments. Moreover, these methods struggle to handle the inherent azimuth or zenith angle ambiguities [7] in the surface estimation from the polarization images, *e.g.*, relying on a pre-computed coarse depth or assuming known reflection types.

In this paper, we aim to design an end-to-end neural network to predict the depth from polarimetric stereo images, which learns to exploit the polarization information for accurate depth estimation of complex scenes. However, it is nontrivial to design such a model due to the following challenges. First, since the polarimetric and photometric domains maintain different characters of the scene, a straightforward combination may not fully exploit the multi-domain information. Moreover, similar to traditional polarimetric 3D modeling methods, we need to solve the inherent ambiguities in the polarization information, including the azimuth and zenith angle ambiguities. Lastly, there are few real polarimetric stereo datasets that are sufficient for network training.

To deal with these challenges, we propose a novel DPS-Net with a cascaded dual-GRU (Gated Recurrent Unit) module that tightly integrates the polarization constraints and learned shape priors for robust polarimetric stereo depth estimation. Specifically speaking, to better exploit the multi-domain knowledge, we first extract per-pixel features and construct both RGB and polarization correlation volume pyramids for stereo matching, and explicitly introduce the geometric constraint from the polarization into the net-

*Corresponding author.

work. In order to handle the inherent ambiguities of polarization information, we adopt the iso-depth constraints instead of direct surface normal constraint in the network to bypass both the zenith angle ambiguity and π-ambiguity of azimuth angle, and construct a unified iso-depth cost with different reflection to handle the $\pi/2$-ambiguity of azimuth angle. In addition, along the gradient descent direction of the iso-depth cost, we further calculate the virtual disparity and retrieve the virtual feature from the correlation pyramids for the information integration in the followed refinement stage. To fully exploit the multi-domain similarity (which refers to the similarity of the RGB and the polarization images) and iso-depth cost for stereo depth estimation, we design a cascaded dual-GRU architecture in our network. In the first GRU of each iteration, the disparity increment is regressed directly from the multi-domain correlation feature. Then, the disparity is further optimized in the second GRU based on the iso-depth cost and the virtual correlation feature. Through the cascaded dual-GRU architecture and the recurrent optimization process, our method can balance the diverse inputs well and resist the disturbance of polarimetric noises under uncontrolled illumination. At last, we present a synthetic dataset simulating the polarimetric stereo pattern and a real dataset captured by a polarization stereo system for the network training.

Our contributions can be summarized as follows. Firstly, we propose the first learning-based framework for polarimetric stereo depth estimation, which exploits both prior knowledge and polarization information. Secondly, we present a novel cascaded dual-GRU architecture and integrate both the multi-domain similarity measurement and the iso-depth constraint from the polarization cue to solve the inherent ambiguities in the polarization and enable high-quality and robust stereo estimation. Thirdly, we present both synthetic and real polarimetric stereo datasets, which can also benefit the community. We conduct extensive experiments, and the results show that our method performs better than the state-of-the-art learning-based methods and traditional polarimetric methods.

2. Related Work

Learning-Based Stereo Depth Estimation. Stereo depth estimation, one of the fundamental and long-lasting problems in computer vision, has been continuously studied with the help of algorithms from traditional pixel-level matching to learning-based ones. For most traditional methods, there are a matching stage and a filter stage, which can be classified into local and global methods. The local methods [4, 17, 27] use a support window to determine the pixel-level match, while the global methods [36, 21] follow the pattern of the energy minimization problem.

The deep neural network benefits the prediction of disparity to a considerable degree. Mayer et al. proposed the first end-to-end stereo matching network DispNet [26]. After that, there are a variety of learning-based methods being presented and introducing the cost volume or the correlation volume into the network, such as GC-Net [20] and PSM-Net [5]. Then DSM-Net [42], a domain-invariant stereo matching network, is proposed, adopting the domain normalization and a structure-preserving filter to overcome the dependency on the training dataset. In LEA-Stereo [6], a hierarchical NAS framework is presented. It enables the network to choose among operations with the winner-take-all strategy and find an optimal architecture.

In order to compress the model and accelerate the estimation, many works have refined the 3D convolutional layers with other innovative layers, including GA-Net [45], AANet [41], EDNet [48], and HITNet [37]. Moreover, in the Lac-Gwc Net [24], a pairwise feature is adopted to reveal the neighbor relationships, and then the disparity is refined by a cost self-reassembling procedure. Based on the work about RAFT [38], a multi-level convolutional GRU-based network is proposed to exploit the visual similarity and update the disparity recurrently called RAFT-stereo [23]. Not only does it saves time and memory cost, but its depth estimation also achieves a higher level. These methods are derived from the volumetric approach to leverage contextual information for stereo matching.

Depth Estimation and Shape Recovery with Polarization. Polarization can provide surface normal information, which is helpful for depth estimation, especially in the textureless and the highlight region. Many works related to depth estimation or shape recovery have introduced polarization information in recent years. Cui et al. [8] present a traditional method of depth estimation for multi-view stereo with polarization. They estimate depth and resolve the polarimetric ambiguity by graph optimization and iso-depth contour tracing. Zhu et al. [51] also leverage polarization images with RGB stereo to present a depth estimation.

Some works take further steps to recover normal information with the help of stereo polarization images. Fukao et al. [13] construct both the regular RGB cost volume and the polarization cost volume for normal estimation. In order to solve the unknown reflectance and illumination under the sky, Ichikawa et al. [19] estimate surface normals by decoding the polarization pattern of the sky. Ding et al. [11] extend the classical Helmholtz stereopsis to the polarimetric case, achieving accurate surface recovery. As for the shape estimation and SVBRDF acquisition, Valentin et al.[10] present an improved U-Net applied for single-view polarization imaging under flash illumination. Lei et al. [22] first present a scene-level normal estimation network. Recently, PANDORA [9], a polarimetric neural inverse rendering approach for 3D Reconstruction and reflectance separation, has been proposed and exhibits excellent performance for object-level reconstruction.

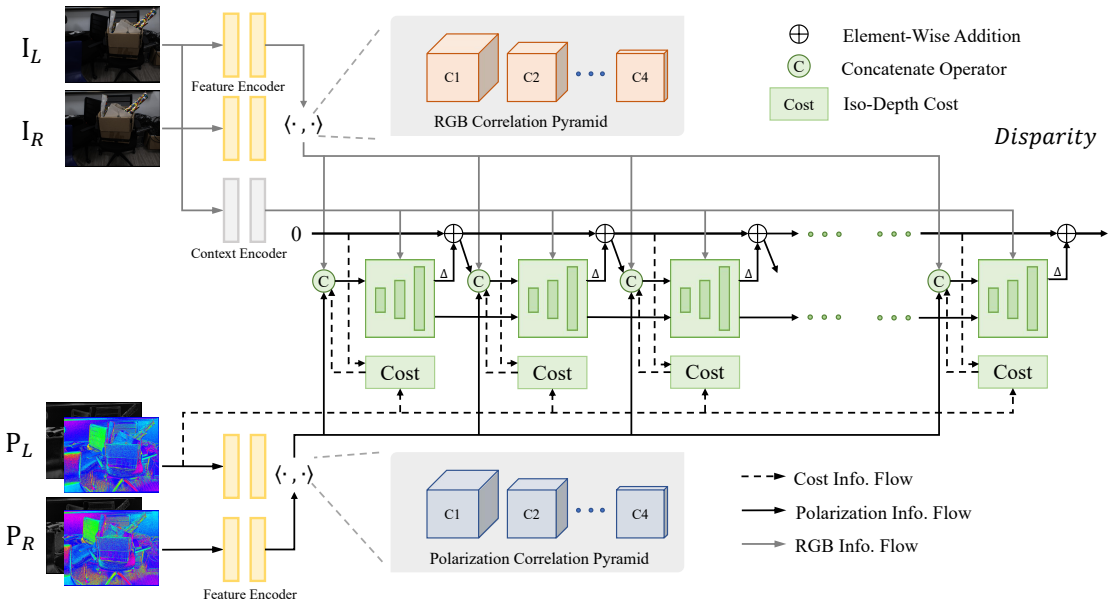

Figure 1. The pipeline of the proposed depth estimation method: First, the multi-domain features are extracted, and the RGB and polarization correlation pyramids are constructed. Secondly, the iso-depth cost is calculated from the polarization cue. Finally, taking advantage of the cascaded dual-GRU architecture, the disparity is recurrently updated.

3. Preliminaries

It is well known that the light wave is a transverse wave whose oscillation direction is perpendicular to the propagation direction. Based on the property of the transverse wave, the polarimetric images with different phase angles can be obtained via a rotating polarizer. As proved in [7], the relationship between the polarization angle ϕ_c and the intensity of polarimetric images satisfies the following equation,

$$\begin{aligned} I\left(\phi_c\right) &= I_{\max}\cos^2\left(\phi_c-\phi\right)+I_{\min}\sin^2\left(\phi_c-\phi\right), \\ &= \bar{I}+\rho\bar{I}\cos\left(2\phi_c-2\phi\right), \end{aligned} \tag{1}$$

where I_{max} and I_{min} are the maximum and minimum intensities, $\bar{I}=\left(I_{max}+I_{min}\right)/2$ is the average intensity value. The angle of linear polarization (AoLP) is defined as ϕ, corresponding to the maximum intensity. The degree of linear polarization (DoLP) is defined as ρ, which relates to the relative variant of intensities.

The light becomes partially polarized when it reflects from the surface of the object. For the common case, diffuse reflection or specular reflection dominates at each pixel. The relationship between the DoLP, the AoLP, the zenith angle θ, and the azimuth angle φ satisfies different equations under two reflections. For diffuse case, we have

$$\begin{gathered} \rho_d=\frac{\left(\eta-1/\eta\right)^2\sin^2\theta}{2+2\eta^2-\left(\eta+1/\eta\right)^2\sin^2\theta+4\cos\theta\sqrt{\eta^2-\sin^2\theta}}, \\ \phi_d=\varphi \text{ or } \phi_d=\varphi+\pi, \end{gathered} \tag{2}$$

where η is the refractive index of the surface material. For specular reflection, we have

$$\begin{gathered} \rho_s=\frac{2\sin^2\theta\cos\theta\sqrt{\eta^2-\sin^2\theta}}{\eta^2-\sin^2\theta-\eta^2\sin^2\theta+2\sin^4\theta}, \\ \phi_s=\varphi\pm\frac{\pi}{2}. \end{gathered} \tag{3}$$

Surface normals can be estimated by solving the azimuth and zenith angles from the above equations, while multiple solutions exist for both azimuth and zenith angles due to the unknown reflection types and the nonlinear equation, which is also known as azimuth angle ambiguity and zenith angle ambiguity. Please refer to the supplementary material for more details on polarization ambiguities. So it is crucial to deal with ambiguities when utilizing polarization.

4. Method

As shown in Fig. 1, our entire pipeline can be divided into four steps. Firstly, we extract the multi-domain features from RGB and polarization images, respectively. Then, we construct the 3D correlation volume of RGB and polarization for all pairs of pixels at the same height. Next, we compute the iso-depth cost according to the polarization constraints. Finally, the correlation features and iso-depth cost are retrieved recurrently by the GRU-based update operator. The disparity map is updated, and the reflection ambiguity is resolved in the updating step.

4.1. Multi Domain Feature and Correlation Volume

Feature Extraction. We extract features from RGB and polarization images to construct the multi-domain correlation volume. The RGB feature encoder and the polarization feature encoder are formed by two convolutional networks with the same backbone. The 256 channels of RGB features and polarization features are extracted from the 2-channel polarization images and the 3-channel RGB images, respectively.

We extract the context features from RGB images by the context encoder, which is used to produce the input state and the hidden state of the update operator. RGB images provide more consistent contextual information than polarization images under different illumination conditions. Therefore, we extract the context features from the RGB images to ensure the consistency of the context information.

Correlation Volume. Similar to RAFT-Stereo [23], we construct correlation volumes for both the RGB domain and the polarization domain. Furthermore, the 3D polarization correlation volume can measure polarization similarity as a complement to RGB information. Especially in the case of low texture and few features, the introduced polarization cue plays an important role. We compute the visual similarity between the polarization and RGB images according to the dot product between the feature vectors. Then, we construct multi-domain 3D correlation volumes based on the visual similarity of all pixel pairs with the same height as the following,

$$\begin{aligned} \mathbf{IC}_{ijk} &= \sum_h \mathbf{f}^I_{ijh} \cdot \mathbf{g}^I_{ikh}, \quad \mathbf{IC} \in \mathbb{R}^{H\times W\times W}, \\ \mathbf{PC}_{ijk} &= \sum_h \mathbf{f}^P_{ijh} \cdot \mathbf{g}^P_{ikh}, \quad \mathbf{PC} \in \mathbb{R}^{H\times W\times W}, \end{aligned} \tag{4}$$

where $\mathbf{IC}_{ijk}$ denotes RGB correlation volume and $\mathbf{PC}_{ijk}$ denotes Polarization correlation volume. We represent the h-th dimension of stereo correlation features that belong to the RGB domain and polarization domain by $\{\mathbf{f}^I_{ijh}, \mathbf{g}^I_{ikh}\}$ and $\{\mathbf{f}^P_{ijh}, \mathbf{g}^P_{ikh}\}$. Compared with the 4D cost volume in most stereo depth estimation networks [20, 5], 3D correlation volume is more efficient and takes up less memory.

We construct the 4-layer correlation volume pyramids for polarization and RGB domains, respectively. By pooling the last dimension of the correlation volume with scaled kernel size, we obtain the layers in the correlation pyramids. Similar to RAFT-Stereo [23], we use a lookup operator for fast indexing. With the help of a linear interpolator, the retrieved value can be efficiently obtained by looking up the correlation volume at each pyramid layer. Finally, by concatenating retrieved values, we get the correlation feature of the given disparity.

4.2. Iso-Depth Cost and Ambiguity Solver

The Iso-Depth Constraints of the Normal. Note that the iso-depth contour is orthogonal to the azimuth component of the normal vector of the point on the contour. After differentiating the depth and substituting the depth with the disparity, we can obtain an approximate representation of the azimuth angle as:

$$\tan(\varphi) = \frac{f_y}{f_x} \frac{(\mathrm{d}_{0,-1} - \mathrm{d}_{0,1})(\mathrm{d}_{-1,0} + \mathrm{d}_{1,0})}{(\mathrm{d}_{-1,0} - \mathrm{d}_{1,0})(\mathrm{d}_{0,-1} + \mathrm{d}_{0,1})}, \tag{5}$$

where $\mathrm{d}_{i,j}$ represents the disparity of neighbor pixel $P(u+i, v+j)$ relative to pixel $P(u, v)$. f denotes the focal length.

The Iso-Depth Cost of Polarization. As mentioned above, we construct a unified iso-depth cost to explicitly utilize the geometric constraints of the polarization and handle the π-ambiguity and $\pi/2$-ambiguity between the azimuth angle φ and the AoLP ϕ. Our proposed iso-depth cost $\mathbf{C}(\varphi)$ bypasses the π-ambiguity similar to [8] and integrates the azimuth cost under the specular and diffuse reflection through the minimizing operator. Furthermore, the ambiguity of reflection types $\mathbf{R}(\varphi)$ can also be resolved as follows:

$$\begin{aligned} &\mathbf{C}_s(\varphi) = \left[sin\left(\phi\right) sin\left(\varphi\right) + cos\left(\phi\right) cos\left(\varphi\right)\right]^2, \\ &\mathbf{C}_d(\varphi) = \left[sin\left(\phi\right) cos\left(\varphi\right) - cos\left(\phi\right) sin\left(\varphi\right)\right]^2, \\ &\mathbf{C}(\varphi) = \min\left\{\mathbf{C}_s(\varphi), \mathbf{C}_d(\varphi)\right\}, \\ &\mathbf{R}(\varphi) = \arg\min\left\{\mathbf{C}_s(\varphi), \mathbf{C}_d(\varphi)\right\}, \end{aligned} \tag{6}$$

where the costs of the diffuse and the specular reflection are noted as $\mathbf{C}_d$ and $\mathbf{C}_s$. The derivation of iso-depth cost is detailed in the supplementary.

The Virtual Disparity and Virtual Correlation Features. In order to suppress the cost disturbance caused by polarization noise and fully exploit the iso-depth constraint, we calculate the virtual disparity from the iso-depth cost and generate the related virtual correlation feature. The virtual disparity and the virtual features can be utilized in the optimization GRU (introduced in Sec. 4.3) of the dual-GRU architecture for disparity refinement. For the sake of distinction, we call the recurrently updated disparity in the update block the actual disparity d_a. The virtual disparity d_v can be obtained based on the gradient of the iso-depth cost $\mathbf{C}(\varphi, d)$ and the iteratively updated step length α by the following equation:

$$d_v = d_a - \alpha \nabla_d \mathbf{C}(\varphi, d). \tag{7}$$

4.3. Hybrid GRU-based Update Operator

We present a hybrid GRU-based update operator to exploit the multi-domain information and update the disparity recurrently. The framework is illustrated in Fig. 2. The multi-domain contextual information and geometric information are fused by the update block. We extract the multi-

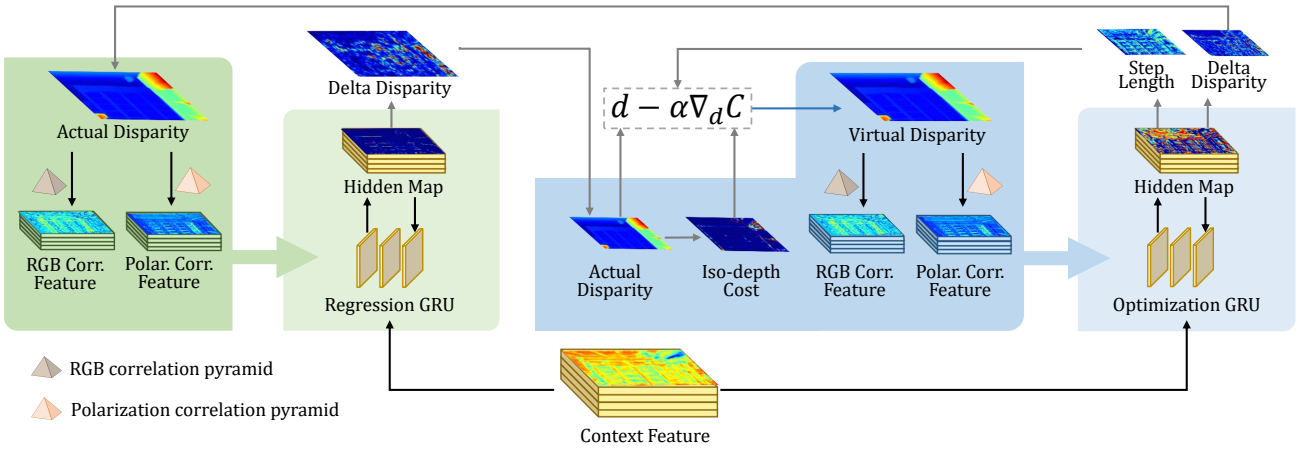

Figure 2. The hybrid GRU-based update operator is shown in this figure to interpret how multi-domain correlation features and iso-depth cost are processed and fused. The cascaded dual-GRU architecture consists of a regression GRU that provides an initial disparity prediction and an optimization GRU that further refines the disparity.

domain correlation features through two correlation volume pyramids to leverage the multi-domain visual similarity. Moreover, to simulate the traditional optimizer, we design a cascaded dual-GRU paradigm to recurrently regress and optimize the disparity based on the actual and virtual correlation features. Specifically, the virtual disparity and the related virtual correlation features are generated from the iso-depth cost to guide the optimization procedure in an indirect fashion.

Cascaded Dual-GRU Architecture. We design a cascaded dual-GRU architecture to combine the iso-depth cost with the visual similarity. The cascaded dual-GRU architecture is composed of a regression GRU and an optimization GRU, which implement a regress-and-optimize recurrently updating paradigm. In the regression GRU, the disparity is regressed intuitively from the multi-domain correlation features of the actual disparity, which leverage the similarity of the RGB and polarization images to produce an initial prediction for the subsequent optimization. In the optimization GRU, the iso-depth cost is injected directly and introduced in an indirect manner by the virtual correlation feature. The optimization GRU fuses the geometric constraint and further rectifies the disparity.

We introduce the iso-depth cost with the help of the virtual correlation features rather than updating the disparity directly by the increment calculated from the cost gradient, which avoids the following difficulties in the traditional optimization process.

On the one hand, it is challenging to explicitly transform the matching problem coupled with the contextual information into an optimization problem. On the other hand, it is still difficult to handle that formulated optimization due to the severe non-convexity. Considering forehead facts, we adopt a hybrid optimization GRU to avoid the local convergence in optimization and resist the adverse effects of polarimetric noise. Moreover, the cascaded scheme can balance the visual similarity and the geometric constraint well. We can gradually refine the disparity in different iterations through the regress-and-optimize cascaded paradigm. Details of the cascaded dual-GRU architecture are deferred to the supplementary.

Multi-Domain Inputs. We compose the input of the GRU utilizing the multi-domain information. Specially, we adopt different input settings for the two GRUs in cascaded dual-GRU architecture. The regression GRU is provided with the actual multi-domain features, while the optimization GRU is introduced into the virtual multi-domain features. Before injecting into the GRU, we use encoders composed of two layers of convolution to process the multi-domain inputs respectively. Finally, the multi-domain correlation features, along with other encoded features, are concatenated to form the inputs of GRU, as shown in Fig. 2.

Update. The GRU module is employed as the core component of the update operator module to implement updating of the disparity and the cost step. Both the regression GRU and the optimization GRU contain a hidden state. The hidden state is updated continuously during different iterations based on the encoded hybrid input. What's more, we can obtain the disparity increment and the optimization step by decoding the hidden state through different Head Nets. Finally, the disparity updated by the new increment is upsampled to recover the full-resolution disparity map.

Loss Function. We train our DPS-Net by minimizing the L_1 difference between the predicted disparity and the ground truth over the full predictions sequence. We encourage the predictions corresponding to different iterations by

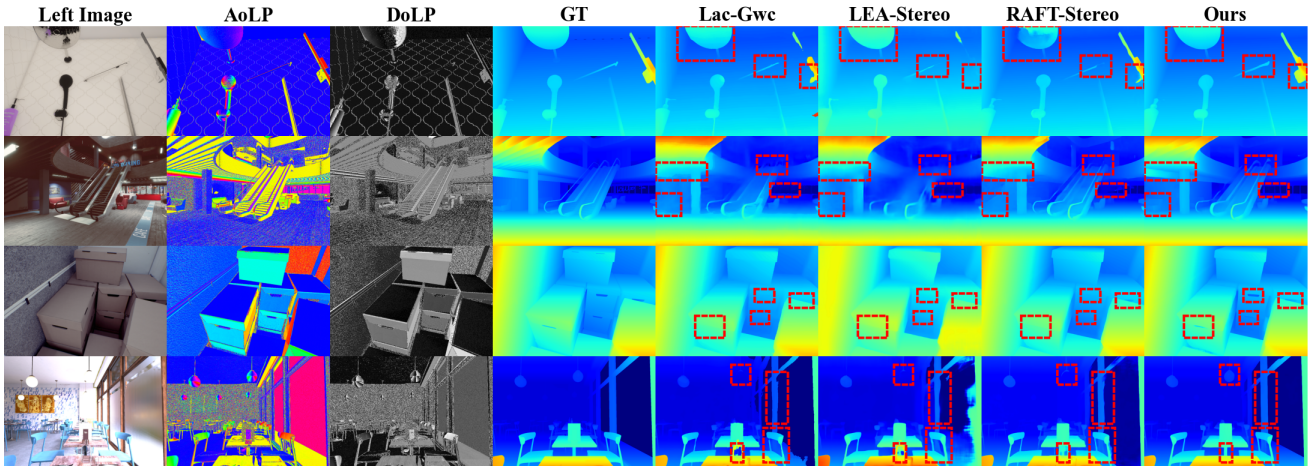

Figure 3. Qualitative results on IPS dataset. **From left to right:** input left images, AoLP images, DoLP images, the ground truth of disparity, disparity results from the existing stereo methods, and those from our method. Please zoom in for a further comparison.

the exponential weights γ^{N-i}:

$$\mathbf{L} = \sum_{i=1}^{N} \gamma^{N-i} \left\| d_{gt} - d_i \right\|_1 . \tag{8}$$

5. Experiments

Since we cannot train and test our method with the existing stereo benchmark datasets due to the lack of polarimetric information, we utilize an existing indoor robotics stereo dataset (i.e., IRS dataset [39]) with accurate surface normals to generate a synthetic polarimetric stereo dataset named IPS dataset. We also capture a real polarimetric dataset consisting of both indoor scenery and outdoor scenery named the RPS dataset for the fine-tuning of our network. Both the IPS and RPS datasets are utilized to demonstrate the effectiveness of our DPS-Net.

Implementation Details. We implemented DPS-Net using PyTorch. We employ an AdamW optimizer for 100 epochs on the synthetic dataset and 150 epochs on the real dataset. We set the learning rate by a one-cycle schedule with an initial learning rate of 0.002. Additionally, we pre-train and fine-tune the network on a single 3090 GPU, spending 50 GPU hours and 4 GPU hours. To preserve raw physical information of the polarization data, we only augment the data by randomly cropping images to a resolution of 320x720.

Benchmark Metrics. We report the average endpoint error (AvgErr), RMS of the endpoint error (RMSE), and the bad pixel ratio with the 2-pixel threshold (bad 2.0) to quantify the quality of depth estimation. We compare our approach with other approaches using the benchmark metrics on the IPS and the RPS dataset.

Stereo Depth Estimation Methods for Comparison. We compare our network with other SOTA learning-based methods, including RAFT-Stereo [23], LEA-Stereo [6], and Lac-Gwc [24] in Sec. 5.1 and Sec. 5.2. Since all those methods are designed to estimate depth utilizing the stereo RGB images, we only use RGB stereo images in both datasets to evaluate the comparison methods. Among those methods, RAFT-Stereo adopts the recurrently GRU-based network, while the LEA-Stereo and Lac-Gwc utilize the feed-forward CNN. Therefore, we conduct further ablation studies on different iterations for RAFT-Stereo and our method in supplementary. We also perform an experiment to compare with the traditional stereo depth estimation method utilizing polarimetric information [51] on IPS and RPS in Sec. 5.3. In [51], coarse depth guidance is estimated from the stereo images first. Then the ambiguity is resolved by the graphical model, and the depth is optimized by a linear function.

5.1. Experiment on Synthetic Dataset

In order to generate a large-scale polarimetric dataset for pre-training our model, we choose the IRS dataset as the raw dataset to generate the synthetic data, which provides accurate surface normal and contains rich indoor scenarios compared to other stereo datasets, such as KITTI, ETH3D, SceneFlow, etc. We generate the polarization images based on the surface normal to manufacture our dataset IPS (see supplementary). We pre-train DPS-Net and other stereo methods on the training set of the IPS dataset and then evaluate the final models of the last epoch on the test set. Specifically, our approach and RAFT-Stereo[23] are trained for 100 epochs with random initial parameters. LEA-Stereo[6] and Lac-Gwc[24] are fine-tuned based on the provided pre-trained models for 50 epochs.

Fig. 3 visualizes the RGB and the polarization input, the ground truth of the disparity, and the disparity results of different methods. It can be seen that our method obtains better results than other methods. Moreover, our approach can

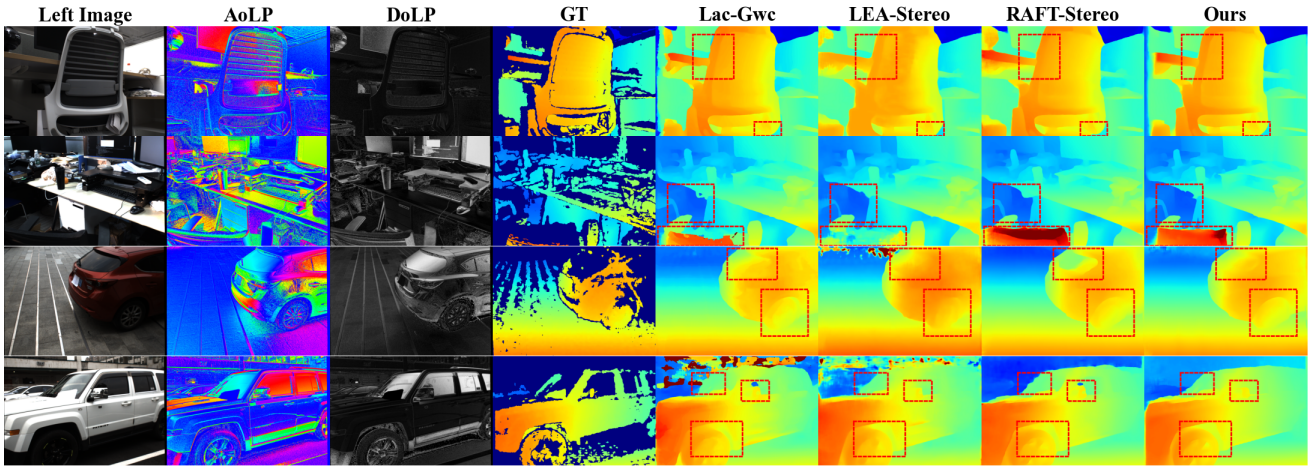

Figure 4. Qualitative results of our DPS-Net and other methods on the RPS dataset. Our method generates a more accurate and sharp disparity. Clearer object boundaries and better overall structures also can be restored. Please zoom in for details.

Method	AvgErr	RMSE	bad 2.0
Lac-Gwc[24]	1.2135	3.4503	8.2061
LEA-Stereo[6]	1.6094	3.5475	11.247
RAFT-Stereo[23]	0.9266	2.6755	6.9791
Ours	**0.5790**	**1.8616**	**3.9705**

Table 1. Comparison to the state-of-the-art stereo matching methods on the IPS dataset. Our method beats others with superior accuracy and makes the error slump drastically in all the metrics.

recover an accurate disparity in the low texture and featureless region as expected. Compared with other methods, our DPS-Net can generate a sharper disparity on the boundary of the objects. Moreover, we can estimate the disparity of slender structures such as thin rods and wires, which are quite challenging for other methods. The sharp and dense disparity maps demonstrate our approach's ability to fuse the RGB and polarization and handle the cases of complex scenes. More results are shown in the supplementary.

The quantitative comparison results are listed in Table 1. It clearly illustrates that our algorithm significantly outperforms the SOTA stereo depth estimation methods in all metrics and gives significantly lower errors than LEA-Stereo and Lac-Gwc. Compared with RAFT-Stereo which also uses a GRU-based network structure, our method has obviously lower average errors, demonstrating the vital role of polarization information for stereo depth estimation.

5.2. Experiment on Real Datasets

We set up an imaging system equipped with stereo polarimetric cameras and an RGB-D camera for data collection. The stereo polarization images are captured by the stereo polarimetric cameras, and the ground truth of the depth is provided by the RGB-D camera. The RPS dataset comprises over 1,000 pairs of indoor stereo images and 300 pairs of outdoor stereo images (see supplementary) and is divided into training set, validation set, and test set. Then we fine-tune the proposed method and the SOTA methods on the training set of RPS for 150 epochs and select the model with minimal endpoint error on the validation data. Finally, we evaluate the selected model on the test set using the AvgErr, bad 2.0, and Runtime metrics.

Method	AvgErr	bad 2.0	Runtime(s)
Lac-Gwc[24]	0.6919	3.6735	0.702
LEA-Stereo[6]	0.7518	5.0487	0.359
RAFT-Stereo[23]	0.6807	3.8864	0.352
Lac-Gwc-RBGP	0.6674	3.4693	0.798
LEA-Stereo-RGBP	0.7517	5.1257	0.577
RAFT-Stereo-RGBP	0.6244	3.4306	0.958
Ours	**0.6187**	**3.3541**	**0.255**

Table 2. Comparison to the state-of-the-art stereo matching methods on the RPS dataset. Our recovery is the precisest and fastest outcome, fitting with the ground truth much better than others.

Fig. 4 shows the qualitative results of different methods on the RPS dataset. In contrast to other methods, our method can produce more accurate disparity with sharper object boundaries and finer overall structures. Our method has a better performance in the textureless and the highlight regions with the help of multi-domain information. More results are shown in the supplementary. As shown in the first three rows and the last row of Table 2, we can see that our method also outperforms other methods with the smallest AvgErr on the RPS dataset. Besides, Table 2 also illustrates that our approach is more efficient and takes shorter runtime than other methods. The experiment results on the RPS dataset demonstrate that our method can effectively handle both indoor and outdoor scenarios.

These existing stereo methods are targeted at pure RGB

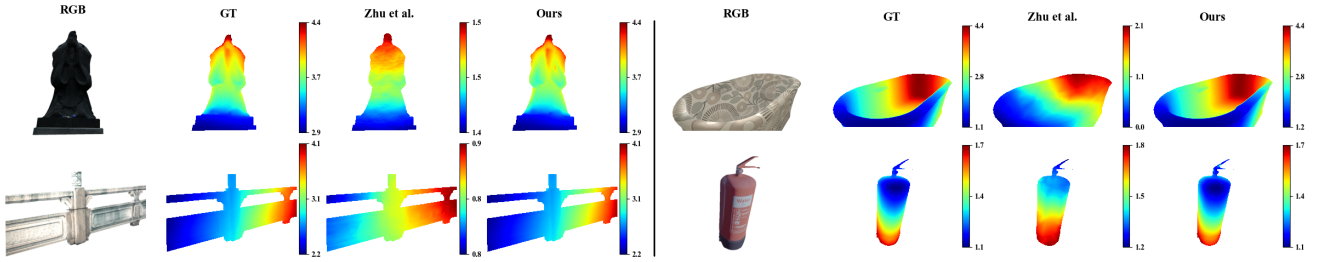

Figure 5. Qualitative results of our DPS-Net and traditional polarimetric method [51]. The RPS dataset examples are exhibited on the left, and the IPS dataset examples are displayed on the right. Each image has a respective color bar depending on the depth range of the depth image. Our results fit with the ground truth well. The predictions of [51] have consistent variation but different ranges with ground truth.

stereo applications. To further compare the performance of utilizing the multi-domain information, we modify these methods and feed the same RGB and polarization images as our method. Specifically, we add the polarization correlation pyramid based on the RAFT-Stereo[23] pipeline and inject both the RGB and polarization features into the GRU. We also replace the pure RGB input in the LEA-Stereo[6] and the Lac-Gwc[24] with the multi-domain input. As shown in Table 2, though the performance of these methods has been improved with the help of the extra polarization information, our method still outperforms other methods in accuracy and efficiency, which demonstrate the effectiveness of our cascaded dual-GRU architecture. To be noted, our ablation ("(RGB+POL)" in Table 3) is similar to the RAFT-Stereo-RGBP. The difference is that we use RAFT as the backbone, and it is light-weighted and more efficient than RAFT-Stereo, which has multi-level GRUs.

5.3. Comparison with Other Polarimetric Methods

Our method focuses on the scene-level depth estimation application, while the method in [51] is an object-level method. To fit the smooth object assumption in [51], we select the suitable object in both the IPS and the RPS dataset and mask the object manually. As mentioned in [51], we produce the initial coarse depth map by the classical SGM [16] and calculate the light source using the method in [34]. Qualitative results are shown in Fig. 5. There are consistent trends of disparity variation between different methods. Our DPS-Net has achieved closer results to the ground truth. However, the depth scale of the result of [51] has a large deviation though we try our best to meet the requirements in [51]. The deviation may mainly be caused by the different application conditions, including the uncontrolled light source, the material with an unknown refractive index, and complicated reflectance contrary to Lambert's assumption.

We also conduct comparison experiments with recent polarimetric normal estimation methods including SPW [22] and DeepSfP [3] to demonstrate the additional normal estimation performance of our methods. DeepSFP applies deep learning techniques to recover the shape from polarimetric images of a single view. SPW performs the scene-level normal estimation through the network. We adopt the same pre-training and finetuning process for DeepSfP and SPW as our method. As shown in Fig. 6, our method can generate more clear and more accurate surface normal results, which further validates the effectiveness of our method. Though we do not introduce normal supervision in our network, it still achieves better performance for normal estimation. The performance of SPW [22] and DeepSfP [3] decreases on our real dataset that contains complicated geometry structures and diverse materials. Please refer to the supplementary material for more normal estimation results.

5.4. Ablation Study

To better characterize our method, we perform the ablation study on the RPS dataset and measure the depth estimation quality by AvgErr and bad 2.0.

Will the polarization similarity help the stereo-matching? We set up three variants of the update block to study the implication of multi-domain correlation features. As shown in Table 3, the three update blocks share a similar backbone, only distinguished by the different correlation features input. Concretely, the (RGB+POL) variant is fed with the multi-domain correlation features, while the RGB-only variant and the POL-only variant are injected into single-domain correlation features of RGB or polarization. It can be seen that the RGB-Only variant has a better performance than the POL-Only variant, which implies that the RGB image contains more reliable similarity information in contrast to the noisy polarization images under uncontrolled illumination. The results also show a notable improvement of the (RGB+POL) variant, outperforming the RGB-only and POL-only variants by a clear margin, which validates the effectiveness of multi-domain similarities.

Will properly introducing the iso-depth cost facilitate stereo depth estimation? The proposed cascaded dual-GRU architecture in Sec. 4.3 archives a higher accuracy than the variants in the first three rows of Table 3, which reveals that the iso-depth constraints can further benefit the depth estimation and improve the robustness to the differ-

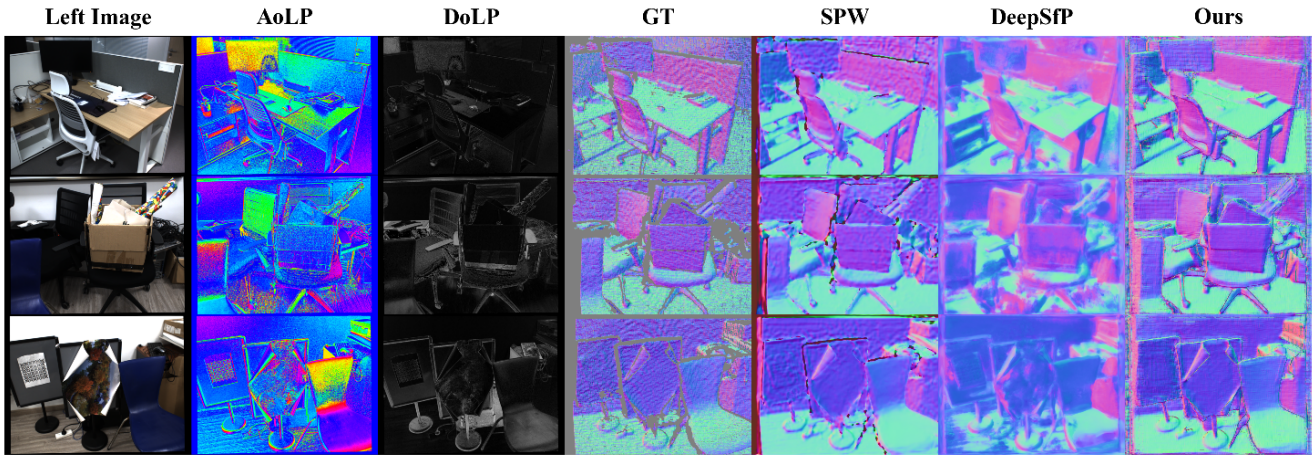

Figure 6. Qualitative normal estimation results of our method, SPW[22] and DeepSfP[3] on the RPS dataset. Our DPS-Net generates better surface normal not only in the inner area but also at the boundary of the object.

Method	AvgErr	bad 2.0
RGB-Only	0.6845	4.3582
POL-Only	1.1278	10.537
(RGB+POL)	0.6338	3.6850
(RGB+POL+Cost)	0.6649	4.0896
(RGB+POL) $\oplus$(Cost)	1.3745	15.813
(RGB+POL+vRGB+vPOL)	0.6285	3.6029
Cascaded Dual-Convs	0.6384	3.7660
Cascaded Dual-GRU	**0.6187**	**3.3541**

Table 3. Comparison to the DPS-Net baseline with Ablation Studies. It clearly reveals that the cascaded dual-GRU architecture of DPS-Net is most effective than other variants. The bracket is used for indicating a complete GRU. The symbol '$\oplus$' indicates the cascaded GRUs. The symbol '$+$' denotes multiple inputs that are injected into the same GRU.

ent environments. To further analyze the effectiveness of the cascaded dual-GRU architecture, we build three other modifications of the update block with different cost usage modes. In the (RGB+POL+Cost) variant, the cost is injected into a single GRU along with the actual correlation features. Compared with (RGB+POL) variant, the metrics of the (RGB+POL+Cost) variant deteriorate slightly, possibly caused by the noisy polarization and the deviation of the iso-depth cost. In the (RGB+POL)$\oplus$(Cost) variant, a similar dual-GRU scheme as cascaded dual-GRU architecture is adopted, where the virtual correlation features are replaced by the iso-depth cost. The accuracy is severely decreased due to the amplification of the noise in the independent GRU. In the (RGB+POL+vRGB+vPOL) update block, both the actual and the virtual correlation features are input into a single GRU. The performance improved slightly, which illustrates that the iso-depth cost can be exploited better through the correlation feature. In contrast, the approach adopting the cascaded dual-GRU architecture significantly outperforms all of the other variants, which indicates that the presented architecture is effective.

Will the adopted GRU operator be useful for stereo matching? To further analyze the effectiveness of the GRU operator, we conduct an ablation study by replacing each GRU with three convolution layers and designing the cascaded dual-convs variant. The cascaded dual-convs variant still adopts the same cascaded architecture and the same recurrently updating procedure as the origin cascaded dual-GRU architecture. As shown in Table 3, the cascaded dual-GRU architecture achieves better performance than the variant without GRU, which further demonstrates the effectiveness of the GRU operator and our proposed network.

6. Conclusions

In this research, we propose an end-to-end network for polarimetric stereo depth estimation. The multi-domain visual similarity is extracted by two correlation volume pyramids in our network. We integrate the iso-depth constraint to the network carefully to incorporate the geometric prior encoded in polarization and deal with the inherent ambiguities. To fuse the visual similarity and the iso-depth cost and balance the hybrid input of GRU simultaneously, we present a novel cascaded dual-GRU architecture. In experiments, both synthetic and real datasets are presented and applied for the evaluation of different methods. We show that our approach achieves competitive performance compared to other state-of-the-art methods.

Acknowledgments. This work was partially supported by NSF of China (No. 62102356).

References

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

[2] Gary A Atkinson and Edwin R Hancock. Surface reconstruction using polarization and photometric stereo. In *Proceedings of the International Conference on Computer Analysis of Images and Patterns*, pages 466–473. Springer, 2007.

[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 *Proceedings of the European Conference on Computer Vision*, pages 554–571. Springer, 2020.

[4] Stan Birchfield and Carlo Tomasi. Depth discontinuities by pixel-to-pixel stereo. *International Journal of Computer Vision*, 35(3):269–293, 1999.

[5] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

[6] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. *Advances in Neural Information Processing Systems*, 33:22158–22169, 2020.

[7] A. E. Collett. Field guide to polarization. In *SPIE*, 2005.

[8] Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, and Jan Kautz. Polarimetric multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1558–1567, 2017.

[9] Akshat Dave, Yongyi Zhao, and Ashok Veeraraghavan. Pandora: Polarization-aided neural decomposition of radiance. *arXiv preprint arXiv:2203.13458*, 2022.

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

[11] Yuqi Ding, Yu Ji, Mingyuan Zhou, Sing Bing Kang, and Jinwei Ye. Polarimetric helmholtz stereopsis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5037–5046, 2021.

[12] Carlos Hernandez Esteban, George Vogiatzis, and Roberto Cipolla. Multiview photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):548–554, 2008.

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

[14] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020.

[15] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019.

[16] Heiko Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 807–814. IEEE, 2005.

[17] Heiko Hirschmüller, Peter R Innocent, and Jon Garibaldi. Real-time correlation-based stereo vision with reduced border errors. *International Journal of Computer Vision*, 47(1):229–246, 2002.

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

[19] Tomoki Ichikawa, Matthew Purri, Ryo Kawahara, Shohei Nobuhara, Kristin Dana, and Ko Nishino. Shape from sky: Polarimetric normal recovery under the sky. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14832–14841, 2021.

[20] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 66–75, 2017.

[21] Andreas Klaus, Mario Sormann, and Konrad Karner. Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure. In *Proceedings of the IEEE International Conference on Pattern Recognition*, volume 3, pages 15–18. IEEE, 2006.

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

[23] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *Proceedings of the IEEE International Conference on 3D Vision*, pages 218–227. IEEE, 2021.

[24] Biyang Liu, Huimin Yu, and Yangqi Long. Local similarity pattern and cost self-reassembling for deep stereo matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1647–1655, 2022.

[25] Fotios Logothetis, Roberto Mecca, Fiorella Sgallari, and Roberto Cipolla. A differential approach to shape from polarisation: A level-set characterisation. *International Journal of Computer Vision*, 127:1680–1693, 2019.

[26] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016.

[27] Xing Mei, Xun Sun, Weiming Dong, Haitao Wang, and Xiaopeng Zhang. Segment-tree based cost aggregation for stereo matching. In *Proceedings of the IEEE/CVF Confer-*

ence on Computer Vision and Pattern Recognition, pages 313–320, 2013.

[28] Daisuke Miyazaki, Takuya Shigetomi, Masashi Baba, Ryo Furukawa, Shinsaku Hiura, and Naoki Asada. Polarization-based surface normal estimation of black specular objects from multiple viewpoints. In *Proceedings of the International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, pages 104–111. IEEE, 2012.

[29] Olivier Morel, Fabrice Meriaudeau, Christophe Stolz, and Patrick Gorria. Polarization imaging applied to 3d reconstruction of specular metallic surfaces. In *Machine Vision Applications in Industrial Inspection XIII*, volume 5679, pages 178–186. International Society for Optics and Photonics, 2005.

[30] Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. Efficiently combining positions and normals for precise 3d geometry. *ACM Transactions on Graphics*, 24(3):536–543, 2005.

[31] Geoffrey Oxholm and Ko Nishino. Multiview shape and reflectance from natural illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2155–2162, 2014.

[32] Jaesik Park, Sudipta N Sinha, Yasuyuki Matsushita, Yu-Wing Tai, and In So Kweon. Multiview photometric stereo using planar mesh parameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1161–1168, 2013.

[33] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1):7–42, 2002.

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

[35] Jian Sun, Heung-Yeung Shum, and Nan-Ning Zheng. Shading-aware multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, 2016.

[36] Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):787–800, 2003.

[37] Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14362–14372, 2021.

[38] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*, pages 402–419. Springer, 2020.

[39] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. *arXiv preprint arXiv:1912.09678*, 2019.

[40] Chenglei Wu, Bennett Wilburn, Yasuyuki Matsushita, and Christian Theobalt. High-quality shape from multi-view stereo and shading under general illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 969–976. IEEE, 2011.

[41] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020.

[42] Jie Yang, Kaichun Mo, Yu-Kun Lai, Leonidas J Guibas, and Lin Gao. Dsm-net: Disentangled structured mesh net for controllable generation of fine geometry. *arXiv preprint arXiv:2008.05440*, 2(3), 2020.

[43] Luwei Yang, Feitong Tan, Ao Li, Zhaopeng Cui, Yasutaka Furukawa, and Ping Tan. Polarimetric dense monocular SLAM. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3857–3866, 2018.

[44] Qingxiong Yang. A non-local cost aggregation method for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1402–1409. IEEE, 2012.

[45] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019.

[46] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. In *Proceedings of the European Conference on Computer Vision*, pages 420–439. Springer, 2020.

[47] Ke Zhang, Jiangbo Lu, and Gauthier Lafruit. Cross-based local stereo matching using orthogonal integral images. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(7):1073–1079, 2009.

[48] Songyan Zhang, Zhicheng Wang, Qiang Wang, Jinshuo Zhang, Gang Wei, and Xiaowen Chu. Ednet: Efficient disparity estimation with cost volume combination and attention-based spatial residual. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5433–5442, 2021.

[49] Jinyu Zhao, Yusuke Monno, and Masatoshi Okutomi. Polarimetric multi-view inverse rendering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[50] Zhenglong Zhou, Zhe Wu, and Ping Tan. Multi-view photometric stereo with spatially varying isotropic materials. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1482–1489, 2013.

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