19th International Conference on Machine Vision Applications (MVA)
Kyoto, Japan, July 26-28, 2025.

Simple Yet Effective Way to Use Polarimetric Information in Stereo Matching

Jinyu Zhao Yusuke Monno Masatoshi Okutomi
Institute of Science Tokyo
Meguro-ku, Tokyo 152-8550, Japan
{jzhao, ymonno}@ok.sc.e.titech.ac.jp, mxo@ctrl.titech.ac.jp

Abstract

Polarimetric information holds great potential in improving stereo matching due to its strong correlation with scene geometry. One limitation of polarimetric stereo matching research is the lack of an evaluation dataset allowing comprehensive assessment of different ways to use polarimetric information. To address this limitation, in this study, we create a synthetic dataset for polarimetric stereo matching using a physically based renderer and systematically investigate effective ways to use polarimetric information for stereo matching. Based on the investigation, we propose a stereo matching network, featuring a simple yet effective way to use polarimetric information based on Stokes parameters inputs and dual encoders. Experimental results demonstrate that our network achieves state-of-the-art performance, especially showing strong robustness to noise.

1 Introduction

Stereo matching, crucial for 3D scene reconstruction, estimates the disparity from stereo image pairs. Traditional methods [1, 2, 3] attempt to find correspondences between stereo images, but struggle to recover texture-less regions. Deep learning methods, such as PSMNet [4], GwcNet [5], and RAFT-Stereo [6] have significantly improved accuracy and robustness through end-to-end training and iterative refinement.

Polarization has attracted attention in recent years in various 3D reconstruction scenarios including stereo matching [7], multi-view stereo [8, 9], SLAM [10], normal estimation [11, 12, 13], and inverse rendering [14], as it captures the orientation and degree of light's oscillation, providing information highly related to scene geometry. Specifically, the angle of polarization (AoP) and the degree of polarization (DoP) of reflected light correlate with the azimuth and zenith angles of an object's surface normal, respectively, allowing the inference of normal information from polarization.

Most learning-based polarimetric 3D reconstruction methods use AoP and DoP images, which are concatenated to the standard RGB image and fed to a single encoder for feature extraction [11, 13]. A recent method [7] applies dual encoders to extract features separately from the RGB image and the polarimetric (AoP and DoP) images. As seen in these methods, AoP and DoP images are most commonly used for feature encoding of polarization, because of their relations to the surface normal. However, AoP is vulnerable to noise, especially in regions with small DoPs. This limits the robustness of polarimetric 3D reconstruction methods. Another limitation is the lack of suitable evaluation datasets allowing comprehensive assessment of different ways to use polarimetric information in each application.

In this paper, we address the task of polarimetric stereo matching. Our main contributions are fourfold: (i) We construct a synthetic polarimetric stereo dataset using Mitsuba2 renderer [15] based on a polarimetric bidirectional reflectance distribution function. This dataset allows comprehensive and systematic investigation of effective ways to use polarimetric information; (ii) Through the investigation using the dataset, we propose a stereo matching network based on a simple yet effective way to encode polarimetric information utilizing Stokes parameters inputs and dual encoders. We demonstrate that a very simple way of using Stokes parameters, instead of AoP and DoP, surprisingly improves the performance and the robustness; (iii) We experimentally demonstrate that our proposed network achieves state-of-the-art performance on our synthetic and existing real-world datasets, especially exhibiting strong robustness to noise; (iv) We make our dataset and source code publicly available at http://www.ok.sc.e.titech.ac.jp/res/Polar3D/.

2 Polarimetric Stereo Dataset Generation

The most representative datasets for polarimetric stereo matching are the ones constructed in [7]. While the authors of [7] provide both synthetic and real-world datasets, there are still some limitations. For instance, the synthetic dataset is generated based on the random assignment of specular- and diffuse-dominant regions using the segmentation mask applied to the standard RGB image. This means that the dataset is not generated based on a physically-based polarimetric rendering. The ground-truth disparity of the real-world dataset is obtained using a depth sensor. This limits the acquisition of ground-truth disparities for low-light and specular regions, resulting in a sparse ground-

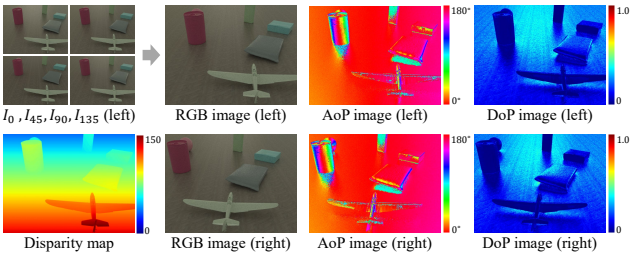


Figure 1: A sample scene of our synthetic polarimetric stereo dataset.

truth disparity map. Since polarization is expected to be effective for such challenging regions, the sparse real-world dataset prevents the evaluation of the full potential of polarization. These limitations motivate us to generate a new synthetic dataset that is based on a physically-based polarimetric rendering and dense ground-truth disparities.

We created a synthetic dataset using Mitsuba2 renderer [15] for physically based rendering. We set the stereo camera, whose focal length is 700.0 pixels with a baseline of 1.0, to render stereo images at 640×480 resolution. We randomly selected 5-7 objects from 40 and 17 objects [16, 17] for training/validation and testing, respectively, without overlapping the objects between the training and the testing. We set a floor under the objects in half of the scenes and a wall behind those objects in the other half of the scenes with a random orientation from $-30°$ to $30°$ to avoid the existing of infinite depth (i.e., zero disparity) in the simulation space. The refractive indexes of the objects were randomly set from 1.4 to 1.6 with roughnesses from 0.0 (completely smooth) to 0.3 (extremely rough) and albedos from 0.0 to 1.0 for each RGB channel to represent typical dielectrics. The light source was randomly set from 10 environment maps [18], including indoor and outdoor.

To render the polarimetric images, we placed a polarizer in front of each camera and rotated them to the directions of 0°, 45°, 90°, and 135° to derive polarimetric information from four directions: I_0, I_{45}, I_{90} and I_{135}, as in [9, 14]. The Stokes vectors representing the polarization state of the light can be expressed as Eq. 1, excluding the component of circular polarization which has little influence in our scene setting.

$$\mathbf{s} = \begin{bmatrix} s_0 \\ s_1 \\ s_2 \end{bmatrix} = \begin{bmatrix} I_0 + I_{90} \\ I_0 - I_{90} \\ I_{45} - I_{135} \end{bmatrix}. \tag{1}$$

The AoP (ϕ) and the DoP (ρ) can be calculated as

$$\phi = \frac{1}{2}\tan^{-1}\frac{s_2}{s_1}, \quad \rho = \frac{\sqrt{s_1^2 + s_2^2}}{s_0}. \tag{2}$$

The dataset includes 900, 100, and 100 scenes for training, validation, and testing, respectively. Our dataset consists of the ground-truth disparity map, the four-directional polarimetric images (RGB), the intensity (s_0) image (RGB), the s_1, s_2, AoP, and DoP images calculated according to Eq. 1 and Eq. 2, for which we used the average intensity of RGB channels. A sample scene of our synthetic polarimetric stereo dataset is shown in Fig. 1.

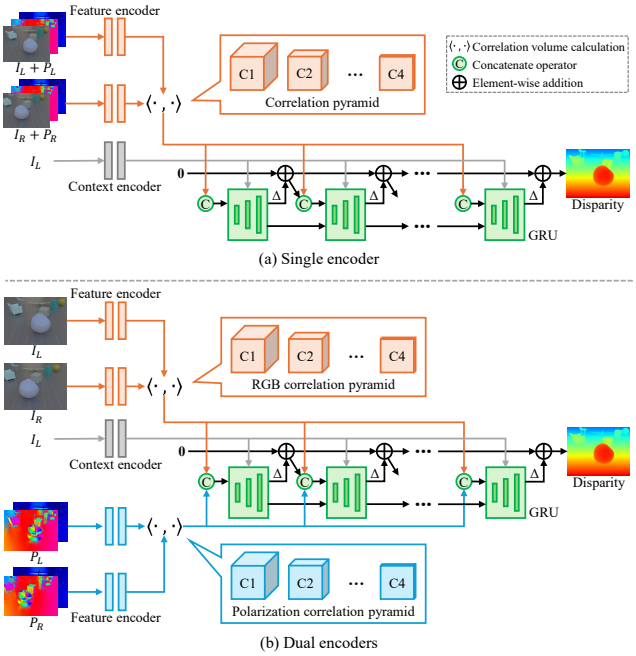


Figure 2: The architectures adopting different feature encoders: (a) A single encoder processing RGB and polarimetric information jointly; (b) Dual encoders handling RGB and polarimetric information separately.

3 Network Architectures

Figure 2 shows the overview of network architectures investigated in this paper. The architectures are based on RAFT-Stereo [6] and composed of three steps: (i) Feature extraction from input rectified stereo images (RGB and polarimetric images), (ii) construction of the 3D correlation volume for all pairs of pixels in the horizontal direction, and (iii) a gated recurrent unit (GRU)-based update operator to perform updates on the disparity map.

Feature extraction: We integrate polarization into stereo matching, adopting the architectures shown in Fig. 2, which includes two types of feature encoders:

(a) Single encoder: A single encoder processes all channels containing RGB information (I_L, I_R) and polarimetric information (P_L, P_R) jointly, potentially capturing cross-modal correlations efficiently. We combine RGB information and the following two types of polarimetric information together as: (i) RGB+$\phi + \rho$ and (ii) RGB+$s_1 + s_2$.

(b) Dual encoders: Specialized features of RGB information (I_L, I_R) and polarimetric information (P_L, P_R) can be extracted from independent encoders, whose potential is demonstrated in DPS-Net [7]. In addition to the original RGB encoder, we input the following two types of polarimetric information into another independent encoder: (i) RGB; $\phi + \rho$ and (ii) RGB; $s_1 + s_2$.

Correlation volume calculation: We calculate 3D correlation volume like RAFT-Stereo [6] from feature maps $\mathbf{f}, \mathbf{g} \in \mathbb{R}^{H \times W \times W}$ extracted from input images using Eq. 3.

$$\mathbf{C}_{ijk} = \sum_h \mathbf{f}_{ijh} \cdot \mathbf{g}_{ikh}, \quad \mathbf{C} \in \mathbb{R}^{H \times W \times W}, \tag{3}$$

where $\mathbf{C}_{ijk}$ denotes the correlation volume, and $\{\mathbf{f}_{ijh}, \mathbf{g}_{ikh}\}$ represent h-th dimension of stereo features that belong to the RGB or the polarization domain.

GRU-based update operator: We adopt the same strategy in RAFT-Stereo [6] to predict a series of disparity fields. During each iteration, we use the current estimate of disparity to index the correlation volume, producing a set of correlation features. These features are concatenated with the disparity and context features and then injected into the GRU which updates the hidden state for disparity update prediction. We train the network by minimizing the L_1 difference between the predicted disparity $\mathbf{d}_i$ and the ground truth disparity $\mathbf{d}_{gt}$ with exponentially increasing weights with N-times iterations:

$$\mathcal{L} = \sum_{i=1}^{N} \gamma^{N-i} ||\mathbf{d}_{gt} - \mathbf{d}_i||_1, \quad \text{where } \gamma = 0.9. \tag{4}$$

4 Experimental Results

4.1 Implementation Details

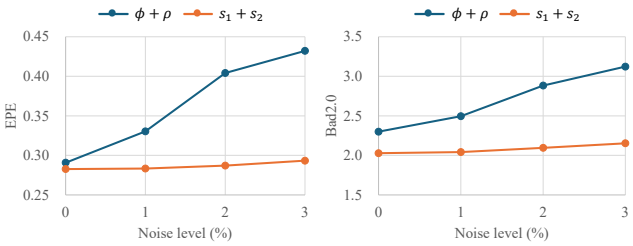
We trained the network models on an NVIDIA RTX 4090 GPU for 1000 epochs on both our synthetic and existing real-world datasets. We cropped the images to the resolution of 400×600 (to cover more vertical information in our settings) and 320×720 (the same setting as DPS-Net [7]) for the synthetic dataset and the real dataset, respectively. We set the initial learning rate to 0.0002, and use AdamW optimizer during the training which is implemented at a batch size of 8.

4.2 Comparison of Encoder Configurations

We evaluate the performance of different encoder configurations using our synthetic dataset and standard stereo matching metrics: end-point error (EPE) and the percentage of pixels with a disparity error greater than 2 pixels (Bad2.0). Table 1 summarizes the EPE and Bad2.0 results. It is clear that polarimetric information is effective in stereo matching, since the results with polarimetric input perform better than the ones with RGB-only input. From the results of a single encoder and dual encoders with polarimetric information, we can see that separating RGB and polarimetric information for feature encoding by dual encoders is a better way than merging them into the same encoder, as better EPE and Bad2.0 results are achieved for both representations of polarimetric input (i.e., ϕ+ρ and s_1+s_2).

Table 1: Results for different encoder configurations.

	Input	EPE	Bad2.0
Single encoder	RGB	0.326	2.62
	RGB+ϕ+ρ	0.318	2.59
	RGB+s_1+s_2	0.313	2.55
Dual encoders	RGB; ϕ+ρ	0.291	2.30
	RGB; s_1+s_2	**0.283**	**2.03**



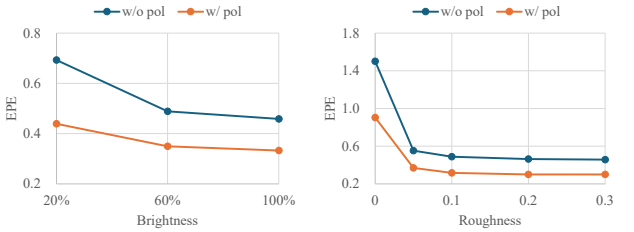
(a) EPE results (b) Bad2.0 results

Figure 3: The results with different noise levels: (a) EPE results; (b) Bad2.0 results.

To further showcase the benefit of introducing $s_1 + s_2$ compared to traditional $\phi + \rho$, we added Gaussian noises to four-directional polarimetric images with standard deviations of $1\%, 2\%$ and 3% of the mean intensity to simulate the sensor noise at the test time. The EPE and Bad2.0 results with different noise levels are shown in Fig. 3, where dual encoders are used. The results highlight the effectiveness of $s_1 + s_2$, which is much robuster to noise than $\phi + \rho$. Given those results, our final proposal is to use the $s_1 + s_2$ input with the dual encoders, which is further evaluated in the later experiments.

4.3 Strengths in Dim and Specular Scenes

Polarimetric information is generally considered to be efficient in dim scenes without enough photometric information (due to dark lighting or albedo) and in areas where strong specular reflections exist. To evaluate polarization's strength over RGB information in stereo matching, we create an additional test dataset with different lighting intensities (20%, 60%, and 100% of the general one) to discuss the performance for introducing polarization. We assign the refractive indexes of all objects to 1.5 to avoid the influence from varying levels of polarization, and the roughnesses to 0.0, 0.05,



(a) EPE results of different brightnesses (b) EPE results of different roughnesses

Figure 4: Comparison of EPE results of different settings: (a) Different brightnesses; (b) Different roughnesses.

0.1, 0.2 and 0.3, respectively, to control the strength of specular reflection (stronger with lower roughness).

Figure 4 shows the EPE results of different brightnesses and roughnesses settings, demonstrating the efficiency of polarization in dim scenes with low brightness settings, as the increase of error is suppressed when the scene becomes darker. In addition, in strong specular scenes with low smoothness settings, polarization is demonstrated to be useful in reducing the influence of specular reflection, and the effect becomes more obvious as the specular reflection becomes stronger.

4.4 Comparison with Existing Methods

We compare our proposed method with the baseline, RAFT-Stereo [6], and the state-of-the-art polarimetric method, DPS-Net [7], using our synthetic dataset without and with Gaussian noise (a standard deviation of 2.0% of the mean intensity) and the real-world dataset provided in [7].

Table 2 summarizes the EPE and Bad2.0 results. The results for the noise-free synthetic dataset and the real-world dataset show that polarimetric methods (DPS-Net and ours) achieve better accuracy in both EPE and Bad2.0 than the RGB-only method (RAFT-Stereo). The results using the synthetic dataset with and without Gaussian noise also demonstrate the robustness of our method against noise over DPS-Net. Figure 5 shows comparisons of the estimated disparity results visually, including regions without ground-truth disparities that are difficult to evaluate numerically. Polarimetric methods perform better especially in estimating the continuous surfaces of the computer case, the bag and the back of the chair, and the flat surface of the display stand with consistent disparity in the horizontal direction. Furthermore, our method derives more accurate results using a simpler network without an additional GRU for polarimetric cost calculation compared to DPS-Net, with around 50%'s improvement in calculation speed (e.g., DPS-Net estimates at the speed of 8.37 fps while our method reaches 12.59 fps using the real-world dataset for the test), which demonstrates the efficiency of our method to handle polarimetric information.

Table 2: Comparisons of the EPE and Bad2.0 results.

	Synthetic		Synthetic (noisy)		Real-world	
	EPE	Bad2.0	EPE	Bad2.0	EPE	Bad2.0
RAFT-Stereo	0.326	2.62	0.328	2.64	0.708	4.18
DPS-Net	0.288	2.10	0.395	2.77	0.667	3.79
Proposed	**0.283**	**2.03**	**0.287**	**2.10**	**0.631**	**3.74**

5 Conclusion

In this paper, we have proposed a simple yet effective polarimetric stereo matching method. We have explored better integration of polarimetric information to extend RAFT-Stereo [6] using a newly generated synthetic dataset to enable comprehensive evaluation of polarization in stereo matching, which remains underexplored in previous studies. The effectiveness of using polarization in the form of Stokes vectors is demonstrated compared to the commonly used AoP and DoP information, and stronger robustness against noise is further proved. We believe that this study provides actionable design insights for robust polarimetric stereo matching in adverse environments.

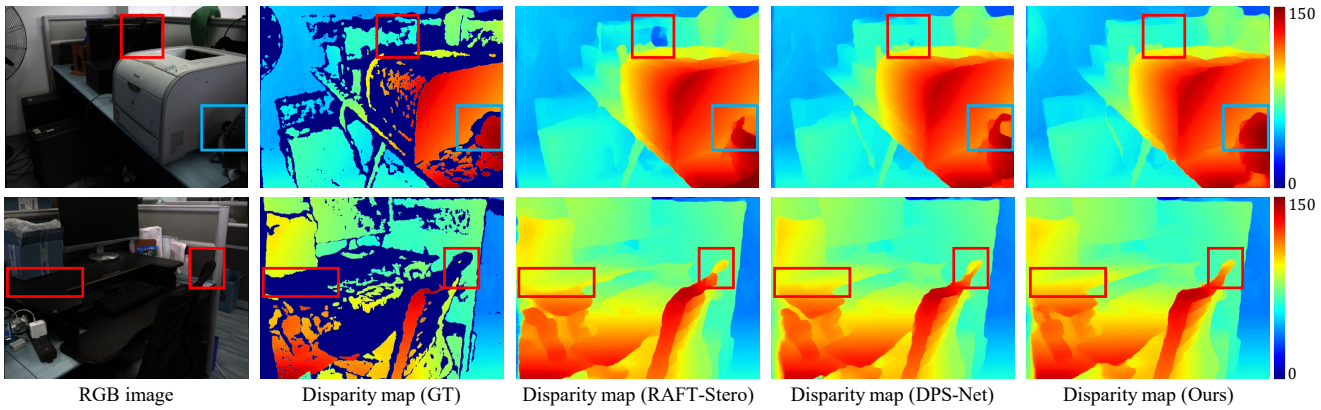


Figure 5: Comparisons of existing methods and our proposed method.

References

[1] Jian Sun, Nan-Ning Zheng, and Heung-Yeunget Shum: "Stereo matching using belief propagation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.25, no.7, pp.787–800, 2003.

[2] Heiko Hirschmuller: "Accurate and efficient stereo processing by semi-global matching and mutual information," In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.807–814, 2005.

[3] Qingxiong Yang: "A non-local cost aggregation method for stereo matching," In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.1402–1409, 2012.

[4] Jia-Ren Chang and Yong-Sheng Chen: "Pyramid stereo matching network," In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.5410–5418, 2018.

[5] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li: "Group-wise correlation stereo network," In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.3273–3282, 2019.

[6] Lahav Lipson, Zachary Teed, and Jia Deng: "RAFT-stereo: Multilevel recurrent field transforms for stereo matching," In *Int. Conf. on 3D Vision (3DV)*, pp.218–227, 2021.

[7] Chaoran Tian, Weihong Pan, Zimo Wang, Mao Mao, Guofeng Zhang, Hujun Bao, Ping Tan, and Zhaopeng Cui: "DPS-Net: Deep polarimetric stereo depth estimation," In *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, pp.3569–3579, 2023.

[8] Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, and Jan, Kautz: "Polarimetric multi-view stereo," In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.1558–1567, 2017.

[9] Jinyu Zhao, Jumpei Oishi, Yusuke Monno, and Masatoshi Okutomi: "Polarimetric patchmatch multi-view stereo," In *Proc. of IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pp.3476–3484, 2024.

[10] Luwei Yang, Feitong Tan, Ao Li, Zhaopeng Cui, Yasutaka Furukawa, and Ping Tan: "Polarimetric dense monocular SLAM," In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.3857–3866, 2018.

[11] Yunhao Ba, Alex Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi: "Deep shape from polarization," In *Proc. of European Conf. on Computer Vision (ECCV)*, pp.554–571, 2020.

[12] Yoshiki Fukao, Ryo Kawahara, Shohei Nobuhara, and Ko Nishino: "Polarimetric normal stereo," In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.682–690, 2021.

[13] Chenyang Lei, Chenyang Qi, Jiaxin Xie, Na Fan, Vladlen Koltun, and Qifeng Chen: "Shape from polarization for complex scenes in the wild," In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp.12632–12641, 2022.

[14] Jinyu Zhao, Yusuke Monno, and Masatoshi Okutomi: "Polarimetric multi-view inverse rendering," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.45, no.7, pp.8798–8812, 2022.

[15] Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob: "Mitsuba 2: A retargetable forward and inverse renderer," *ACM Trans. on Graphics (TOG)*, vol.38, no.6, pp.1–17, 2019.

[16] YCB Benchmarks: `https://www.ycbbenchmarks.com/`

[17] Artec 3D: `https://www.artec3d.com/3d-models/`

[18] Bernhard Vogl: `https://dativ.at/lightprobes/`