DPS-Net: Deep Polarimetric Stereo Depth Estimation -Supplementary Material-

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

In this supplementary material, we first introduce more details of the proposed cascaded dual-GRU architecture (Sec. A), the polarization ambiguities (Sec. B), the iso-depth cost (Sec. C), the imaging system (Sec. D) and the synthetic data (Sec. E). Moreover, we provide more qualitative results of stereo depth estimation (Sec. F), polarimetric normal estimation (Sec. G), and additional ablation experiments on different iteration numbers of GRU (Sec. H). Finally, we discuss the limitations of our work in Sec. I.

A. Detail of Cascaded Dual-GRU Architecture

As described in Sec. 4.3 of our main paper, we propose a cascaded dual-GRU architecture to fuse the iso-depth constraint and the multi-domain correlation features. The cascaded dual-GRU architecture is composed of a regression block and an optimization block, where a similar encoder-decoder scheme is employed.

In the regression block, the actual correlation features and the actual disparity are encoded by different encoders, then concatenated with the contextual features to form the input of the regression GRU. In the hybrid GRU module, the hidden state is updated at each iteration as follows,

$$\begin{aligned} z_t &= \sigma(Conv_{3\times3}([h_{t-1}, x_t], W_z)), \\ r_t &= \sigma(Conv_{3\times3}([h_{t-1}, x_t], W_r)), \\ \tilde{h}_t &= tanh(Conv_{3\times3}([r_t \odot h_{t-1}, x_t], W_h)), \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t, \end{aligned} \tag{1}$$

where the hidden state and the multi-domain input are noted as h_t and x_t. Finally, the disparity is decoded by the Head Net. Specifically, both the encoders and the Head-Net consist of a two-layers convolution. The detail of the architecture of the regression block is shown in Fig. A.

In the optimization block, similar to the regression block, the virtual correlation features, along with the iso-depth cost, the actual disparity, and the virtual disparity, are processed by different encoders. The hidden state is recurrently updated by the GRU as Eq. 1 and decoded to generate the increment of the disparity and the step length. The detail of the architecture of the optimization block is shown in Fig. B.

*Corresponding author.

B. Details of Polarization Ambiguities

Normally, for each pixel of the captured image, the diffuse reflection or the specular reflection dominates. Moreover, the DoLP and AoLP satisfy different formulas for the different reflections. For the diffuse reflection, the relationship between the polarization and the surface normal is as follows:

$$\rho_d = \frac{(\eta - 1/\eta)^2 \sin^2\theta}{2 + 2\eta^2 - (\eta + 1/\eta)^2 \sin^2\theta + 4\cos\theta\sqrt{\eta^2 - \sin^2\theta}}, \tag{2}$$

$$\phi_d = \varphi \text{ or } \phi_d = \varphi + \pi, \tag{3}$$

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

$$\rho_s = \frac{2\sin^2\theta\cos\theta\sqrt{\eta^2 - \sin^2\theta}}{\eta^2 - \sin^2\theta - \eta^2\sin^2\theta + 2\sin^4\theta}, \tag{4}$$

$$\phi_s = \varphi \pm \frac{\pi}{2}. \tag{5}$$

Surface normal can be estimated from DoLP and AoLP by solving the above equations. However, it is prone to be proved that the ambiguity of the normal is introduced in the solving process caused by the unknown reflection and the multi-solution of the nonlinear equation. The polarization can be further classified into azimuth angle ambiguity and zenith angle ambiguity. By solving Eq. 2 and Eq. 4, we can get four possible solutions for the specular case and two possible solutions for the diffuse case. In our method, we bypass the zenith angle due to the unknown refractive index. As for the azimuth ambiguity, we can get a total of four possible solutions from Eq. 2 and Eq. 4. For the sake of distinction, the azimuth ambiguity with a $\pi/2$ shift caused by the different reflection is called the $\pi/2$-ambiguity. Concretely, the azimuth vector is parallel to the AoLP vector

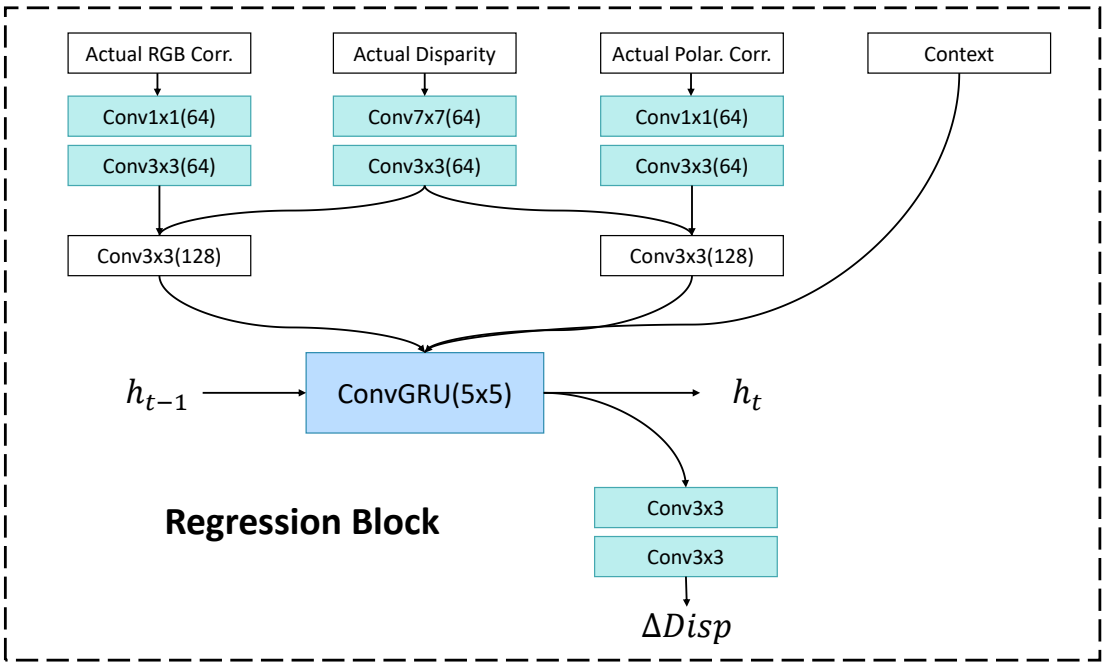

Figure A. The detailed architecture of the regression block in our network. The multi-domain correlation feature is encoded first, and the disparity increment is decoded from the recurrently updated hidden state.

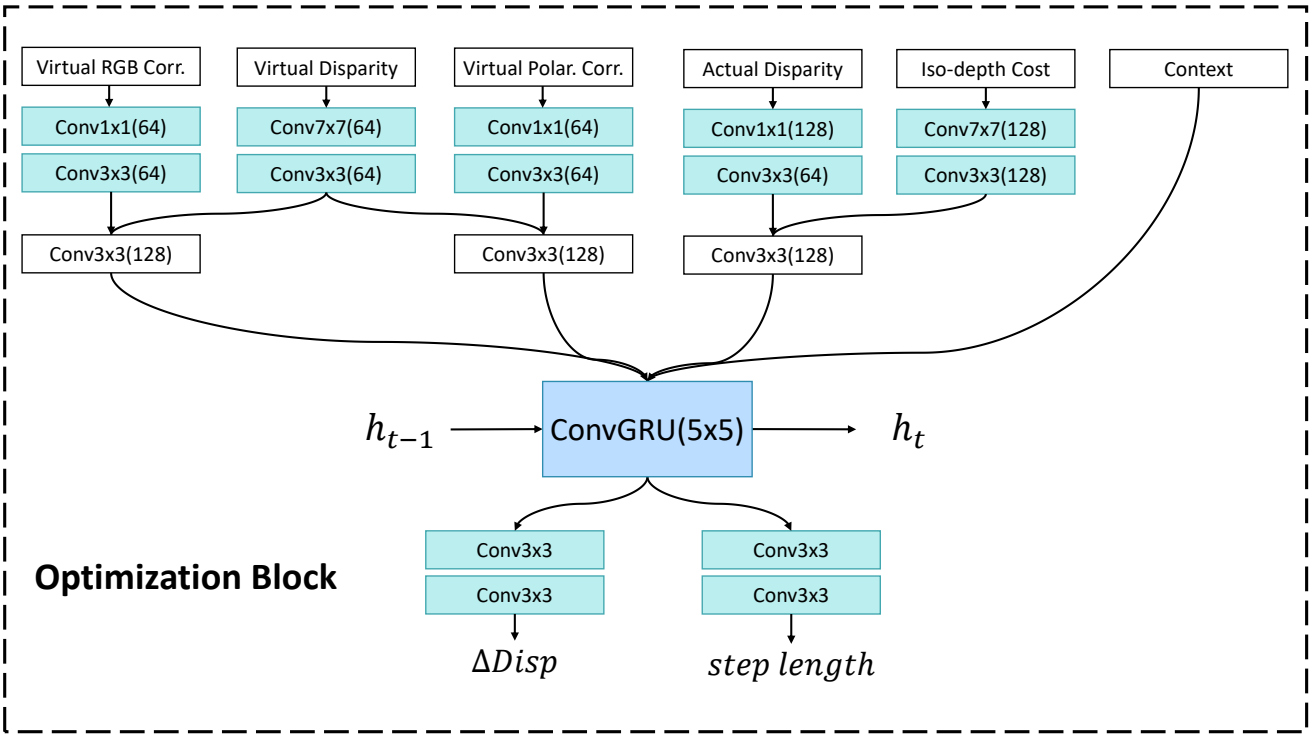

Figure B. The detailed architecture of the optimization block in our network. Both the encoded virtual correlation features and the encoded iso-depth cost are injected into the GRU. The optimization GRU produces the disparity increment and the step length to optimize and rectify the disparity.

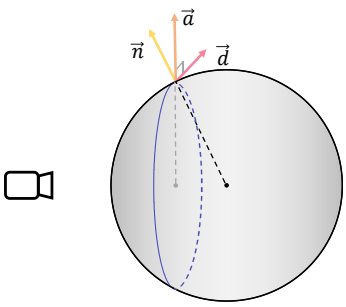

Figure C. Illustration of the geometric relationship between iso-depth counter and azimuth vector. The normal vector $\vec{n}$ is orthogonal to the iso-depth counter. The azimuth vector $\vec{a}$ is coplanar with the iso-depth counter and orthogonal to the iso-depth counter.

for the diffusion reflection case, while the azimuth vector is orthogonal to the AoLP vector for the specular reflection case. Additionally, it can be seen that there is azimuth angle ambiguity for the same reflection in Eq. 2 or Eq. 4, which is noted as the π-ambiguity.

C. Details of Iso-depth Cost Derivation

In this section, we elaborate on the derivation of iso-depth cost as described in Eq. 6 of the main paper.

The normal can be decomposed as the azimuth angle φ and the zenith angle θ:

$$\mathbf{n} = \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix} = \begin{bmatrix} \cos\varphi \sin\theta \\ \sin\varphi \sin\theta \\ \cos\theta \end{bmatrix}. \tag{6}$$

We can represent the azimuth angle by the depth as:

$$\tan(\varphi) = \frac{D(x, y+1) - D(x, y-1)}{D(x+1, y) - D(x-1, y)}, \tag{7}$$

where x, y denotes the 3D coordinates in the camera frame. As illustrated in Fig. C, the iso-depth contour is orthogonal to the azimuth vector, which can also be concluded from Eq. 7.

The surface normal $\vec{n}$ can be further expressed by the depth D through the central difference method:

$$\vec{n} = (P_{1,0} - P_{-1,0}) \times (P_{0,1} - P_{0,-1}). \tag{8}$$

It should be noted that the neighbourships of Eq. 8 and Eq. 7 are based on different reference coordinate systems. The Eq. 8 is based on the image coordinate system, where the $\mathrm{P}_{i,j}$ denotes the 3D position of neighbor pixel $(u+i, v+j)$ relative to pixel (u, v). In contrast, the Eq. 7 is based on the 3D coordinate system, where the depth $D(x+i, y+j)$ denotes the depth of the neighbor point relative to point $\{x, y, D(x, y)\}$, which is utilized to illustrate the geometric relationship between the iso-depth contour and the azimuth vectors.

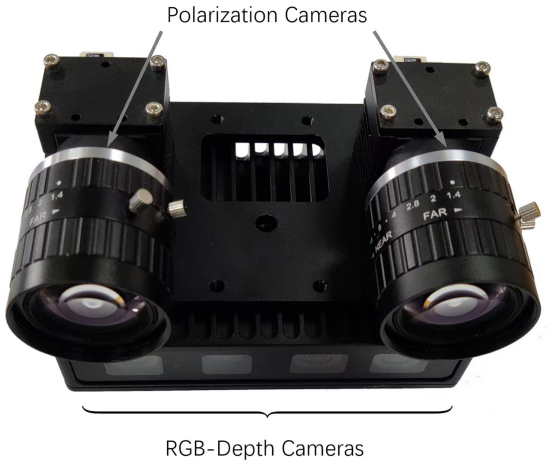

Figure D. The photograph of our imaging system. The stereo polarimetric cameras are fixed on the top plane of the RGB-D camera. The extrinsic parameters between the cameras are calibrated.

By exploiting a suitable change of variables according to the pinhole model, we can represent the azimuth angle as:

$$\tan(\varphi) = \frac{f_y}{f_x} \frac{(D_{0,1} - D_{0,-1})(D_{1,0} + D_{-1,0})}{(D_{1,0} - D_{-1,0})(D_{0,1} + D_{0,-1})}. \tag{9}$$

Substituting the disparity into the iso-depth constraint, we can further get

$$\tan(\varphi) = \frac{f_y}{f_x} \frac{(\mathrm{d}_{0,-1} - \mathrm{d}_{0,1})(\mathrm{d}_{-1,0} + \mathrm{d}_{1,0})}{(\mathrm{d}_{-1,0} - \mathrm{d}_{1,0})(\mathrm{d}_{0,-1} + \mathrm{d}_{0,1})}, \tag{10}$$

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

Finally, recall Sec. 3 in the main paper and Sec. B in supplementary that there are π-ambiguity and $\pi/2$-ambiguity for the azimuth angle. We bypass the π-ambiguity by the cross-product operator and resolve the $\pi/2$-ambiguity by the argmin operator. The iso-depth is formulated as follows,

$$\begin{aligned} \mathbf{C}_s(\varphi) &= \left[sin\left(\phi\right) sin\left(\varphi\right) + cos\left(\phi\right) cos\left(\varphi\right)\right]^2, \\ \mathbf{C}_d(\varphi) &= \left[sin\left(\phi\right) cos\left(\varphi\right) - cos\left(\phi\right) sin\left(\varphi\right)\right]^2, \\ \mathbf{C}(\varphi) &= \min\left\{\mathbf{C}_s(\varphi), \mathbf{C}_d(\varphi)\right\}, \\ \mathbf{R}(\varphi) &= \arg\min\left\{\mathbf{C}_s(\varphi), \mathbf{C}_d(\varphi)\right\}. \end{aligned} \tag{11}$$

D. Imaging System for Data Collection

We devise an efficient imaging system to capture the polarimetric stereo images and the depth images simultaneously. The imaging system is shown in Fig. D. The imaging system consists of two polarization cameras and an RGB-D camera.

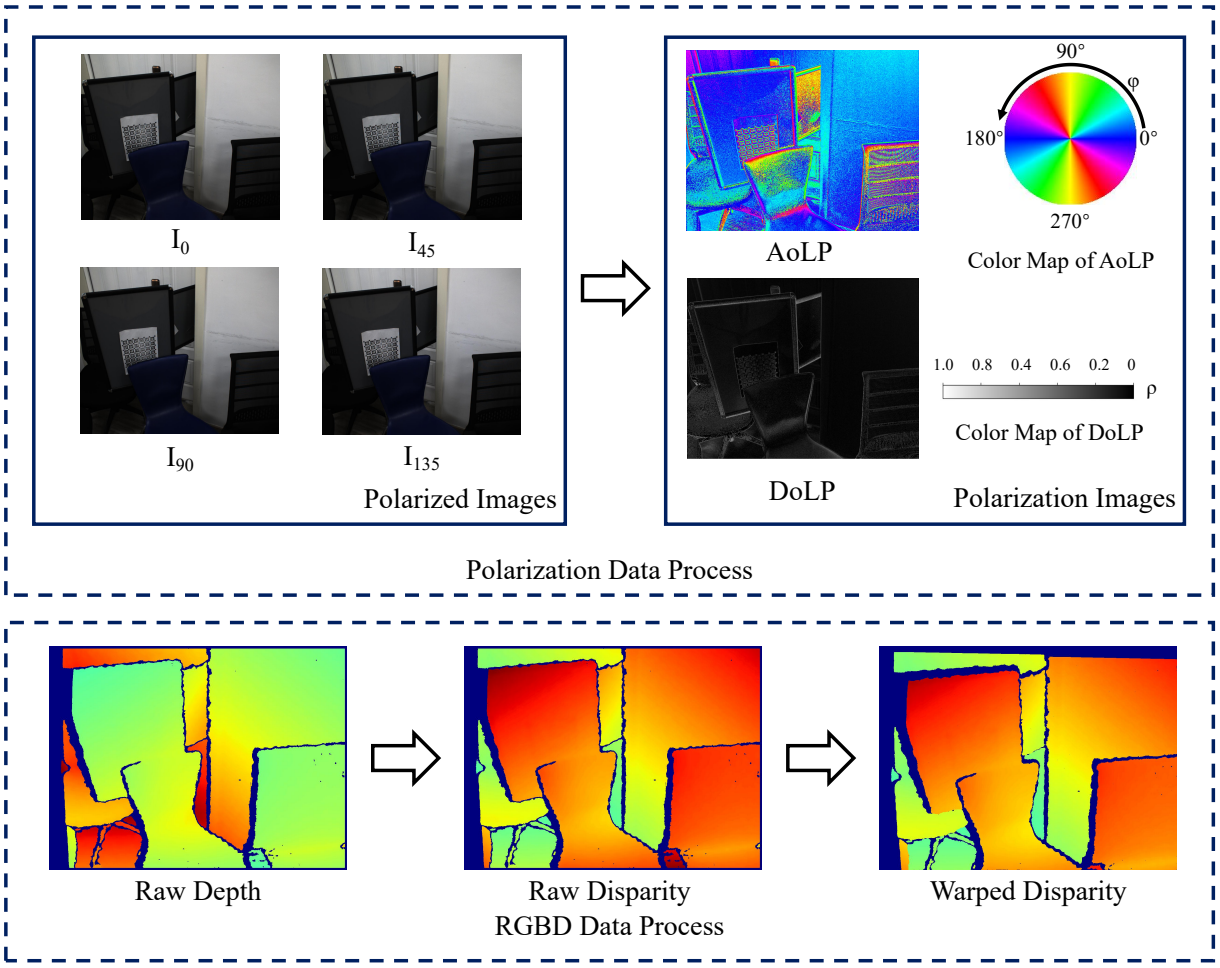

Figure E. The diagram of the real data processing procedure. The DoLP and AoLP are generated from the raw polarized images. The ground truth of the disparity is retrieved from the raw depth images. The color maps of AoLP and DoLP are shown in the top-right corner.

Two Lucid PHX050S-Q polarization cameras are used to capture the stereo polarization images, and a plate with high machining accuracy is utilized to fix the stereo cameras. Each polarization camera can capture four polarization images with different polarizer angles in a single shot. Then the per-pixel DoLP and the AoLP value are calculated as follows,

$$\rho = \frac{\sqrt{(I_0 - I_{90})^2 + (I_{45} - I_{135})^2}}{(I_0 + I_{90})}, \tag{12}$$

$$\phi = \frac{1}{2} arctan(\frac{I_{45} - I_{135}}{I_0 - I_{90}}), \tag{13}$$

where the DoLP and the AoLP are noted as ρ and ϕ, the intensity of four raw polarization images are represented as I_0, I_{45}, I_{90} and I_{135}.

We capture the depth image with an RGB-D camera to generate dense ground truth of the disparity. Compared with LiDAR, the RGB-D camera can provide dense depth, which avoids the uncertain deviation introduced by the depth completion processing. The RGB-D camera can provide an accurate depth image within the valid operating range, which is suitable for our dataset. We calibrate the intrinsic parameters between the polarimetric cameras and the RGB-D camera. The depth image is warped from the RGB-D camera to the left polarimetric camera to align with the left image.

The collecting procedure of real data is illustrated in Fig. E.

E. Additional Details for Data Synthesis

We synthesize the polarimetric data of IPS from the accurate normal map provided by the IRS[11] and take the RGB image of the IRS dataset as the average intensity image in our dataset. We design a procedure for synthesizing the polarimetric data as illustrated in Fig. F.

Firstly, the normal is directly retrieved and converted to the azimuth angle and the zenith angle according to Eq. 6.

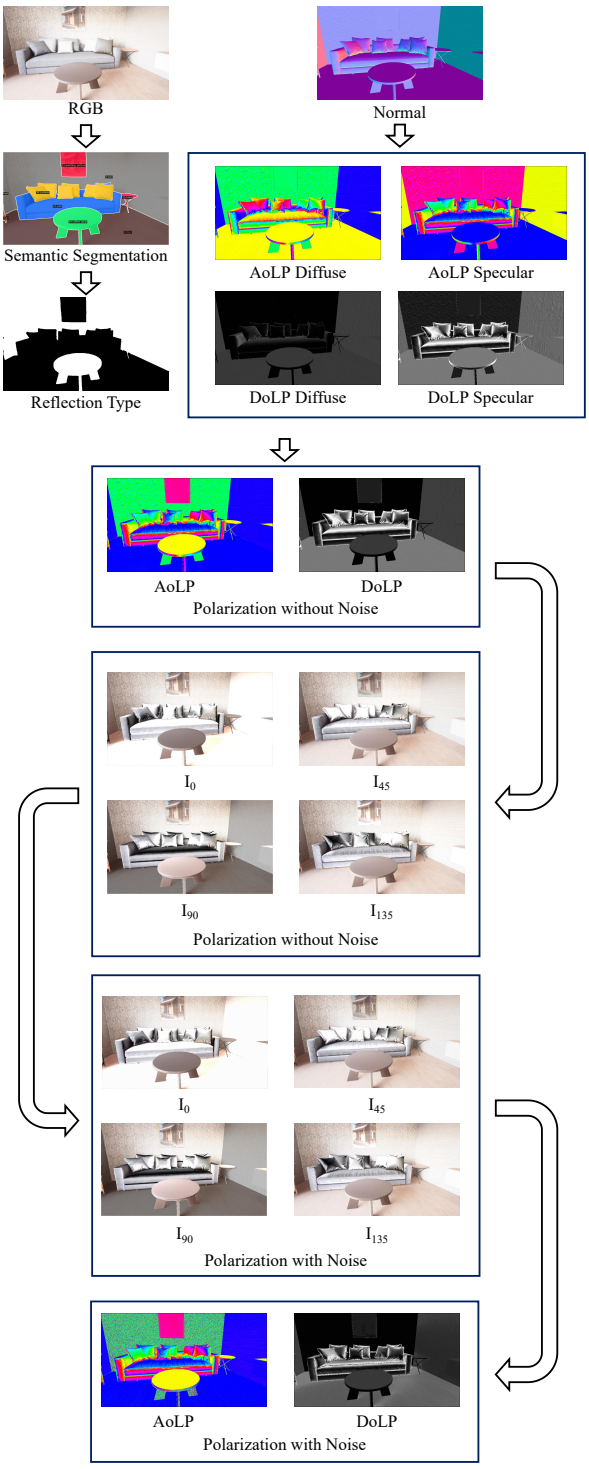

Figure F. The diagram of the data synthesizing procedure. We first generate the DoLP and AoLP of different reflections. Then we produce the polarization image with the mixed reflection according to the segmentation result. Lastly, we add noise to the polarization image.

Given the pixel-wise azimuth angle and zenith angle, we calculate the AoLP image and the DoLP image dominated by the specular reflection or diffuse reflection, respectively.

To further simulate the polarization with the mixed reflection in the natural environment, we segment the instances in the RGB image by a semantic segmentation method in [5]. Different instances are randomly given the label of diffuse reflection or specular reflection, and the pixels belonging to the same instance are labeled by the identical reflection type. Then, we make use of the reflection mask to produce AoLP and DoLP images with mixed reflection types.

In addition, we generate noise to augment the polarization data and better simulate the real data. The AoLP and DoLP are converted to four polarized images with different phase angles according to Eq. 1 of the main paper. We add Gaussian noise into the polarized images and recover the AoLP and DoLP from the noisy polarized images. We add noise to the polarized images instead of directly modifying the AoLP and DoLP to preserve the latent geometric constraints.

It should be noted that we follow most of the existing polarization methods for 3D modeling [1, 2, 6, 10, 12, 14] and generate the synthetic data by modeling per-pixel polarization as either diffuse or specular reflection is dominant. According to the Fresnel Equations, for the mixed reflections, the two components of the polarization are perpendicular or parallel to the reflection plane, and the AoLP of the total polarization appears as the AoLP of the dominant reflection, referring to [2, 12, 13]. As for the DoLP map generated from the dominant reflection model in our IPS dataset, even though it is somewhat simple and may deviate from the real data, it is only exploited to extract contextual features for matching, which has a slight influence on the generalization and performance of our network. Furthermore, after the fine-tuning of the model on the real datasets, the gap between the simulation and reality can be closed, and the polarization feature can be well learned, making our proposed network applicable to the real polarization state.

F. More Qualitative Results of Stereo Depth Estimation and 3D Reconstruction

We present more qualitative results of stereo depth estimation on the IPS and the RPS dataset. We compared our method with state-of-the-art methods, including RAFT-Stereo [7] and LEA-Stereo [8]. As shown in Fig. G and Fig. H, our method has a better performance compared to other methods. Our method can produce more accurate and sharper depth on both the synthesis and the real data. We highlight the challenging regions for depth estimation with bounding boxes in Fig. G and Fig. H.

In addition, we also show the mesh reconstructed by our method, LEA-Stereo[8] and RAFT-Stereo[7] to fur-

ther compare the performance of stereo reconstruction. As shown in Fig. I, our method can generate smoother and more coherent surfaces, while the reconstructed results of LEA-Stereo and RAFT-Stereo are coarser and lack details. The comparison results reveal that our method can capture the 3D geometry well and achieve high-quality reconstruction, which validates the effectiveness of our stereo method.

Besides, we also showcase the ability of our method of recovering high-frequency 3D details. For better visualization, we enlarge the detailed result with the intuitive grayscale colormap. As shown in Fig. J, our results can preserve high-frequency details, especially for the synthetic data.

We propose more comparison results of the traditional polarimetric stereo methods in [14]. We demonstrate the inputs and the results of [14] as well as the results of our method and the ground truth of disparities in Fig. K. Specifically, the input is generated by SGM as mentioned in [4]. In addition, different color bars are related to the disparity ranges of different disparity images.

The method in [14] is designed to recover depth from a polarization and RGB stereo pair, which is similar to our setting. As shown in Fig. K, we can see that both the results of SGM and our method are broadly in line with the ground truth. However, there is still a notable deviation in the disparity range between the ground truth and the results of [14]. The deviation may result from the following reasons. First of all, the approach in [14] is focused on the object-level application, while our method is concerned about the scene-level application. In addition, our data is captured under uncontrolled illumination, including natural sunlight and complicated light source, which may not conform to the Lambert assumption. What's more, the method in [14] assumes the refractive index of the dielectric material is known, which is unknown in our data.

G. Additional Qualitative Results of Polarimetric Normal Estimation

In this section, we compare our method with the recent polarimetric normal estimation methods SPW [6] and DeepSfP [1]. DeepSfP is a learning-based method to implement shape recovery from polarimetric images of a single view. In DeepSfP, a fully convolutional encoder-decoder architecture is adopted to process both the raw polarized images and the ambiguous normal map. In contrast, SPW is a scene-level normal estimation network. SPW employs a multi-head self-attention module and viewing encoding to handle the polarization ambiguities and estimate the normal by a single-view polarization image. We pre-train and fine-tune the SPW and DeepSfP with the same schedule as our method. Specifically, we pre-train the SPW and the DeepSfP for 100 epochs on the IPS dataset and fine-tune them for 150 epochs on the RPS dataset. We utilize the final models of the 150th epoch to generate the normal results for comparison. Additional qualitative results are shown in Fig. L, our method can recover better surface normal.

Iteration		4	6	8	10
RAFT-Stereo	AvgErr	0.852	0.724	0.681	0.672
	Runtime(s)	0.255	0.309	0.352	0.408
Ours	AvgErr	0.658	0.630	0.619	**0.615**
	Runtime(s)	**0.178**	0.218	0.255	0.288

Table A. Comparison to RAFT-Stereo [7] with different iterations. Our method is more efficient and accurate at all iterations.

Method	Separate Backbone	Single Backbone
AvgErr	0.641	0.619
bad 2.0	3.738	3.354

Table B. Comparison to the DPS-Net with additional ablation studies. The ablation results of the separate backbone are listed.

H. Additional Ablation Study of Hybrid GRU-based Update Operator

We evaluate our method and the GRU-based method RAFT-Stereo [7] with different iteration numbers. The results shown in Table A demonstrate that our method can achieve better performance and takes shorter runtime than RAFT-Stereo under different iteration numbers. It can be seen that the accuracy increases with more iterations and saturates beyond a certain number of iterations. Moreover, in Table A, the runtime increases almost linearly with the iteration number while the performance improvement diminishes. Lastly, we choose 8 iterations for the training and inference to balance efficiency and accuracy.

To further analyze our proposed network, we also carry out the ablation on the separate backbones. Concretely, we compare the performance of the networks that adopt separate backbones or the single backbone for the correlation features and the context features extraction. As shown in Table B, utilizing a single backbone can achieve better performance.

I. Limitations

Although the proposed DPS-Net achieves competitive performance for polarimetric stereo depth estimation, there are still several aspects that can be further improved. At first, the noise pattern of our synthetic data may be different from the real data since it is especially difficult to fully simulate the noise caused by mixed reflections under complicated lighting conditions. If more realistic polarization noises are introduced, the generalization of the pre-trained model can be further improved. Moreover, while our method is faster than other state-of-the-art learning-based methods on the real dataset, it still takes about 250ms to infer from the stereo polarization images with a resolution

of 1280x960 on a single NVIDIA 3090 GPU. In order to accommodate real-time application scenarios, our network needs to be further improved. We can optimize the architecture of our network based on the current accelerating method presented in [9, 3]. At last, in the proposed DPS-Net, we do not consider the material information which may benefit the depth estimation due to the lack of reliable material information in our dataset. In our future work, we plan to introduce material cues into our network and consider the consistency relevant to the material to improve the performance of stereo reconstruction and disambiguation.

References

[1] Yunhao Ba, Alex Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. Deep shape from polarization. In *Proceedings of the European Conference on Computer Vision*, pages 554–571. Springer, 2020.

[2] 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.

[3] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*, 2021.

[4] 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.

[5] Jitesh Jain, Anukriti Singh, Nikita Orlov, Zilong Huang, Jiachen Li, Steven Walton, and Humphrey Shi. Semask: Semantically masked transformers for semantic segmentation. *arXiv preprint arXiv:2112.12782*, 2021.

[6] 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.

[7] 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.

[8] 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.

[9] Haotong Qin, Ruihao Gong, Xianglong Liu, Xiao Bai, Jingkuan Song, and Nicu Sebe. Binary neural networks: A survey. *Pattern Recognition*, 105:107281, 2020.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

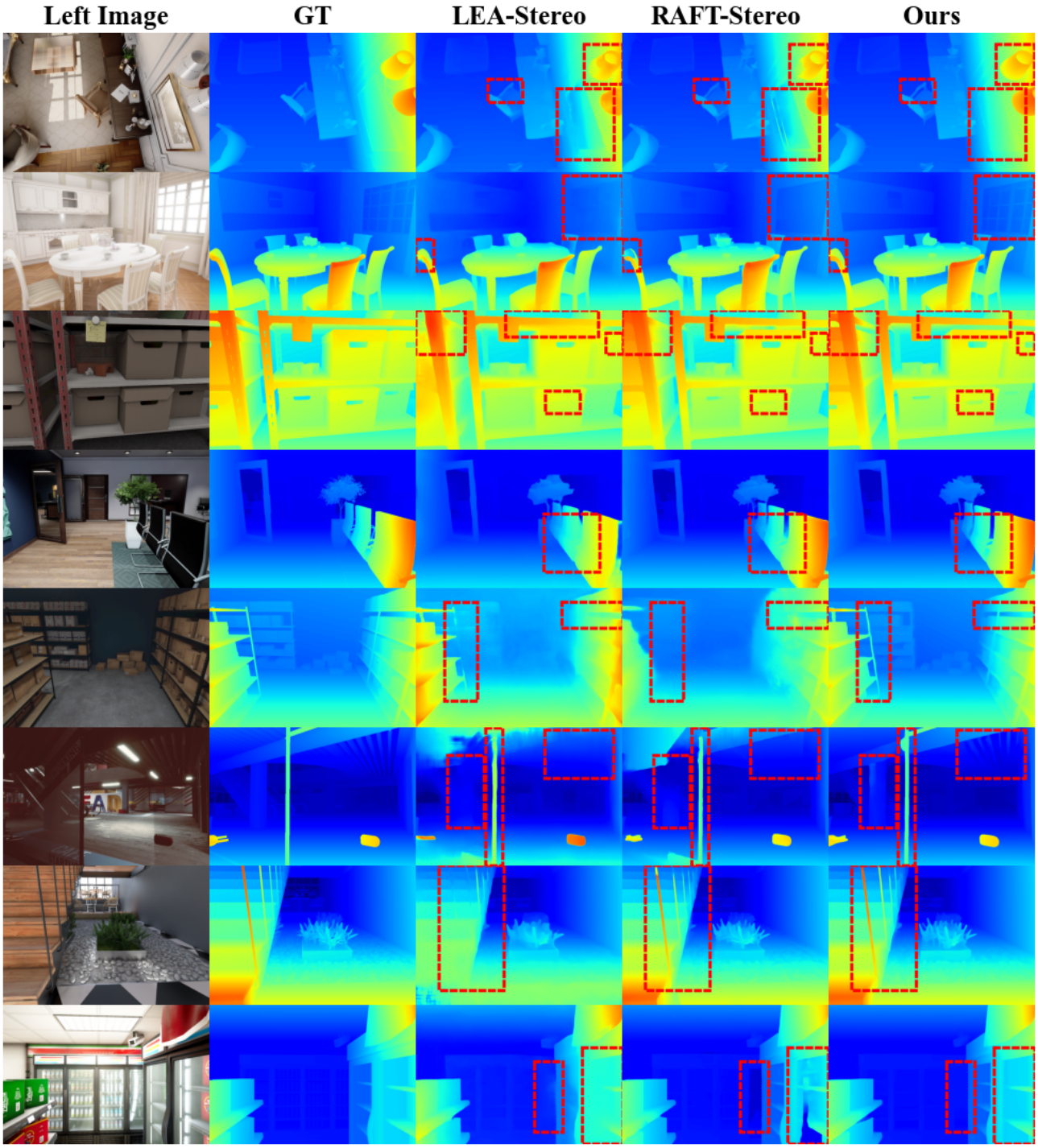

Figure G. Additional qualitative results on the IPS dataset. We display the comparison results of our method, LEA-Stereo[8], and RAFT-Stereo[7]. The bounding box visualizes the challenging region, including the thin structure, featureless area, and the boundary of the objects.

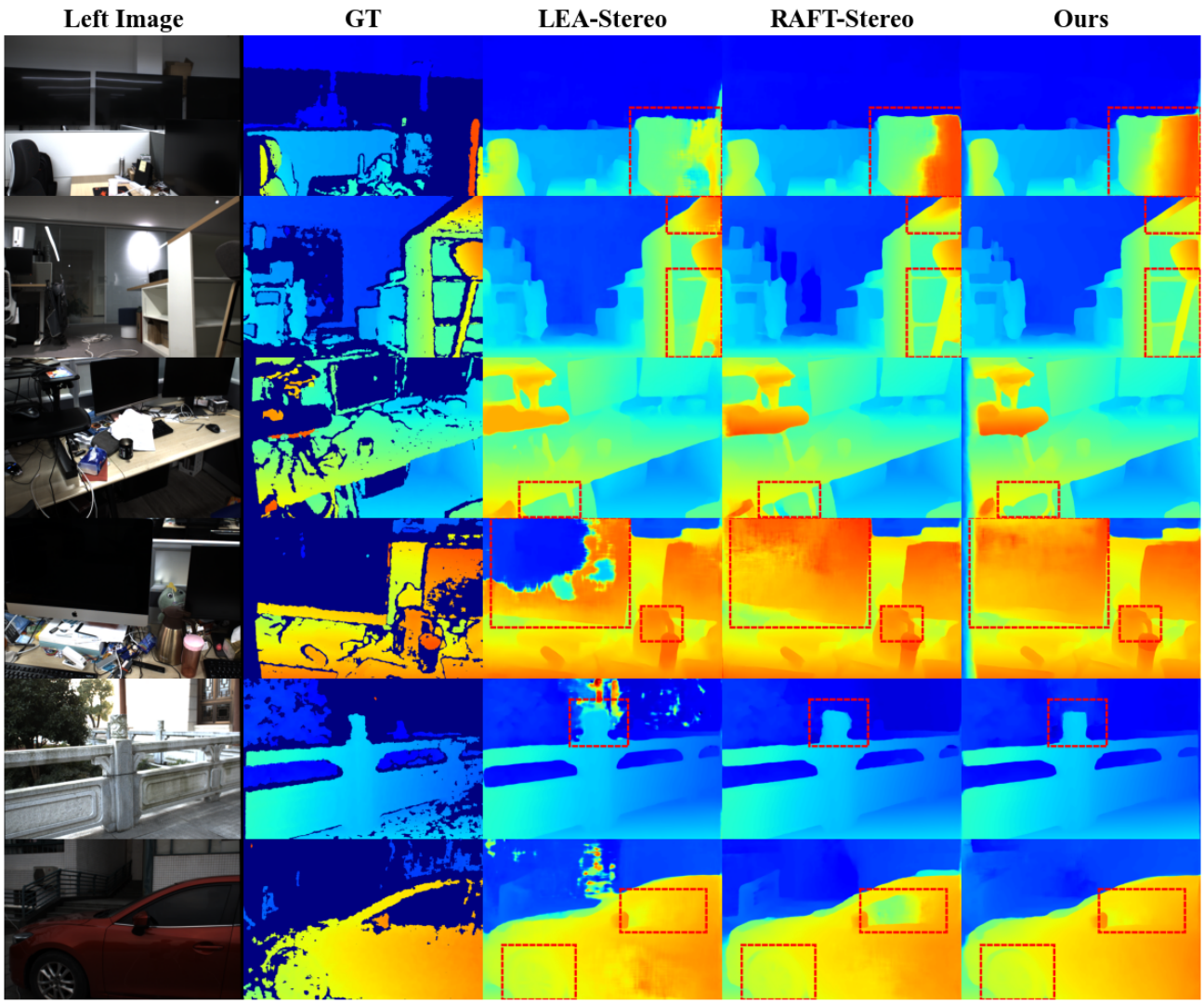

Figure H. Additional qualitative results on the RPS dataset. Our DPS-Net generates sharper object boundaries and is robust to various illumination.

Figure I. Additional stereo reconstruction results on the RPS dataset. We adopt a different observation view from the capturing view to display the qualitative results and evaluate the reconstruction performances. Our DPS-Net can capture the geometry well and reconstruct surfaces with higher quality.

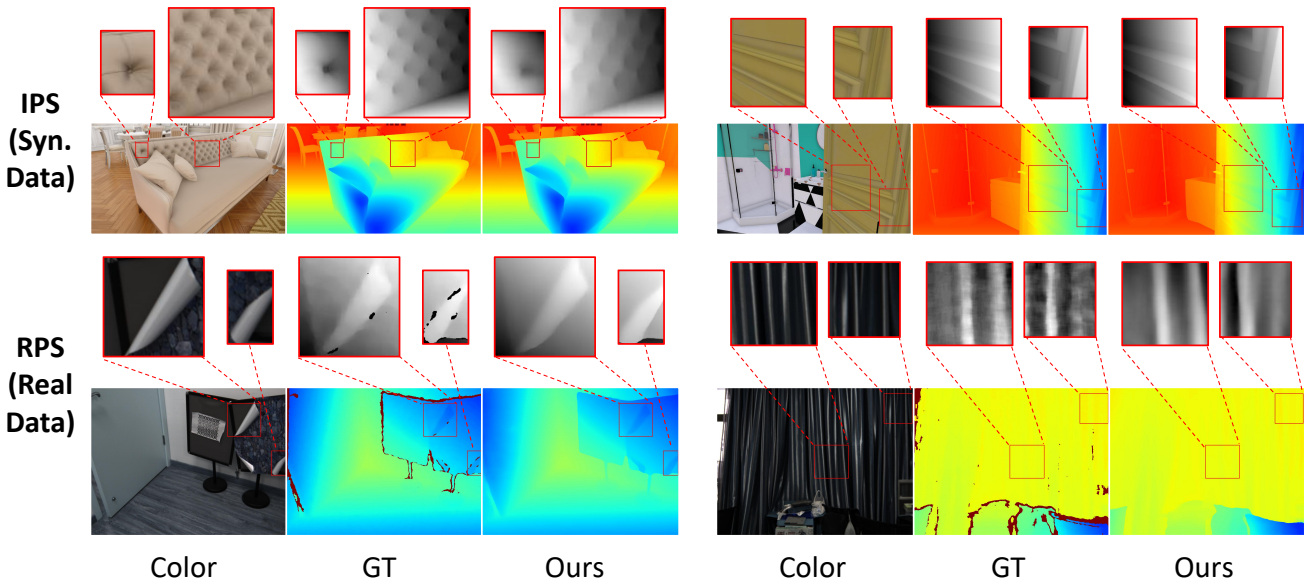

Figure J. Additional high-frequency detail results of the IPS and the RPS datasets. Both the details of the ground truth of the disparity and the results of our method are enlarged. Our DPS-Net can preserve high-frequency details, especially for the synthetic data.

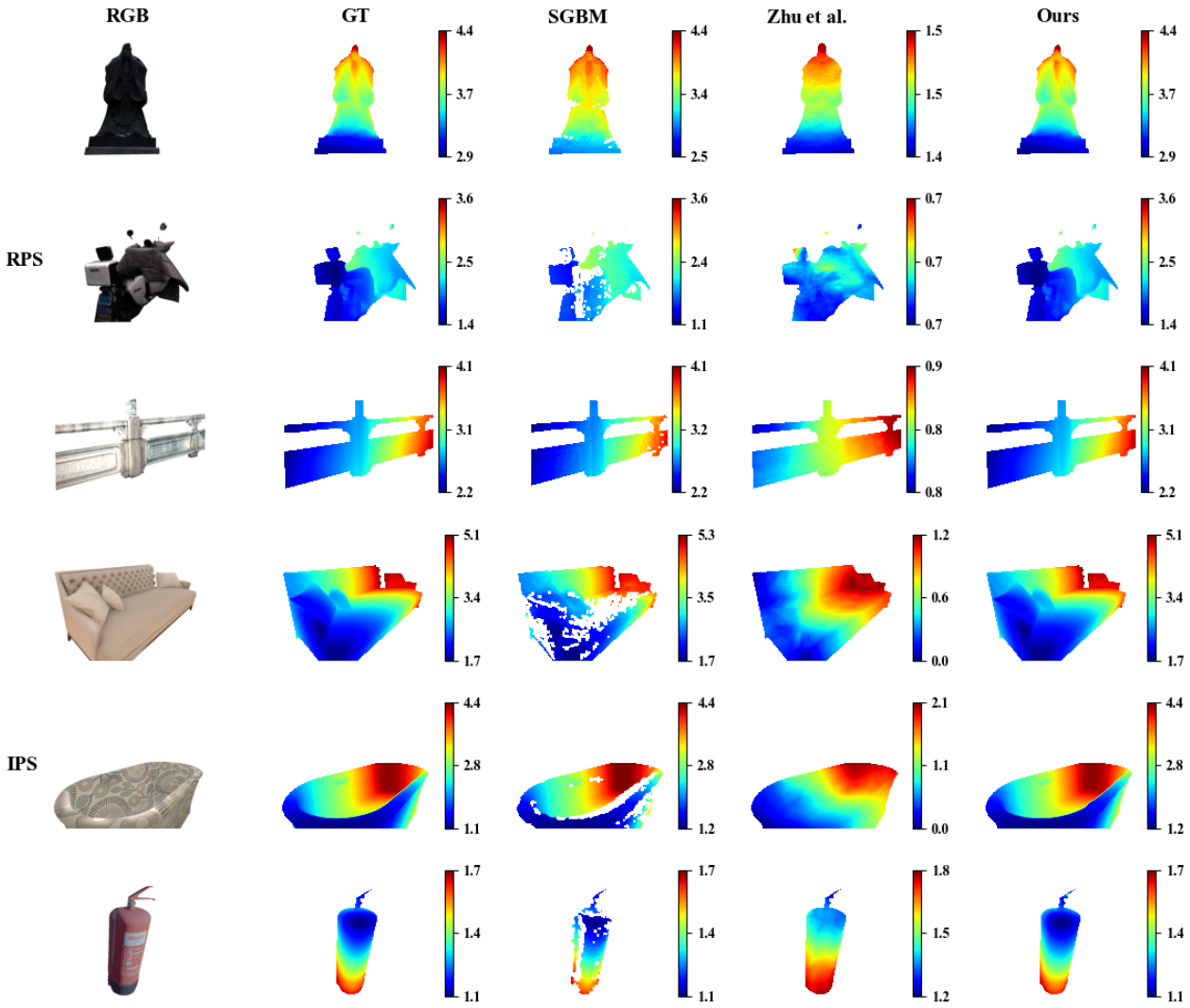

Figure K. More comparison results with traditional stereo depth estimation methods. **From left to right:** input left images, the ground truth of disparities, the results of SGBM[4], the results of [14], and the results of our method. The color bars exhibited are used to show the data range of disparities.

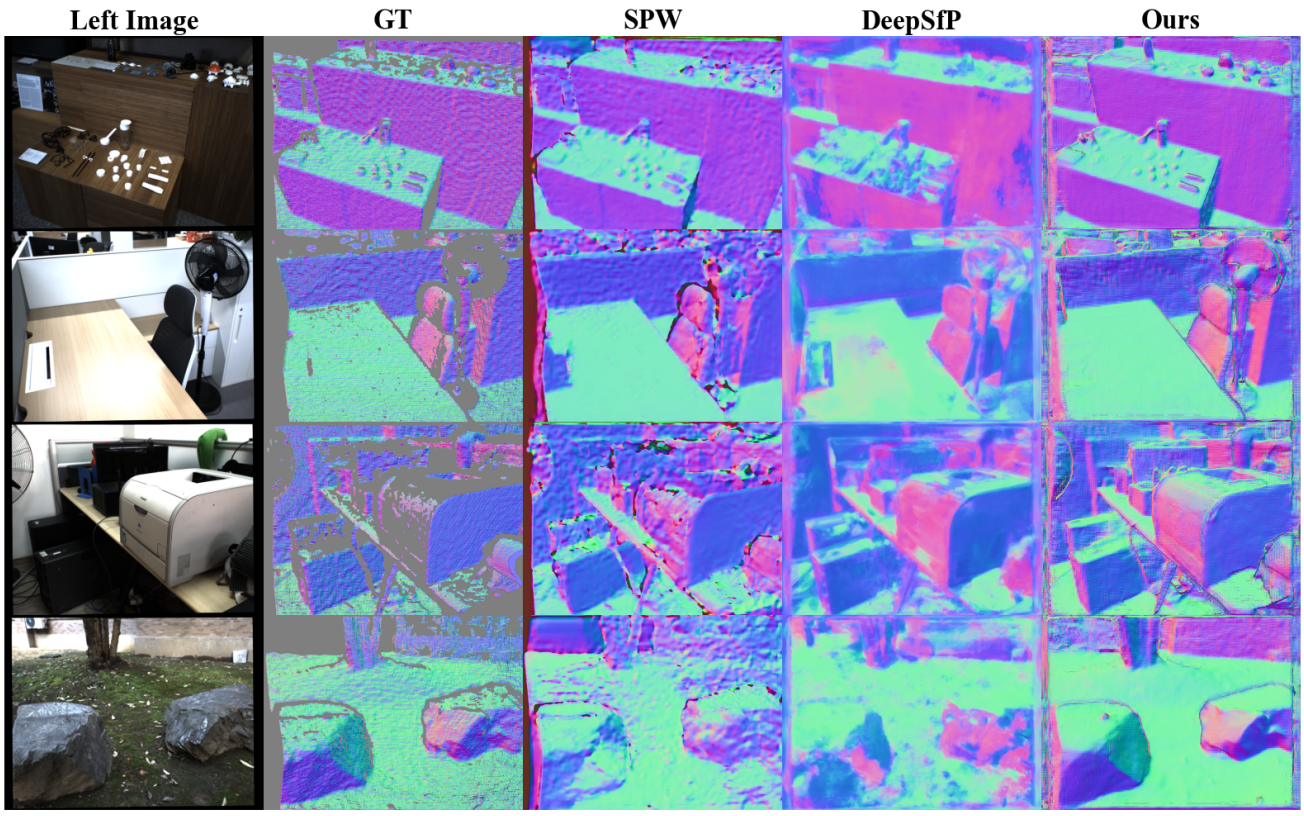

Figure L. Additional qualitative normal estimation results of our method, SPW[6] and DeepSfP[1] on the RPS dataset. Our DPS-Net can achieve better performance.