Glass Segmentation using Intensity and Spectral Polarization Cues (Supplementary Material)

Haiyang Mei[1] Bo Dong[2,⋆] Wen Dong[1] Jiaxi Yang[1] Seung-Hwan Baek[2,3] Felix Heide[2]
Pieter Peers[4] Xiaopeng Wei[1,⋆] Xin Yang[1,⋆]
[1] Dalian University of Technology [2] Princeton University [3] POSTECH [4] College of William & Mary
https://mhaiyang.github.io/CVPR2022_PGSNet

In this Supplemental Material, we present additional information and results in support of the main manuscript. First, we describe the trichromatic AoLP and DoLP image formation process (section 1), and section 2 reviews Brewster's law and quantitatively describe the different behavior of s- and p-polarization based on the Fresnel equations. Next, section 3 provides additional visual examples from our RGBP-Glass dataset. For completeness, we also provide a brief introduction to Conformer networks which are a central building block in PGSNet (section 4). In section 5, we describe the four metrics used for quantitatively evaluating the effectiveness of the segmentation approaches. Sections 6-8 describe additional evaluations and ablation studies of the performance with different Conformer backbones, the MSDP module compared to other mainstream attention schemes, and the effectiveness of the EDA module. Finally, section 9 provides additional qualitative comparisons of PGSNet against state-of-the-art approaches.

1. Trichromatic Polarization Image Formation

Our large-scale polarization glass segmentation dataset, named *RGBP-Glass*, is captured using a trichromatic polarizer-array camera (LucidVision PHX050S) that records four different linear-polarization directions ($0°$, $45°$, $90°$, and $135°$) for each color channel (*i.e.*, R, G, and B). Therefore, unlike a regular Bayer pattern mosaic, the camera captures the Bayer pattern mosaic in a superpixel manner. In particular, a 2×2 square superpixel is under each color filter, and each pixel of the superpixel records one of the four different linear polarization directions.

To obtain trichromatic polarization cues, we first estimate the first three Stokes parameters for each color channel, denoted as $S_0^X, S_1^X, S_2^X, X \in \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}$, as

$$S_0^X = I_0^X + I_{90°}^X, \quad S_1^X = I_0^X - I_{90°}^X, \quad S_2^X = I_{45°}^X - I_{135°}^X \quad X \in \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}, \tag{1}$$

where I_x denotes the image captured by a linear polarizer at the angle x. With the three Stokes parameters, we can estimate the DoLP and AoLP for each color channel as

$$\mathrm{DoLP}^{\mathrm{X}} = \frac{\sqrt{(S_1^X)^2 + (S_2^X)^2}}{S_0^X}, \quad \mathrm{AoLP}^{\mathrm{X}} = \frac{1}{2}\arctan\left(\frac{S_2^X}{S_1^X}\right) \quad X \in \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}. \tag{2}$$

The process is also illustrated in Figure 1.

2. Brewster's Law

A single ray of natural light actually consists of two independent oppositely polarized components, which represent the s- and p-polarization states. The *s* and *p* notation come from the German words for perpendicular (*senkrecht*) and parallel (*paralelle*) [2]. Intuitively, the p-polarized component's electric field lies *in* (*i.e.*, parallel) the plane of incidence, and the s-polarized component's electric field is normal (*i.e.*, perpendicular) to the plane of incidence.

Around 1812, David Brewster found that the change in magnitude of the two polarization components differs when reflected of a flat glass surface. Notably, the p-polarized component vanishes completely at a particular angle of incidence i,

⋆ Xin Yang (xinyang@dlut.edu.cn) and Xiaopeng Wei are the corresponding authors. Xin Yang and Bo Dong lead this project.

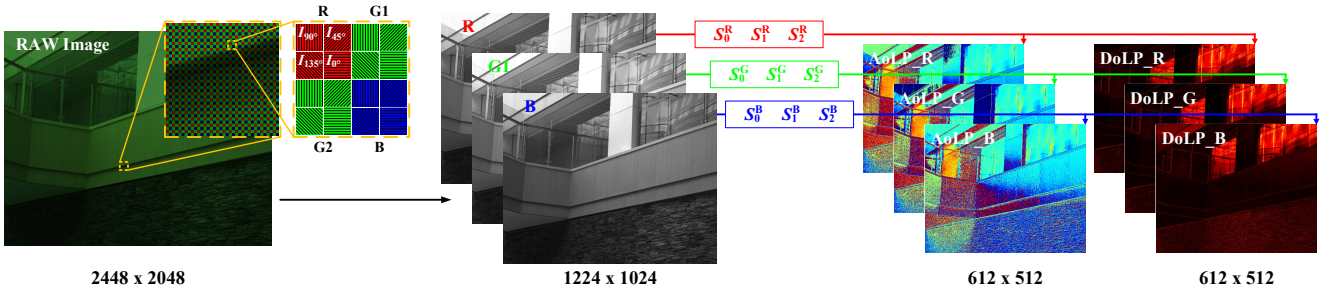

Figure 1. The trichromatic polarization image formation process.

which is called Brewster's angle. Brewster also discovered the relationship between the incident angle i and refracted angle r as, $i + r = 90°$; see Figure 2. The magnitude of Brewster's angle depends on the ratio refractive indices

$$\theta_B = \arctan\left(\frac{n_2}{n_1}\right), \tag{3}$$

where n_1 is the refractive index of the initial medium through which the light propagates (the "incident medium"), and n_2 is the index of the other medium. This equation is known as Brewster's law.

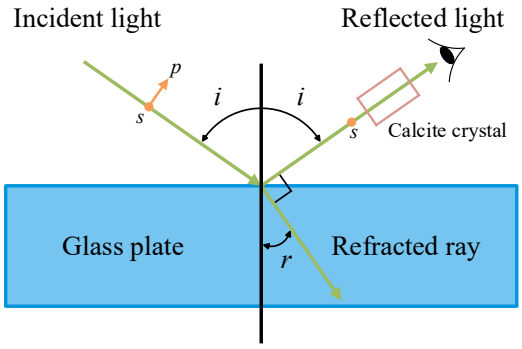

Figure 2. Polarization relation between reflected and refracted rays.

An unpolarized ray hitting the surface at the Brewster's angle, becomes fully s-polarized after reflection. The magnitude of both the reflected s- and p-polarized component at a media surface can be calculated based on Fresnel equations. In Figure 3, we show the s- and p-polarization magnitude of glass in terms of angle of incidence. Most materials have a relative refractive index that varies with wavelength, resulting in wavelength-dependent s- and p-polarization curves and Brewster's angles. This wavelength-dependent phenomenon underlines the motivation of the proposed spectral polarimetric processing in PGSNet.

3. RGB-P Glass Segmentation Dataset

Our RGBP-Glass dataset contains 4,511 images, of which 3,207 are used for training PGSNet and 1,304 are used for testing. For each image, we provide the corresponding estimated AoLP and DoLP images of each color channel. In addition, each image comes with ground truth glass masks for segmentation tasks and bounding boxes for detection tasks. We compensate for the reduction in light sensitivity due to the linear-polarization filters on the sensor by using an $f/1.6$ aperture and manually adjusting the exposure time based on the illumination conditions. In addition to the examples shown in our main manuscript, Figure 4 provides more examples of our RGBP-Glass dataset. Specifically, in the first four rows, we show examples with strong and similar polarization cues in both AoLP and DoLP across all three color channels; The second group (*i.e.*, rows 5-8) demonstrate the case where polarization cues differ between the color channels; The last four rows are the examples of weak intensity and polarization cues.

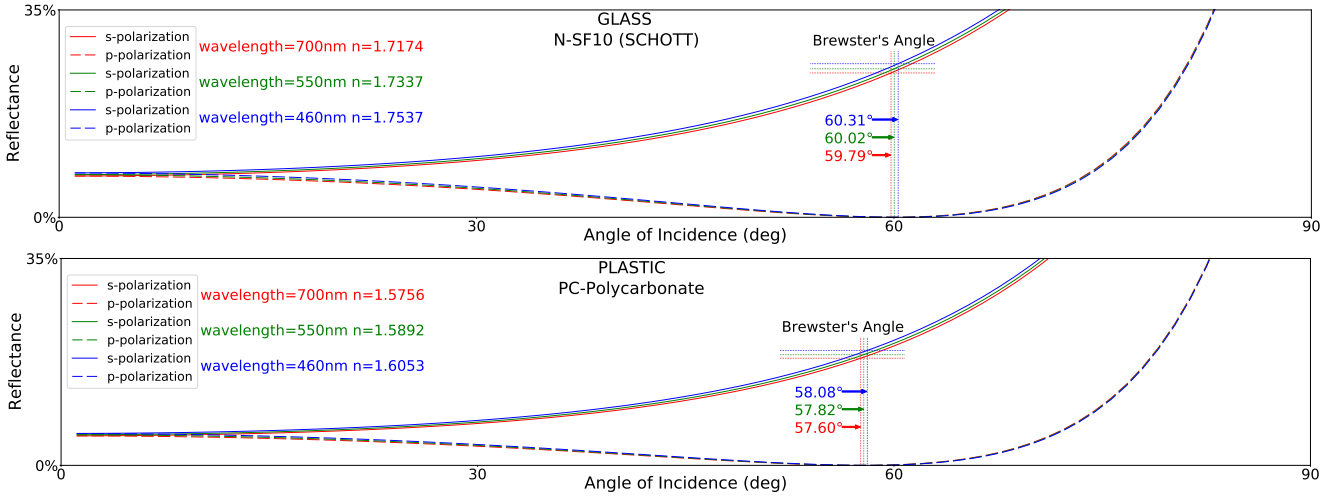

Figure 3. Reflectance plots of the s- and p-polarization component with respect to incident angle.

4. Brief Intoduction to Conformers

Conformer [15], a dual network structure, consists of a convolutional branch and a transformer branch. In particular, the ResNet [8] architecture is used for the convolutional branch, and the transformer branch follows the design of ViT [3]. With the specially designed feature coupling unit (FCU), the local features offered by the convolutional branch and the global features provided by the transformer branch are fused under different resolutions in an interactive fashion. As such, the discriminability of both local and global features is enhanced significantly.

As discussed in the main paper, both local and global contextual cues from the three input domains (*i.e.*, RGB, AoLP, and DoLP) are equally crucial to the glass segmentation task. In the proposed PGSNet, for each input domain, we leverage a Conformer as our backbone network for extracting both local and global features for downstream processing.

5. Evaluation Metrics

For our evaluation, we adopt four widely used metrics for quantitatively assessing the glass segmentation performance: intersection over union (IoU), weighted F-measure (F_β^w) [12], mean absolute error (MAE), and balance error rate (BER) [14].

Intersection over union (IoU) is a widely used metric in segmentation tasks, which is defined as:

$$IoU = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}(G(i,j) * P_b(i,j))}{\sum_{i=1}^{H}\sum_{j=1}^{W}(G(i,j) + P_b(i,j) - G(i,j) * P_b(i,j))}, \tag{4}$$

where G is the ground truth mask in which the values of the glass region are 1 while those of the non-glass region are 0; P_b is the predicted mask binarized with a threshold of 0.5; and H and W are the height and width of the ground truth mask, respectively.

Weighted F-measure (F_β^w) is adopted from the salient object detection tasks. F-measure (F_β) is a measure on both the precision and recall of the prediction map. Recent studies [4, 5] have suggested that the weighted F-measure (F_β^w) [12] can provide more reliable evaluation results than the traditional F_β. Thus, we report F_β^w in the comparison.

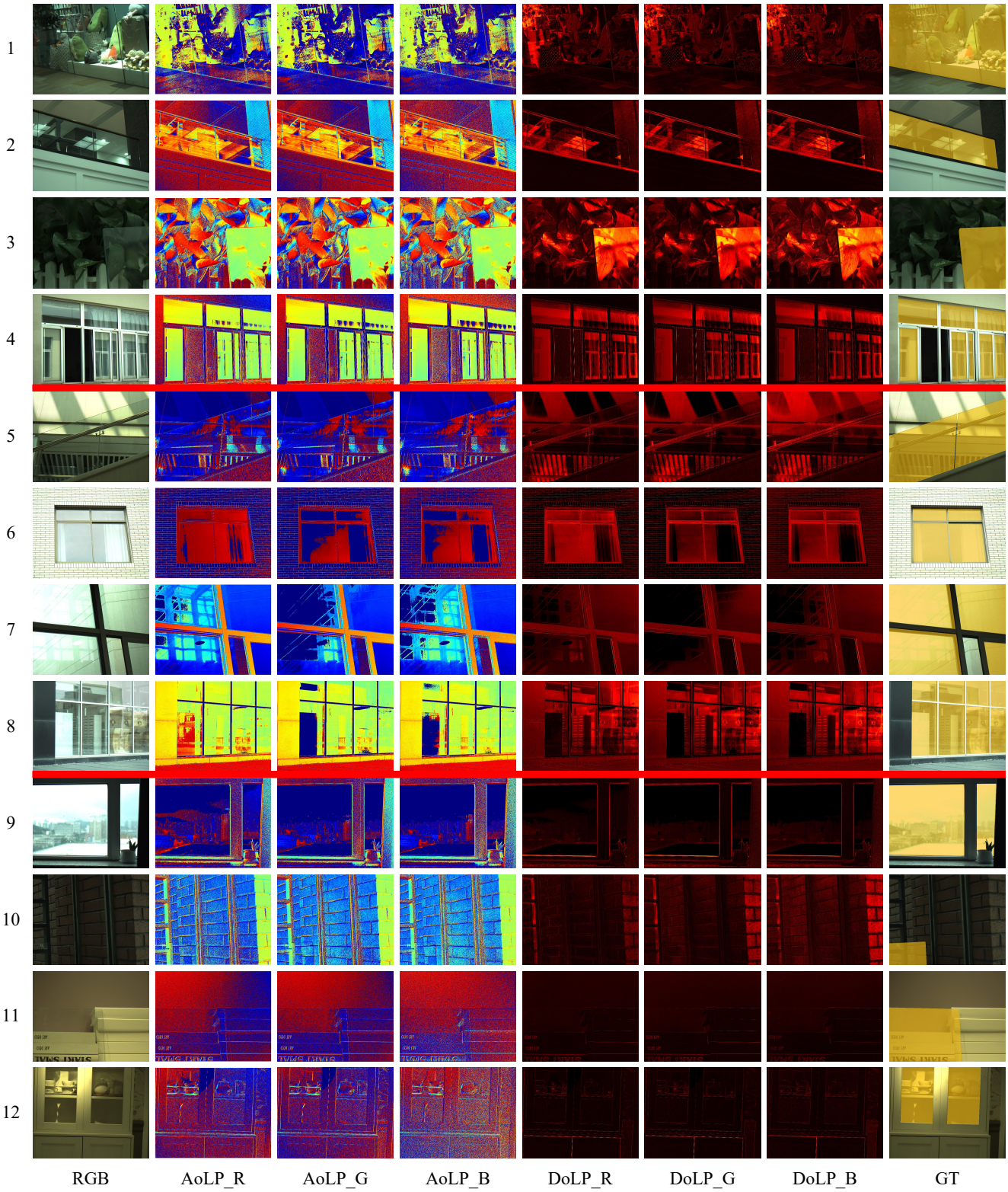

Figure 4. RGBP-Glass examples.

Mean absolute error (MAE) is widely used in foreground-background segmentation tasks, which calculates the element-wise difference between the prediction map P and the ground truth mask G:

$$MAE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} |P(i,j) - G(i,j)|, \tag{5}$$

where $P(i,j)$ indicates the predicted probability score at location (i,j).

Balance error rate (BER) is a standard metric used in shadow detection tasks, defined as:

$$BER = (1 - \frac{1}{2}(\frac{TP}{N_p} + \frac{TN}{N_n})) \times 100, \tag{6}$$

where TP, TN, N_p, and N_n represent the numbers of true positive pixels, true negative pixels, glass pixels, and non-glass pixels, respectively.

6. Performance Analysis

Our main focus in designing PGSNet was on accuracy rather than efficiency. For completeness, we list in Table 1 the performance compared to prior work below as well as comparisons with other variants of Conformer backbone (the last three rows of Table 1). From this we conclude that PGSNet is computationally more expensive than prior methods (albeit within a reasonable range). Furthermore, with modest loss in accuracy, **switching from the Conformer-B to the Conformer-T backbone can yield a 4x speedup**.

Methods	Backbone	FLOPs (G)	IoU↑	F_β^w↑	MAE↓	BER↓
GDNet [13]	ResNeXt-101	271.533	77.64	0.807	0.119	11.79
TransLab [19]	ResNet-50	61.264	73.59	0.772	0.148	15.73
Trans2Seg [20]	ResNet-50	49.034	75.21	0.799	0.122	13.23
GSD [11]	ResNeXt-101	92.697	78.11	0.806	0.122	12.61
EAFNet [18]	ResNet-18	**18.925**	53.86	0.611	0.237	24.65
P Mask R-CNN [10]	ResNet-101	56.594	66.03	0.714	0.178	18.92
SETR [24]	ViT-Large	240.100	77.60	0.817	0.114	11.46
SegFormer [21]	MiT-B5	70.240	78.42	0.815	0.121	13.03
PGSNet	Conformer-T	94.677	77.60	0.806	0.117	11.61
PGSNet	Conformer-S	153.236	78.77	0.821	0.110	11.13
PGSNet	Conformer-B	290.615	**81.08**	**0.842**	**0.091**	**9.63**

Table 1. Performance Analysis of competing approaches and PGSNet with different backbones settings.

7. MSDP vs. PPM/ASPP/non-local

Conceptually, our MSDP module is similar to PPM [23], ASPP [1], non-local [16], and other similar mainstream attention methods. At a high level, one can see our MSDP module as a combination of PPM and non-local. Comparing MSDP, PPM, ASPP, and non-local (see below), we can see in Table 2 that **the MSDP module outperforms existing methods with similar computational costs**.

Methods	FLOPs (G)	IoU↑	F_β^w↑	MAE↓	BER↓
PGSNet w/ PPM [23]	290.524	79.15	0.821	0.106	10.53
PGSNet w/ non-local [16]	**290.445**	79.25	0.824	0.101	10.36
PGSNet w/ ASPP [1]	290.659	79.73	0.826	0.102	10.33
PGSNet w/ MSDP (Ours)	290.615	**81.08**	**0.842**	**0.091**	**9.63**

Table 2. Comparison between our MSDP module and other attention schemes.

8. Effectiveness of EDA module

As discussed in the main manuscript, our Early Dynamic Attention (EDA) module is effective in balancing the different spectral components in both the DoLP and AoLP. Please refer to Table 3 (A vs. G) in the main manuscript. Figure 5 provides visual confirmation of these numerical results that show AoLP and DoLP examplars where the EDA module helps to distinguish the glass region.

In particular, we show the normalized AoLP/DoLP image before and after EDA module. The normalized image, I_N, is mathematically defined as:

$$I_N = \frac{I - I_{min}}{I_{max} - I_{min}},$$

where I is an AoLP/DoLP image; I_{min} and I_{max} represent the minimum and maximum value of image I, respectively.

Based on these three examples, we can see the glass areas in the AoLP domain become more evident after the EDA module. In contrast, the glass areas are distinguishable in the original DoLP images, and thereby EDA module changes them subtly.

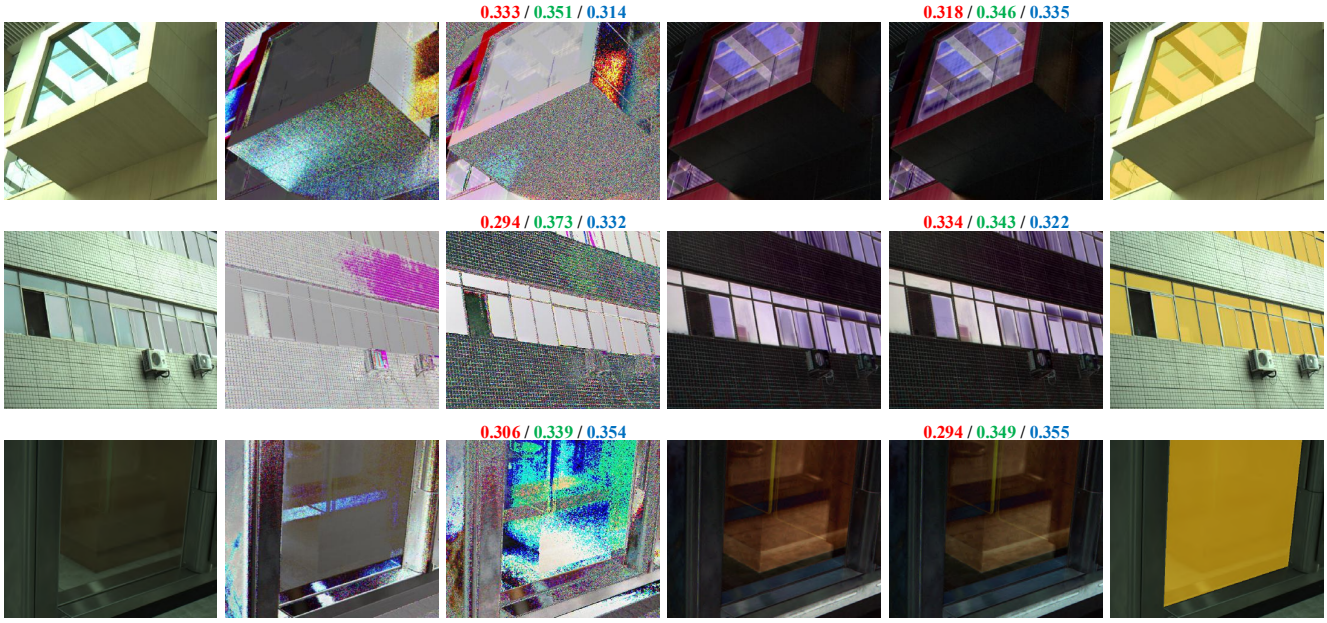

Figure 5. The visual example that reflects the effectiveness of EDA module. The **red**, **green**, and **blue** number indicates the fusion weights generated from the EDA module for the red, green, and blue channel, respectively.

9. Additional Qualitative Comparisons

We provide additional qualitative results of PGSNet compared to state-of-the-art approaches. In particular, besides the five state-of-the-art glass segmentation methods discussed in our main manuscript, we also include the best performers in seven other related tasks retrained on the RGBP-Glass dataset: CCNet [9] a semantic segmentation network; CPD [17] a salient object detection solution; SINet-V2 [6] segments camouflaged objects; PraNet [7] is designed for medical image segmentation; BDRAR [25] detects shadows; MirrorNet [22] focuses on mirror segmentation task; and EAFNet [18] performs RGB-P-based semantic segmentation. Five selected examples and the corresponding estimated masks of the 12 state-of-the-art approaches are shown in Figure 6, Figure 7, Figure 8, Figure 9, and Figure 10, respectively. The proposed PGSNet makes the best estimate in all five examples.

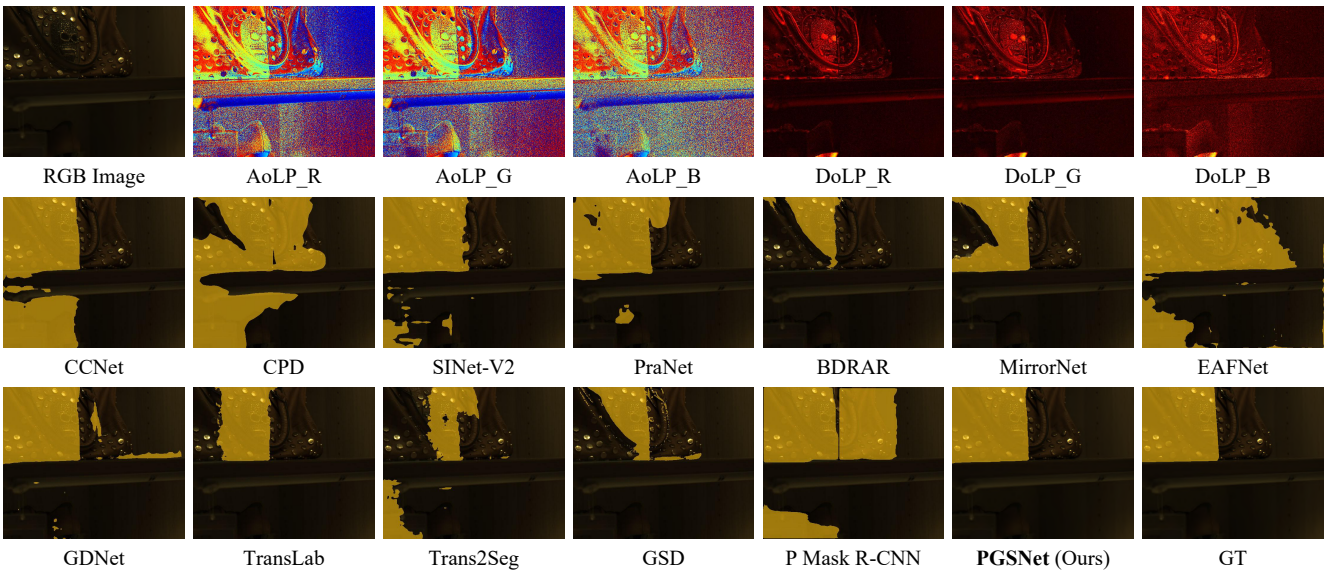

Figure 6. Qualitative comparison of PGSNet against state-of-the-art segmentation methods retrained on the RGBP-Glass dataset.

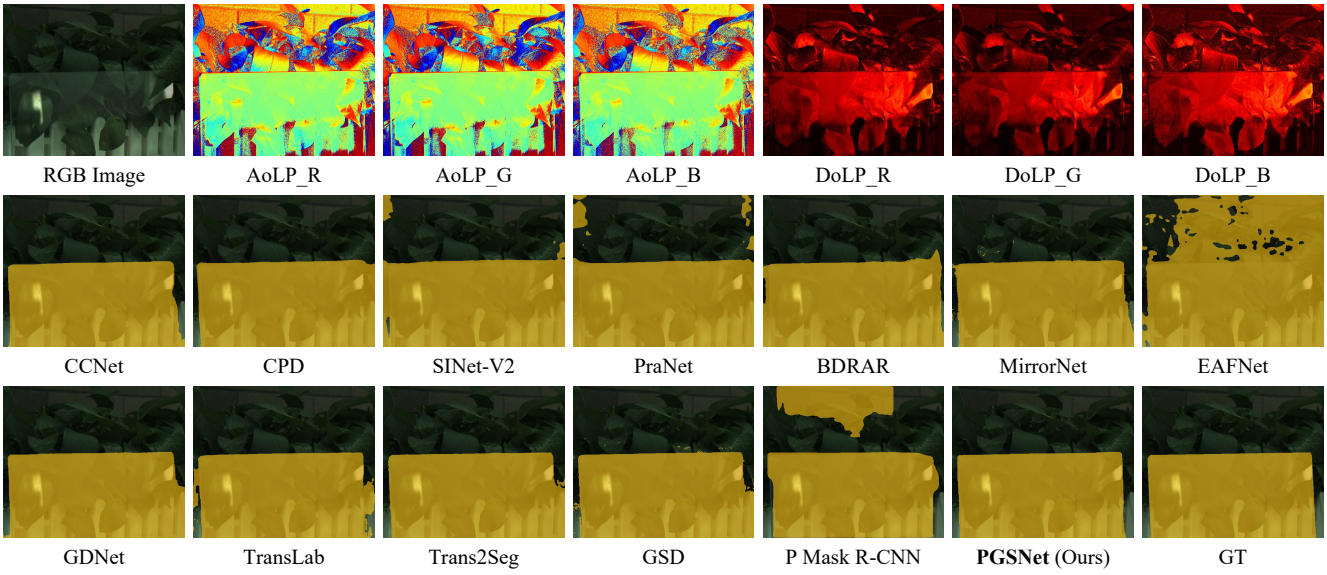

Figure 7. Qualitative comparison of PGSNet against state-of-the-art segmentation methods retrained on the RGBP-Glass dataset.

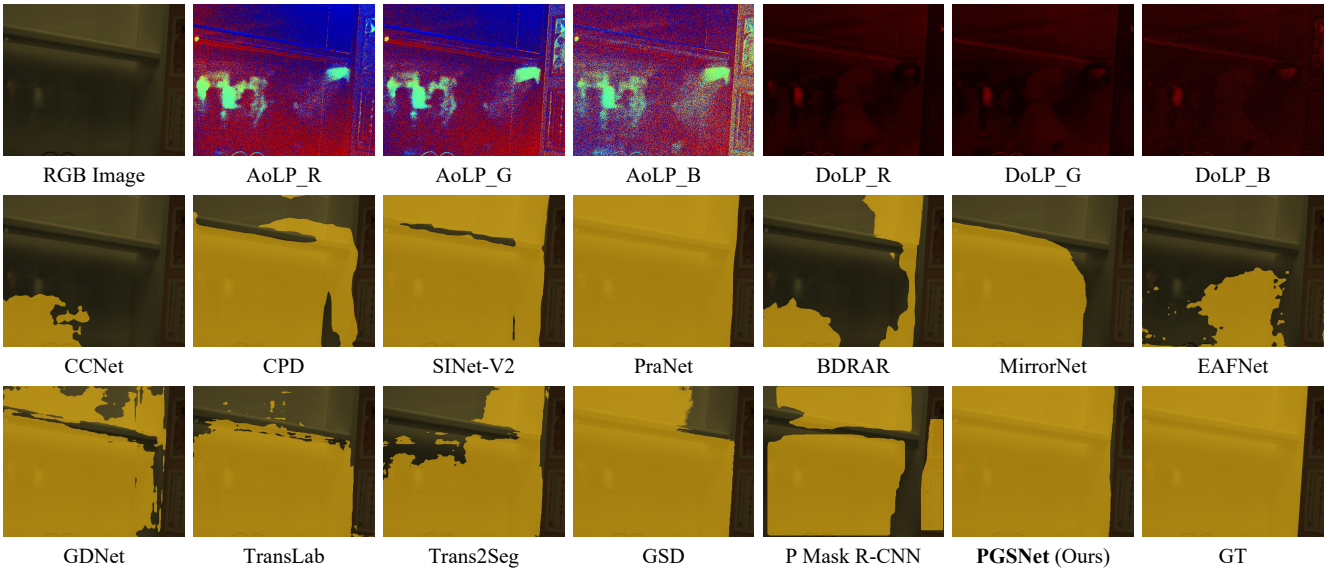

Figure 8. Qualitative comparison of PGSNet against state-of-the-art segmentation methods retrained on the RGBP-Glass dataset.

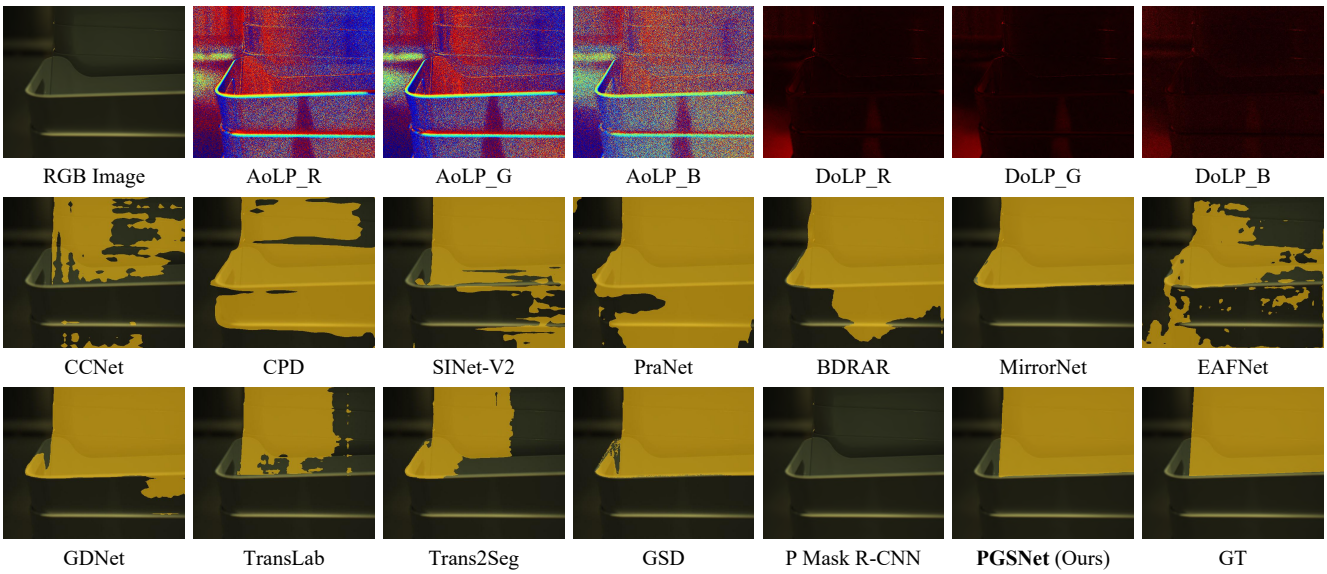

Figure 9. Qualitative comparison of PGSNet against state-of-the-art segmentation methods retrained on the RGBP-Glass dataset.

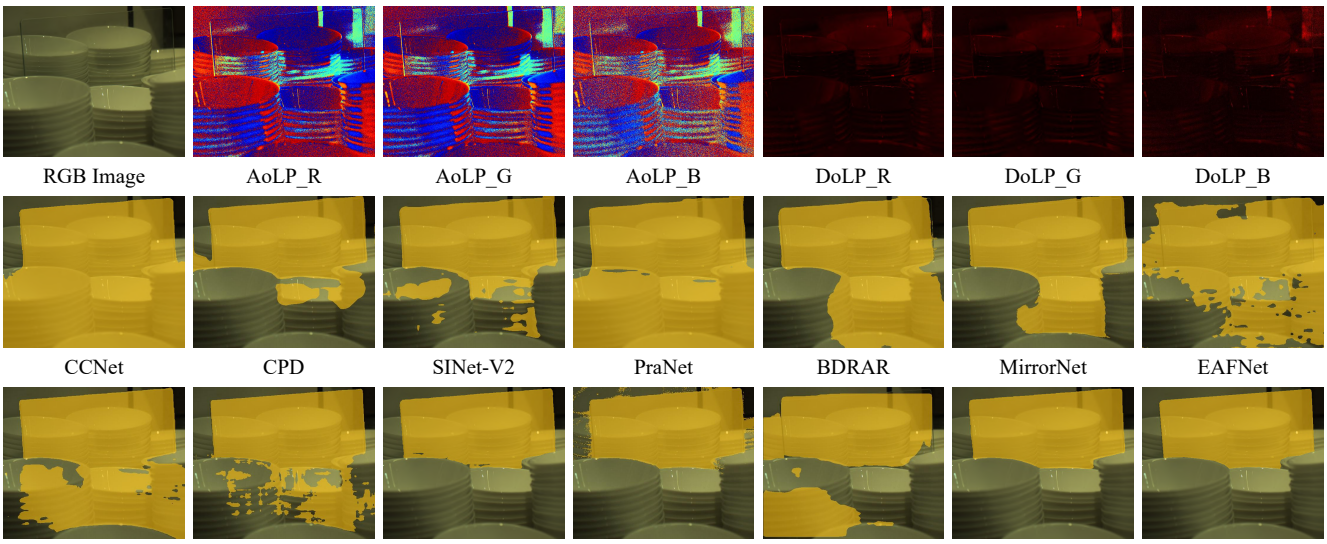

Figure 10. Qualitative comparison of PGSNet against state-of-the-art segmentation methods retrained on the RGBP-Glass dataset.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 2017. 5

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

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020. 3

[4] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, 2017. 3

[5] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, 2018. 3

[6] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE TPAMI*, 2021. 6

[7] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. *MICCAI*, 2020. 6

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[9] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019. 6

[10] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *CVPR*, 2020. 5

[11] Jiaying Lin, Zebang He, and Rynson W.H. Lau. Rich context aggregation with reflection prior for glass surface detection. In *CVPR*, 2021. 5

[12] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, 2014. 3

[13] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson W.H. Lau. Don't hit me! glass detection in real-world scenes. In *CVPR*, 2020. 5

[14] Vu Nguyen, Tomas F Yago Vicente, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. Shadow detection with conditional generative adversarial networks. In *ICCV*, 2017. 3

[15] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *ICCV*, 2021. 3

[16] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 5

[17] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, 2019. 6

[18] Kaite Xiang, Kailun Yang, and Kaiwei Wang. Polarization-driven semantic segmentation via efficient attention-bridged fusion. *Optics Express*, 2021. 5, 6

[19] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In *ECCV*, 2020. 5

[20] Enze Xie, Wenjia Wang, Wenhai Wang, Peize Sun, Hang Xu, Ding Liang, and Ping Luo. Segmenting transparent object in the wild with transformer. In *IJCAI*, 2021. 5

[21] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 2021. 5

[22] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson W.H. Lau. Where is my mirror? In *ICCV*, 2019. 6

[23] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 5

[24] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 5

[25] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *ECCV*, 2018. 6