GlassWizard: Harvesting Diffusion Priors for Glass Surface Detection

Supplementary Material

We present additional implementation details and analysis of our proposed method GlassWizard in this supplementary material.

1. More Experimental Details

1.1. Datasets

In **Stage I** and **Stage II**, we utilize the combination of the training sets of *GDD* [6], *Trans10K-Stuff* [11], *GSD* [4] and *HSO* [14], which is summarized in Table 1.

- The *GDD* dataset [6] includes 2,980 training images and 936 test images, drawn from both outdoor and indoor scenes.
- The *Trans10K* [11] dataset comprises 5,000 training images, 1,000 validation images, and 4,428 test images, categorized into "things" and "stuff." Some works state that "stuff" is more challenging than "things" [9, 14]. Following [14], we focus on a subset consisting of 2,455 training images and 1,771 testing images from the "stuff" category of Trans10K.
- The *GSD* dataset [4] consists of 3,285 training images and 813 test images, all captured from real-world environments.
- The *HSO* dataset [14] contains 3,070 training images and 1,782 test images.

To assess the generalization capability of our method, we conduct zero-shot segmentation using the trained model on the test sets from the *VGSD-D* [5] and *RGBP* [7] datasets. (see Table 2)

- The *VGSD-D* dataset [5] is the first large-scale video glass surface detection dataset, including 12,317 training images and 6,851 test images. These images are from 192 videos and 105 videos, separately.
- The *RGBP* dataset [7] is a large-scale RGB-Polarization dataset, containing 3,206 images in training set and 1,304 images in test set. we utilize solely on the RGB images.

Here, we only adopt the test sets in *VGSD-D* and *RGBP*.

Modality-Customized Adaptation. We test the modality-customized adaptation performance on RGB-D and RGB-T GSD tasks. For RGB-D GSD task, we adopt *TROSD* dataset [10]. For RGB-T GSD task, we use *RGBT* [2] dataset.

- The *RGBT* [2] dataset includes 4,427 training images and 1,124 test images. The images are captured by the FLIR ONE Pro camera. The thermal and RGB images are aligned with the FLIR Thermal Studio software.
- The *TROSD* dataset [10] dataset contains 7,421 images in training set and 3,639 images in test set. *TROSD* utilizes Structure Sensor to capture RGB-D iamges, which is an infrared structured light sensor. It contains 14 different scenes (*e.g.*, living room, bathroom, office) totally.

Mirror Detection. To assess the transferability of our framework, we performed experiments on the Mirror Detection task. We employ the PMD [3] and MSD [13] datasets, training Stage I and Stage II on the PMD and MSD datasets separately.

- The *PMD* [3] dataset includes 5,095 training images and 571 test images.
- The *MSD* dataset [13] dataset contains 3,063 images in training set and 955 images in test set.

1.2. Evaluation Metrics

The IoU is employed to evaluate the degree of overlap between ground truths and prediction maps as

$$IoU = \frac{TP}{TP + FP + FN}, \tag{1}$$

where the TN, TP, and FN denote the true negative, true positive, and false negative pixels, respectively.

F_β is a weighted average of weighted Precision and weighted Recall, which is defined as

$$F_\beta = \frac{(1+\beta^2)Precision^\omega \times Recall^\omega}{\beta^2 Precision^\omega + Recall^\omega}, \tag{2}$$

$$Precision^\omega = \frac{TP^\omega}{TP^\omega + FP^\omega}, \tag{3}$$

$$Recall^\omega = \frac{TP^\omega}{TP^\omega + FN^\omega}, \tag{4}$$

where the TP^ω, FP^ω and FN^ω is obtained by weighting the absolute error.

MAE reveals the error of the predictions and ground truths, and it is defined as

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

where x and y represent the horizontal and vertical coordinates of the pixel, respectively.

Balance Error Rate (BER) computes the average of the false positive and false negative rates, which helps assess the algorithm's ability to correctly classify both object pixels (foreground) and non-object pixels (background).

$$BER = 1 - 0.5 \times (\frac{N_{TP}}{N_P} + \frac{N_{TN}}{N_N}), \tag{6}$$

Table 1. Detailed information of datasets for different tasks.

Dataset	Task	Train	Test	Sum
GDD [6]	Glass segmentation	2,980	936	3,916
Trans10K-Stuff [11]	Transparent object segmentation (stuff)	2,455	1,771	4,226
GSD [4]	Glass segmentation	3,285	813	4,098
HSO [14]	Glass segmentation in home scenes	3,070	1,782	4,852

Table 2. Detailed information of datasets for testing the generalization ability.

Dataset	Task	Train	Test	Sum
VGSD-D [5]	Vidoe Glass Surface Detection	12,317	6,851	19,168
RGBP [7]	RGB-P Glass Surface Detection	3,206	1,304	4,510

Table 3. Detailed information of datasets for modality-customized adaptation.

Dataset	Task	Train	Test	Sum
RGBT [2]	RGB-T GSD	4,427	1,124	5,551
TROSD [10]	RGB-D GSD	7,421	3,639	11,060

Table 4. Detailed information of datasets for mirror detection.

Dataset	Task	Train	Test	Sum
PMD [3]	Mirror Detection	5,095	571	5,666
MSD [13]	Mirror Detection	3,063	955	4,018

where N_{TP}, N_P, N_{TN}, N_N are the number of true positives, true negatives, object and non-object pixels, respectively. A lower BER indicates better performance in terms of achieving a balanced segmentation, where both foreground and background are accurately predicted.

2. More Comparison Studies

Due to some methods not being open-source and their reproduction results not matching the original paper's performance, we report the performance presented in their respective papers for a fair comparison. For fairness, we train and test separately on each dataset (including both Stage I and Stage II) rather than using the training data splits mentioned in the main text. This approach ensures alignment with the pipeline described in their paper.

The results are shown in Table 5. The experimental results demonstrate that our model achieves the best performance, whether using the data splits mentioned in the main text or on individual datasets. This highlights the strengths of our framework in addressing glass object detection and further underscores the advantages of diffusion priors for the glass surface detection task.

3. Failure Case Analysis

Despite the high quality of the prediction maps generated by our GlassWizard in most scenarios, the proposed method may struggles to accurately localize glasses or transparent objects when faced with particularly challenging situations. Figure 1 illustrates some representative failure cases. In the first row, it can be found that GlassWizard produces erroneous predictions when the glass is positioned at a distance and features discontinuous non-transparent frosted sections. Additionally, as demonstrated in the second row, it may misidentify certain reflective surfaces. Another problematic scenario arises when small pieces of glass are set against a cluttered background, as shown in the third row. Lastly, partially opened glass doors are also prone to misidentification, as depicted in the fourth row. Nonetheless, it is important to emphasize that despite the shortcomings of our method in these specific instances, our results still outperform those of competitors.

References

[1] Hao He, Xiangtai Li, Guangliang Cheng, Jianping Shi, Yunhai Tong, Gaofeng Meng, Véronique Prinet, and LuBin Weng. Enhanced boundary learning for glass-like object segmentation. In *ICCV*, pages 15859–15868, 2021. 3

[2] Dong Huo, Jian Wang, Yiming Qian, and Yee-Hong Yang. Glass segmentation with rgb-thermal image pairs. *IEEE Transactions on Image Processing*, 32:1911–1926, 2023. 1, 2

[3] Jiaying Lin, Guodong Wang, and Rynson WH Lau. Progressive mirror detection. In *CVPR*, pages 3697–3705, 2020. 1, 2

[4] Jiaying Lin, Zebang He, and Rynson WH Lau. Rich context aggregation with reflection prior for glass surface detection. In *CVPR*, pages 13415–13424, 2021. 1, 2, 3

[5] Fang Liu, Yuhao Liu, Jiaying Lin, Ke Xu, and Rynson WH Lau. Multi-view dynamic reflection prior for video glass surface detection. In *AAAI*, pages 3594–3602, 2024. 1, 2

[6] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu,

Table 5. Model performance for transparent object segmentation on the GDD [6], Trans10K-Stuff [11], GSD [4] and HSO [14] datasets.

Model	GDD [6]				Trans10K-Stuff [6]				GSD [4]				HSO [14]			
	IoU↑	F_β↑	MAE↓	BER↓	IoU↑	F_β↑	MAE↓	BER↓	IoU↑	F_β↑	MAE↓	BER↓	IoU↑	F_β↑	MAE↓	BER↓
TransLab [11]	0.816	0.892	0.097	9.70	0.871	0.897	0.051	5.44	0.781	0.828	0.069	9.19	0.743	0.781	0.123	12.00
Trans2Seg [12]	0.844	0.905	0.078	7.36	0.750	0.767	0.124	10.73	0.797	0.839	0.069	8.21	0.780	0.817	0.095	9.65
GDNet [6]	0.876	0.937	0.063	5.62	0.887	0.907	0.046	4.72	0.825	0.857	0.058	6.41	0.787	0.817	0.097	9.32
GSDNet [4]	0.881	0.932	0.059	5.71	0.897	0.917	0.042	4.52	0.836	0.903	0.055	6.12	0.789	0.818	0.103	9.79
EBLNet [1]	0.882	0.935	0.056	5.38	-	-	-	-	0.817	0.878	0.059	6.75	-	-	-	-
PGSNet [14]	0.878	0.901	0.062	5.56	0.898	0.917	0.042	4.39	0.836	0.868	0.054	6.25	0.801	0.836	0.089	9.08
GDNet-B [8]	0.878	0.939	0.061	5.52	-	-	-	-	-	-	-	-	-	-	-	-
VBNet [9]	0.907	0.948	0.048	4.70	0.916	0.955	0.032	3.41	0.861	0.921	0.043	5.51	0.831	0.900	0.078	7.65
Ours	**0.921**	**0.961**	**0.041**	**3.86**	**0.930**	**0.965**	**0.028**	**2.91**	**0.891**	**0.942**	**0.035**	**4.14**	**0.867**	**0.929**	**0.062**	**6.06**

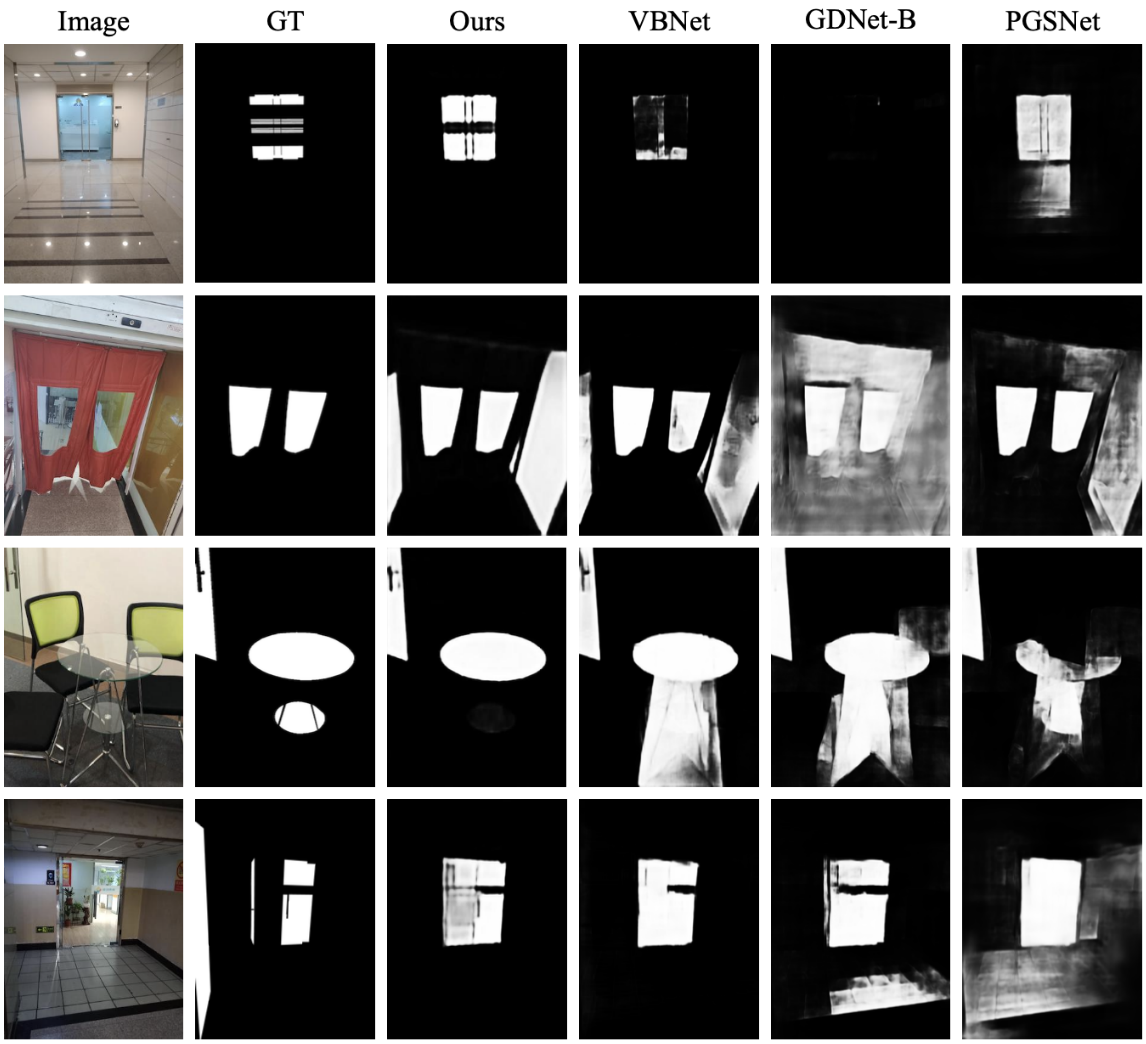

Figure 1. Qualitative comparison of different methods.

Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson W.H. Lau. Don't hit me! glass detection in real-world scenes. In *CVPR*, 2020. 1, 2, 3

[7] Haiyang Mei, Bo Dong, Wen Dong, Jiaxi Yang, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, and Xin Yang. Glass segmentation using intensity and spectral polarization cues. In *CVPR*, pages 12622–12631, 2022. 1, 2

[8] Haiyang Mei, Xin Yang, Letian Yu, Qiang Zhang, Xiaopeng Wei, and Rynson WH Lau. Large-field contextual feature learning for glass detection. *IEEE TPAMI*, 45(3):3329–3346, 2023. 3

[9] Fulin Qi, Xin Tan, Zhizhong Zhang, Mingang Chen, Yuan Xie, and Lizhuang Ma. Glass makes blurs: Learning the visual blurriness for glass surface detection. *IEEE Transactions on Industrial Informatics*, 2024. 1, 3

[10] Tianyu Sun, Guodong Zhang, Wenming Yang, Jing-Hao Xue, and Guijin Wang. Trosd: A new rgb-d dataset for transparent and reflective object segmentation in practice. *IEEE TCSVT*, 33(10):5721–5733, 2023. 1, 2

[11] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In *ECCV*, pages 696–711. Springer, 2020. 1, 2, 3

[12] 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. 3

[13] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson WH Lau. Where is my mirror? In *ICCV*, pages 8809–8818, 2019. 1, 2

[14] Letian Yu, Haiyang Mei, Wen Dong, Ziqi Wei, Li Zhu, Yuxin Wang, and Xin Yang. Progressive glass segmentation. *IEEE TIP*, 2022. 1, 2, 3