Supplementary Material: Leveraging RGB-D Data with Cross-Modal Context Mining for Glass Surface Detection

Anonymous submission

Overview

This appendix is organized as follows:

- We provide more details of our dataset.
- We provide more details of our method.
- More experimental results and ablation studies are included.

RGB-D GSD Dataset

Dataset construction

In Table 1, we show the composition of our proposed dataset. We follow the dataset split of the original datasets. For SUN RGB-D (Armeni et al. 2017), we have identified 630 images from the training set and 573 images from the test set of the original dataset, with glass surfaces. We then reallocate 290 images from the test set to the training set in order to keep the training-test ratio. For 2D-3D-Semantics (Song, Lichtenberg, and Xiao 2015), we follow cross validation fold #2 (*i.e.*, areas 1, 5, 6 for training and areas 2, 4 for testing). For Matterport3D (Chang et al. 2017), we randomly split the selected images into a training set with 992 images and a test set with 214 images. Refer to Table 1 for a summary of the composition of our dataset. Each RGB image is accompanied with a pre-processed depth image and finely annotated ground truth mask. The depth images were taken by different RGB-D cameras models, *e.g.*, Asus Xtion, Kinect v2 (Armeni et al. 2017) and Matterport (Chang et al. 2017) cameras. Although all depth images were encoded in 16-bit grayscale format, the definitions for missing depth are not the same in these three original datasets. For example, in SUN RGB-D (Armeni et al. 2017), un-returned depth signals were set to be the minimum value, which depends on the the depth ranges of individual images. On the other hand, 2D-3D-Semantics (Song, Lichtenberg, and Xiao 2015) assumed invalid depth signals to be the maximum depth value (*i.e.*, $2^{16}-1$).

Table 1: Composition of our proposed RGB-D GSD dataset. We collect glass images from three existing RGB-D datasets. Note that as these datasets were originally created for other tasks, they do not include accurate annotations of glass surface masks. Thus, we annotate the GT masks of the glass surfaces in our dataset construction.

Dataset	Whole	Train	Test
SUN RGB-D	1,203	920	283
2D-3D-Semantics	600	488	112
Matterport3D	1,206	992	214
Total	3,009	2,400	609

Method

As shown in Figure 4 in our main paper, the proposed framework consists of four major components: the backbone network for the input RGB images (in red), the backbone network for the input depth maps (in yellow), the cross-modal context mining (CCM) modules (in blue), and the depth-missing aware attention (DAA) modules (in green). These components are arranged to enable multi-stage feature learning with bottom-up and top-down information flows.

Lighter Depth backbone. The depth backbone network is much simpler and lighter, compared to the RGB one. There are two reasons. First, using a lighter depth backbone network makes our full framework more efficient in both training and test stages. Second, we observe that depth maps contain sparser information. Simply adopting the same network as the RGB image for the depth map may cause a modality gap between the RGB and depth information, which will lead to performance degradation. Refer to Table 2 for the detailed network architecture.

RGB Context Mining Submodule. The CNA mechanism that we use in the submodule consists of an average pooling layer and two convolution layers with a ReLU and sigmoid activation, as:

$$CNA(x) = x \times \sigma(\psi_2(ReLU(\psi_1(\mu(x))))), \qquad (1)$$

where μ, $ReLU$, σ and ψ are the global average pooling (GAP) layer, ReLU, sigmoid function and convolution layers with a 1×1 kernel, respectively. ψ_1 and ψ_2 are 1×1 convolution layers with different weights. x represents the input features. The output of channel-wise attention has the same number of channels as the input features x. Similarly, we can obtain the CXA by adjusting the output channels of the convolution layers that we use.

Depth Context Mining Submodule. The difference be-

Table 2: The architecture of the depth backbone network that we use for the input depth map. It consists of five stages, and each stage contains a convolution layer followed by a pooling layer. Note that each "conv-BR" corresponds a sequence of convolution layer, BatchNorm layer and ReLU activation. K, S and P denote the number of kernels, the number of strides and the padding size, respectively, used in the convolution layer.

Layers Name	Layer Details	Output Size
Convolution	3×3 conv-BR, $K = 8$, $S = 1$, $P = 1$	384×384
Pooling	2×2 max pool, stride 2	192×192
Convolution	3×3 conv-BR, $K = 16$, $S = 1$, $P = 1$	192×192
Pooling	2×2 max pool, stride 2	96×96
Convolution	3×3 conv-BR, $K = 32$, $S = 1$, $P = 1$	96×96
Pooling	2×2 max pool, stride 2	48×48
Convolution	3×3 conv-BR, $K = 64$, $S = 1$, $P = 1$	48×48
Pooling	2×2 max pool, stride 2	24×24
Convolution	3×3 conv-BR, $K = 128$, $S = 1$, $P = 1$	24×24
Pooling	2×2 max pool, stride 2	12×12

tween the RGB context mining submodule and the depth context mining submodule is that the former submodule takes the RGB backbone features x^{RGB} as input, while the later submodule takes the depth backbone features x^{depth}. The model weights of these two submodules are not shared, so that they can focus on context mining in their own modalities.

Implicit Multi-modal Context Mining Submodule. The implicit multi-modal context mining submodule aims to extract multi-modal rich contextual features implicitly by taking the fused multi-modal features as input. Specifically, this submodule takes the RGB backbone features x^{RGB} and the depth backbone features x^{depth} as input. These two input features are first concatenated and forwarded to a convolution layer to obtain the implicit multi-modal input features x^{mul}.

Adaptive Selection. For example, it would be challenging to predict glass surfaces from insufficient visual information (*e.g.*, lack of context and weak reflection), while the depth information may be able to supplement this limitation. In addition, rigidly selecting a particular set of contextual information from different modalities can reduce the generality of the proposed model, as the contextual information from different modalities may different cases in predicting the glass surfaces.

Experiments

Datasets and Evaluation Metrics

For the evaluation, we use four metrics to evaluate the performances of our methods: intersection over union (IoU), F-measure, mean absolute error (MAE), and balance error rate (BER). MAE is formulated as:

$$MAE = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} |P(i,j) - G(i,j)|, \qquad (2)$$

where P is the predicted mask, and G is ground truth. H and W are the width and height of the input image.

F-measure is calculated by a weighted combination of Precision and Recall:

$$F_\beta = \frac{1 + \beta^2 (Precision \times Recall)}{\beta^2 Precision + Recall}, \qquad (3)$$

where β^2 is set to 0.3 as suggested in (Achanta et al. 2009).

The IoU score is calculated as:

$$IoU = \frac{N_{tp}}{N_{tp} + N_{fp} + N_{fn}}, \qquad (4)$$

where N_{tp}, N_{fp} and N_{fn} are the numbers of true positive, false positive and false negative pixels, respectively.

The BER score is a widely used metric in shadow detection to measure the binary prediction from a balance-aware prospective, and is formulated as:

$$BER = 1 - 0.5 \times (\frac{N_{tp}}{N_p} + \frac{N_{tn}}{N_n}), \qquad (5)$$

where N_{tp}, N_{tn}, N_p, N_n are the numbers of true positive, true negative, glass and non-glass pixels, respectively.

Implementation Details

Our proposed network is implemented using Pytorch. The details of our depth backbone network are shown in Table 2. We resize all RGB images, depth maps and the corresponding ground truth masks to the spatial size of 400×400, and then randomly crop them to 384×384. To prevent overfitting, we adopt random horizontal flipping during our training process. We randomly initialize the parameters in all layers except the backbone network for RGB input images.

Ablation Study

Effectiveness of the CCM Module. Table 3 shows the ablation study on the proposed CCM module. Specifically, we keep all other modules in the final model while replacing our proposed CCM module with its variants, where "RGB", "D", "imp.", and "exp." refer to the RGB context mining submodule, depth context mining submodule, implicit multi-modal context mining submodule, and explicit multi-modal context mining submodule, respectively, in the CCM module. We can see that the single-modal variants (*i.e.*, "CCM w/ RGB" and "CCM w/ D") have the worse performances, compared with the other three multi-modal variants. We also observe that the ablated models with the cross-modal context mining submodule (*i.e.*, "CCM w/ RGB + D + imp." and "CCM w/RGB + D + exp.") outperform those without the submodules (*e.g.*, "CCM w/RGB + D"). This indicates the importance of cross-modal context modeling in our CCM module. Finally, our final model performs the best among all ablated models, which shows that the CCM module with cross-modal mining can provide a great performance improvement in glass surface detection.

Effectiveness of the DAA Module. Table 4 shows the ablation study on our proposed DAA module. "DAA w/o Dm" refers to the DAA module without taking the depth-missing map as input. We design three other ablated models: "DAA

Table 3: Ablation study of the CCM module, on our RGB-D GSD dataset. "RGB", "D", "imp.", and "exp." refer to the RGB context mining submodule, depth context mining submodule, implicit multi-modal context mining submodule, and explicit multi-modal context mining submodule, respectively, in the CCM module. "CCM w/ RGB" refers to the CCM module containing only the RGB context mining submodule, while "CCM w/ RGB + D" refers to the CMM module with both RGB and depth context mining submodules.

Methods	IoU↑	F_β ↑	MAE↓	BER↓
CCM w/ RGB	0.708	0.819	0.046	10.44
CCM w/ D	0.695	0.815	0.053	10.92
CCM w/ RGB + D	0.716	0.827	0.047	10.18
CCM w/ RGB + D + imp.	0.736	0.839	0.046	9.66
CCM w/ RGB + D + exp.	0.737	0.841	0.043	9.65
Ours	**0.742**	**0.853**	**0.043**	**9.33**

Table 4: Ablation study of the DAA module, on our RGB-D GSD dataset. "DAA w/o Dm" refers to the DAA module without using the depth missing map as input. "DAA on RGB/Depth/CM" refers to the DAA module applied on the RGB/depth/cross-modal contextual features extracted by the preceding CCM modules in *stage4* and *stage3*. Best results are shown in bold.

Methods	IoU↑	F_β ↑	MAE↓	BER↓
DAA w/o Dm	0.733	0.835	0.045	9.92
DAA on RGB	0.738	0.838	0.044	9.62
DAA on Depth	0.729	0.831	0.048	9.85
DAA on CM	0.739	0.846	0.045	**9.30**
Ours	**0.742**	**0.853**	**0.043**	9.33

on RGB", "DAA on Depth", and "DAA on CM" as adopting the DAA module only on the RGB, depth, and cross-modal contextual features from the CCM modules, to test the effectiveness of the DAA module for extracting features in different modalities. Our final model adopts the DAA module with all three modalities. Experimental results show that the depth-missing information plays a key role in the DAA module, and our DAA module can effectively enhance the feature representation from different modalities.

Qualitative Evaluation

We further demonstrate our method visually in fig:visual, comparing it with five state-of-the-art methods.

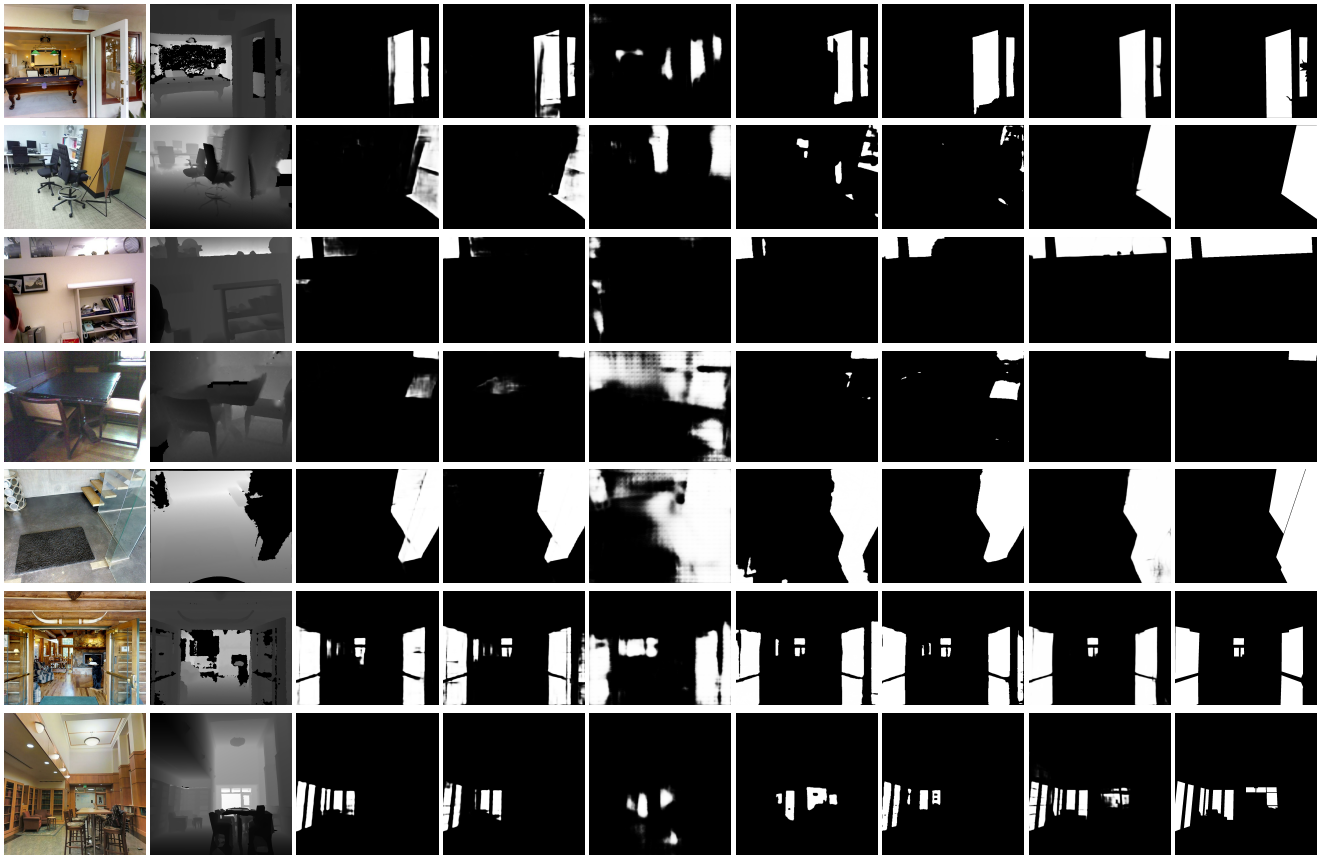

Figure 1: Visual comparison of our method with state-of-the-art methods on images from our RGB-D GSD dataset. CLNet (Zhang et al. 2021) and SPNet (Zhou et al. 2021) are RGB-D salient object detection methods, while GDNet (Mei et al. 2020), GSDNet (Lin, He, and Lau 2021), and EBLNet (He et al. 2021) are RGB-based glass surface detection methods.

References

Achanta, R.; Hemami, S.; Estrada, F.; and Susstrunk, S. 2009. Frequency-tuned salient region detection. In *CVPR*.

Armeni, I.; Sax, A.; Zamir, A. R.; and Savarese, S. 2017. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv e-prints*.

Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision (3DV)*.

He, H.; Li, X.; Cheng, G.; Shi, J.; Tong, Y.; Meng, G.; Prinet, V.; and Weng, L. 2021. Enhanced Boundary Learning for Glass-Like Object Segmentation. In *ICCV*, 15859–15868.

Lin, J.; He, Z.; and Lau, R. W. 2021. Rich Context Aggregation with Reflection Prior for Glass Surface Detection. In *CVPR*.

Mei, H.; Yang, X.; Wang, Y.; Liu, Y.; He, S.; Zhang, Q.; Wei, X.; and Lau, R. W. 2020. Don't Hit Me! Glass Detection in Real-World Scenes. In *CVPR*.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*, 567–576.

Zhang, J.; Fan, D.-P.; Dai, Y.; Yu, X.; Zhong, Y.; Barnes, N.; and Shao, L. 2021. RGB-D Saliency Detection via Cascaded Mutual Information Minimization. In *ICCV*.

Zhou, T.; Fu, H.; Chen, G.; Zhou, Y.; Fan, D.-P.; and Shao, L. 2021. Specificity-preserving RGB-D Saliency Detection. In *ICCV*.