Rich Context Aggregation with Reflection Prior for Glass Surface Detection

Jiaying Lin Zebang He Rynson W.H. Lau[†]
City University of Hong Kong

Abstract

Glass surfaces appear everywhere. Their existence can however pose a serious problem to computer vision tasks. Recently, a method is proposed to detect glass surfaces by learning multi-scale contextual information. However, as it is only based on a general context integration operation and does not consider any specific glass surface properties, it gets confused when the images contain objects that are similar to glass surfaces and degenerates in challenging scenes with insufficient contexts. We observe that humans often rely on identifying reflections in order to sense the existence of glass and on locating the boundary in order to determine the extent of the glass. Hence, we propose a model for glass surface detection, which consists of two novel modules: (1) a rich context aggregation module (RCAM) to extract multi-scale boundary features from rich context features for locating glass surface boundaries of different sizes and shapes, and (2) a reflection-based refinement module (RRM) to detect reflection and then incorporate it so as to differentiate glass regions from non-glass regions. In addition, we also propose a challenging dataset consisting of 4,012 glass images with annotations for glass surface detection. Our experiments demonstrate that the proposed model outperforms state-of-the-art methods from relevant fields.

1. Introduction

Glass surfaces are large glass regions such as windows, glass doors and glass walls. They are very popular and appear almost everywhere in our daily life. Although they are very useful in many ways, their existence can pose a serious problem to computer vision tasks. For example, if we are unable to detect the presence of glass surfaces, a robot or a drone may easily crash into glass surfaces. Thus, it is of vital importance to detect and separate glass surfaces from other objects for better scene understanding.

There have been some works focusing on the transparent object detection problem. As transparent objects tend to have certain shapes or thicker boundaries for defracting light, most existing methods are based on detecting shapes [24] or boundaries via polarization [13] and explicit edge supervision [30]. However, as glass surfaces typically do not process these properties, these methods developed for detecting transparent objects cannot be easily adopted to address the glass surface detection problem. Recently, Mei *et al*. [20] make the first attempt to develop the GDNet model for glass surface detection by integrating multi-level context information. While GDNet has a simple structure, it does not consider any specific properties of glass surfaces and its performance gain is mainly coming from a general context integrating structure based on dilation convolutions.

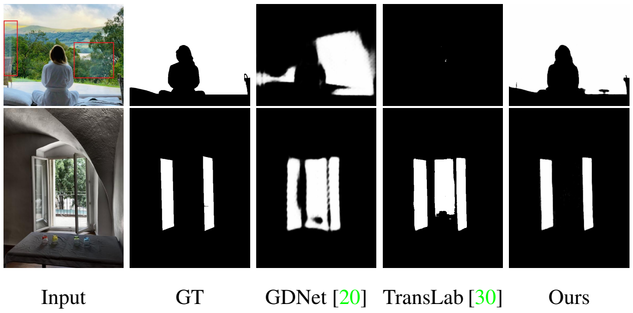


Figure 1: Two popular scenarios where existing methods [20, 30] fail. GDNet [20] is based on extracting/integrating abundant context features for glass surface detection. As it does not consider any specific glass properties, it tends to fail in scenes with insufficient contexts (e.g., top row where the glass surface covers almost the whole image) or with glass-lookalike regions (e.g., bottom row where the center region is not covered by glass). TransLab [30] is based on a boundary-guided network for transparent object detection. It also fails to detect glass surfaces correctly. Our method, which considers reflections and boundaries, can accurately detect the glass surfaces in these complex scenes.

Figure 1 shows two common but challenging scenarios. As the top image does not provide sufficient boundary context, GDNet fails to recognize the glass surface. As the middle region of the bottom image is not a glass surface but looks like one, GDNet gets confused and mis-recognizes it as a glass surface. TransLab [30] also fails to detect the glass surface in the top row as the glass boundary is not obvious, and mis-recognizes the non-glass region in the bottom image as a glass surface.

[†] Rynson Lau is the corresponding author, and he leads this project.

In this paper, we aim to address this challenging glass surface detection problem. Our observations are that humans often rely on the identification of reflections in order to be aware of the presence of glass, and the localization of the boundary in order to determine the extent of the glass surface. These two observations motivate us to explore two visual cues, *i.e.*, glass reflections and boundaries, for glass surface detection. Based on these two observations, we propose a new model for glass surface detection with two novel modules. We propose a Rich Context Aggregation Module (RCAM) for multi-scale boundary feature extraction, and a Reflection-based Refinement Module (RRM) for detecting glass reflections to help determine glass regions. Note that unlike TransLab [30], which requires explicit boundary maps for guiding their network training, our RRM aims to exploit context contrasted features to flexibly locate glass boundaries at different sizes and shapes in different context levels. Figure 1 demonstrates the superiority of the proposed model. Even though the glass surface boundary in the first image is not obvious, our model can detect the glass region fairly accurate. It can also accurately detect the two glass doors but not the middle non-glass region.

In addition, we note that although a glass dataset (GDD) is proposed by Mei *et al*. [20], the images in it are collected from limited scenes and captured mainly based on close-up shots. This can significantly limit the generalization performance of the trained model. To address this limitation, we propose a challenging glass surface dataset that contains close-up, normal and long shots, from diverse scenes with glass surfaces. Our dataset contains a total of 4,012 glass images with corresponding glass surface masks. It covers a diversity of indoor and outdoor scenes. We have conducted extensive experiments to evaluate our model and show that the proposed model outperforms state-of-the-art methods on both GDD and our proposed dataset.

Our main contributions can be summarized as follows:

- We propose a novel model that consists of a Rich Context Aggregation Module (RCAM) for multi-scale boundary features extraction and a Reflection-based Refinement Module (RRM) to extract glass reflections for glass surface detection.
- We have constructed a challenging glass surface dataset, which consists of 4,012 real-world images with glass surface masks, from diverse scenes.
- Our extensive experiments demonstrate the superiority of the proposed model over state-of-the-art methods from relevant fields.

2. Related Work

In this section, we mainly discuss various types of detection works that are relevant to our problem.

2.1. Glass Surface Detection

Recently, Mei *et al*. [20] propose the first computational method and the GDNet model for glass surface detection. It uses a large-field contextual feature integration module to capture low-level and high-level contexts. Although this module is shown to be useful, it does not consider any specific glass properties in the detection. As a result, it can be easily confused by regions that look like glass surface, *e.g*., an opened window.

To address the limitation of this work, we propose in this paper to incorporate two important properties of glass surfaces to address the glass surface detection problem. We first learn glass surface boundaries and then detect reflections within them in order to guide the prediction.

2.2. Transparent Object Detection

There has been some research interest in recent years in addressing a related problem to glass surface detection, transparent object detection. Early works [8, 9] try to detect specific transparent objects, *e.g*., wine glass and glass bottles, using specially designed transparent local patch features or image gradients. [33, 10] focus on general transparent object detection with the help of light-field or RGB-D cameras. Recently, Kalra *et al*. [13] combine polarization with deep learning and propose a polarized CNN for transparent object detection. Comparing with previous methods that require additional input data, Xie *et al*. [30] propose a boundary-guided network to detect transparent objects from a single RGB image, and a transparent object dataset for training.

In short, most methods for transparent object detection leverage multi-modal data and achieve good results based on detecting the shape and boundary of transparent objects. However, unlike transparent objects, glass surfaces usually do not have well-defined shapes and their boundaries are often ambiguous, resulting in the glass surface detection problem being very challenging.

2.3. Salient Object Detection.

Salient object detection aims to detect objects that are most salient to humans. While early methods depend heavily on hand-crafted features and saliency priors [34, 42], recent methods mostly use CNNs to enhance feature extraction and aggregation [11, 38, 25]. He *et al*. [11] proposes a CNN to extract contrast information from superpixels. Wang *et al*. [25] propose a novel pyramid pooling module as well as a multi-stage refinement mechanism to capture detailed spatial information. Zhang *et al*. [38] propose a progressive attention based network for adaptive multi-scale context integration. More recently, BASNet [23] and PAGE-Net [26] leverage boundaries of salient objects to encourage a finer segmentation. Pang *et al*. [22] propose a multi-scale aggregation interaction module to utilize multi-scale features in adjacent levels for salient object detection.

While these methods may perform well in the salient object detection problem, they are not suitable for addressing the glass surface detection problem as the content within a glass surface may not necessarily be salient.

3. Glass Surface Detection Dataset

Although Mei *et al*. [20] propose a glass dataset (GDD) for the glass surface detection problem, it contains primarily close-up shots. Figure 2(a) shows some of these images. To address the limitations of GDD, we have constructed a large-scale glass surface dataset, named **GSD**, which includes 4,012 real images with glass surfaces and corresponding masks. Our proposed dataset contains close-up, medium and long shots from diverse scenes.

Dataset construction. To compile our glass surface dataset, we have collected about half of the images from existing datasets [29, 40, 5, 17] as well as from the Internet (which are under the Creative Commons licenses). The rest are captured by ourselves using several smartphones. The images in our dataset cover diverse scenes, including bathrooms, sitting rooms, shops, classrooms, museums, streets and entrances. After collecting all the images, we then use Labelme[2] to manually create the glass surface masks. Figure 2(b) shows some example glass surface images and the corresponding masks from our dataset. Note that although existing semantic segmentation datasets [40, 21, 5, 2] have fine annotations of various objects, we are not able to use them for two reasons. First, existing datasets seldom explicitly consider the glass surface category. Second, although some of them may contain some related categories, *e.g*., windows and screen doors, we cannot use their annotations as they are rather coarse and mixed with non-glass parts like handles and window frames.

To split the dataset into training set and test set, We randomly split them into a training set with 3,202 images and a test set with 810 images.

Dataset analysis. To give a better understanding of our glass dataset, we conduct statistical analyses on GSD as:

- **Area Distribution.** It is defined as the ratio of the glass area over the image area. As shown in Figure 3(a), our dataset contains glass covering a wide range of area ratios. In general, images that fall in the range of [0, 0.3] contain glass that is relatively far away from the camera. These images tend to provide more context information, as more surrounding objects can be seen. Images that fall in the range of [0.7, 1] contain glass that is relatively close to the camera. These images provide very little context information, and are more challenging for glass surface detection.
- **Contrast Distribution.** Due to the intrinsic transparency property of glass, the content within a glass surface shares very similar semantics as the content behind the glass, and sometimes also the content around it. Here, we analyze the contrasts between glass regions and non-glass regions by computing χ^2 distance between their RGB histograms, similar to [15]. We also compare the distribution with GDD [20], as shown in Figure 3(b). In general, GSD has more images with low color contrasts (< 0.4), and less images with high color contrasts (> 0.4), compared with GDD. This means that our dataset has lower global color contrasts, making it more challenging to detect.
- **Perceptual Similarity.** Perceptual similarity can affect the quality of the dataset [16]. Normally, images with higher perceptual similarity may have similar contexts. When these images are used for network training, it would reduce the robustness of the trained model. Table 1 shows that our dataset has a lower average learned perceptual image patch similarity (LPIPS) [36], compared with GDD [20].
- **Shape Complexity.** Objects with complex topology can be challenging to detection [26, 28]. Table 1 shows that shapes of the glass surfaces in our GSD dataset are more complex than those from GDD [20].

Table 1: Perceptual similarity and shape complexity analysis of our proposed GSD dataset, compared with GDD [20].

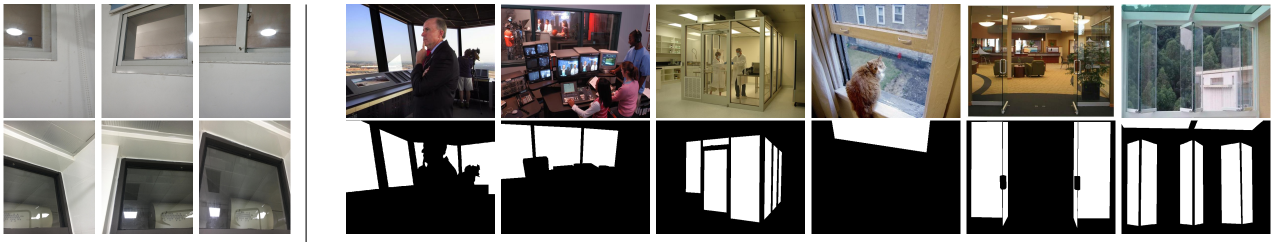
Dataset	Num.	LPIPS [36]	Shape Complexity [19]
GDD [20]	3900	0.76582	1.3420
Ours (GSD)	4102	**0.75870**	**1.9577**

4. The Proposed Method

Our method is based on two observations. We observe that humans often try to identify reflections in order to sense of the existence of glass surfaces, and to localize the boundary (*e.g*., window frames) in order to determine the extent of the glass surface. These observations motivate us to explore two visual cues, *i.e*., glass reflection and boundary, in designing our model for glass surface detection. Figure 4 shows the pipeline of our proposed model. It takes a single image as input and outputs a binary glass mask.

In our model, we first feed the input image to a backbone network [31] to extract multi-scale backbone features. Specifically, we use the outputs of five stages of the network, *i.e*., *conv1*, *res2c*, *res3b3*, *res4b22* and *res5c*, as our backbone features. The deepest features from the final stage (*i.e*., *res5c*) are first fed into the proposed rich context aggregation module (RCAM) to capture multi-scale boundary features, which are then fed into a decoder to generate

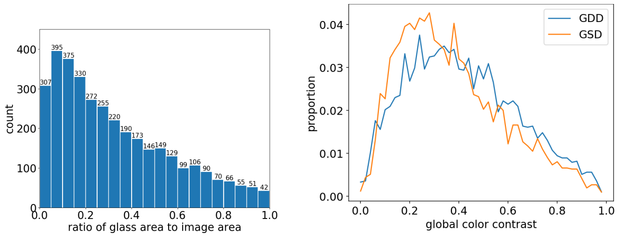
[2] https://github.com/wkentaro/labelme



(a) GDD

(b) GSD

Figure 2: Comparison between GDD [20] and GSD. While GDD mainly includes close-up shots (a), GSD contains close-up, medium and long shots. GSD also covers a diversity of daily-life scenes with glass surfaces of complex shapes (b).



(a) glass area distribution

(b) color contrast distribution

Figure 3: Statistics of our GSD dataset.

a coarse binary mask indicating the location of glass surfaces. This coarse mask serves as an attention map to the previous stage (*i.e*., *res4b22*) to guide the refinement of the glass mask. In this way, the glass mask as well as the glass boundary can be progressively refined by integrating with the earlier backbone features.

After obtaining the finest glass mask, we finally feed tit to the proposed reflection-based refinement module (RRM). This RRM takes as input the concatenation of the backbone features from the first stage, *i.e*., *conv1*, the input image, and the finest glass mask from the first stage. It detects reflections in the input image and then leverages the detected results to refine the input glass mask to produce an output glass surface mask.

4.1. Rich Context Aggregation Module (RCAM)

To localize the glass surface boundaries, we propose the RCAM, which aims to extract the boundary features of glass surfaces using contrasted features from multi-scale contexts. The multi-scale processing is to handle glass boundaries of different scales. Figure 5 shows the structure of the RCAM. It is built upon the basic CCF block, in a way similar to [35], to learn contextual contrasted features ($\mathbb{CCF}$) with different dilation rates r_i and r_j, where r_i $<r_j$. However, unlike [35] which focuses on learning the contrasts between inside and outside of mirrors for mirror detection, we concentrate on extracting information around the glass surface boundaries.

Our RCAM consists of two cascaded stages:

- **Pairing.** Given the input features, we first obtain the corresponding multi-scale contextual features using multi-rate atrous convolutions, which share some similarity with ASPP [4]. However, instead of directly fusing the features with concatenation as in [4], we first separately compute the contrasted features by subtracting contextual features at different scales of all permuted pairs of the multi-scale contextual features. All contrasted features are then concatenated together and fed to the selection stage.

- **Selection.** Since different channels of the features convey different degrees of semantics and different scales of the contrasted features contribute differently depending on the actual size of the glass surface, we employ a two-level attention mechanism in our selection stage to highlight different channels of contrasted features. Unlike the SENet [12], which allocates weights across channels without explicitly considering the semantic similarities among the channels, we decouple the allocation process in a top-down way by grouping channels based on the scales of contrasted contexts. To this end, we employ two independent sub-networks to adaptively enhance the features with context-wise attention and channel-wise attention.

Our proposed RCAM structure takes advantage of all permutations of context contrasted features to segment glass surfaces of any sizes and scales. To cover different scales, we set the dilation rate to 1, 2, 4, and 8, such that we have a total of $\binom{n}{2}$ $\mathbb{CCF}$s. Combining with the original multi-scale contextual features, we can obtain rich contextual contrasted features ($\mathbb{RCCF}$) for selection.

Formally, given the input feature $\mathbb{F} \in \mathbb{R}^{C\times H\times W}$, we can extract a series of context contrasted features, $\mathbb{CCF}_{(r_i,r_j)}$, corresponding to atrous convolutions with dilate rates r_i and r_j, and then concatenate them with the original multi-scale contextual features to obtain rich contextual contrasted features, $\mathbb{RCCF}$. For each $\mathbb{RCCF}_i$, we compute its context-

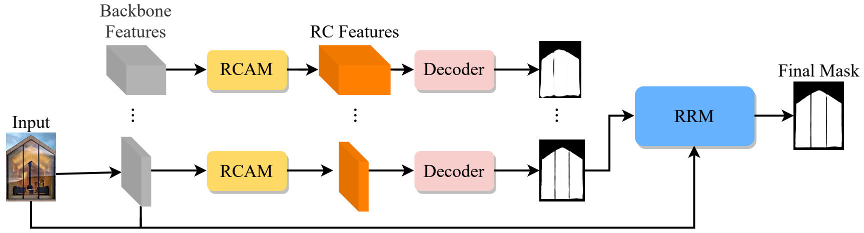


Figure 4: The pipeline of our proposed method. We first use ResNeXt-101 [31] as a backbone to extract multi-scale representations (gray blocks). We then embed a novel RCAM (yellow blocks) in different layers of the backbone to learn multi-scale glass boundary information by extracting corresponding rich context features. Finally, we use a novel RRM (blue block) to extract reflection information to guide the prediction of the output glass surface mask.

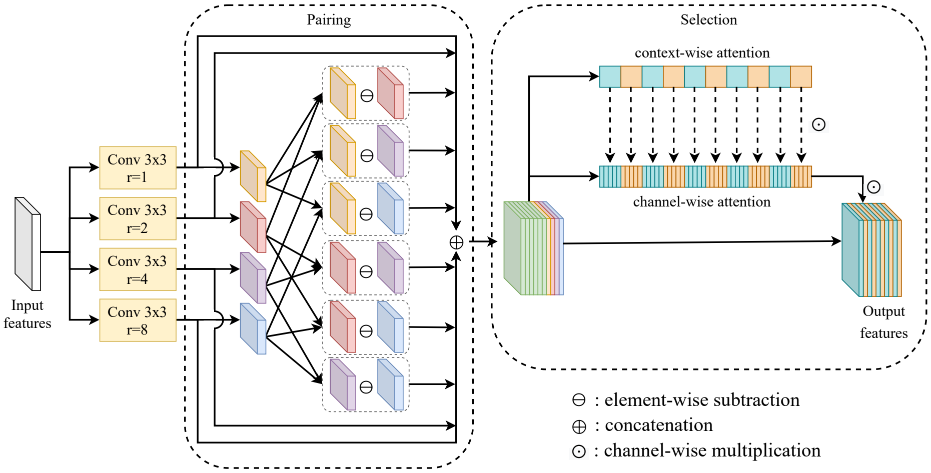


Figure 5: The structure of our proposed RCAM. We first use a series of atrous convolutions with increasing dilation rates to extract multi-scale context features. We then compute contrasted features from all permuted pairs of different context scales, allowing the detection of glass boundary of any size. (Note that different scales of contrasted features are denoted by different colors.) We also propose to use a two-level attention mechanism to explicitly highlight context-wise and channel-wise attentions.

wise attention, $\alpha_i \in [0,1]$, which is shared across all channels within the same context, and its corresponding channel-wise attention, $\beta_i \in [0,1]^C$. The resulting attentive $\widetilde{\mathbb{RCCF}}$ is obtained by re-scaling with two levels of attention as:

$$\widetilde{\mathbb{RCCF}}_i = \alpha_i \cdot (\mathbb{RCCF}_i \odot \beta_i), \qquad (1)$$

where $\odot$ denotes channel-wise multiplication.

Note that our proposed RCAM is different from the original CCL [6] and from the CCFE [35]. The CCL block in [6] only employs one scale context contrast to produce multi-level CCL features in a chained form, while we use multiple atrous convolutions with increasing dilation rates in parallel to capture multi-scale context contrast features in a single level. The CCFE block in [35] consists of four independent branches to focus on contrasts between local and context features. Instead, we emphasize more on the contrasts among multi-scale contexts by considering all permutations using a two-level attention mechanism, which allows us to effectively detect glass boundaries of any scales.

4.2. Reflection-based Refinement Module (RRM)

Only using contextual information for glass surface detection is insufficient especially when the glass surface boundaries are missing or ambiguous. Thus, it is necessary to utilize an additional strong cue to facilitate glass surface detection. Due to the intrinsic reflective property of glass, reflections can often be observed when light is reflected off a glass surface. In consideration of this reflection cue, we propose a novel reflection-based refinement module (RRM) to refine the glass mask guided by the detected reflection. Due to the lack of ground-truth reflection inside the glass surface, we generate a reflection map from each glass surface using an existing reflection separation model [37] in combination with our ground-truth glass mask G.

Specifically, given an input image, we first detect the reflection R_{global} within the glass surface using [37]. We then force the RRM to learn the reflection R_{glass} given the ground-truth glass mask G, during training process. The details of training RRM are introduced in Section 4.3. The RRM has an encoder-decoder architecture. We add skip connections between the counterpart layers of the encoder and decoder to facilitate the training process. It takes as input the concatenation of the input RGB image, backbone features and the predicted glass mask from the first stage (*i.e.*, *conv1*), and outputs a refined mask.

In contrast with the reflection separation model [37] and the reflection removal model [27], which mainly deal with global reflections (*i.e.*, the whole input image is expected to be covered by glass), the RRM aims to locate where the glass surfaces are through detecting local reflections.

4.3. Loss Functions

Our proposed network can be trained to simultaneously predict glass surface reflections and a glass segmentation mask in an end-to-end manner. For the glass surface reflection prediction, we use Mean-Square-Error (MSE) to optimize our RRM. For the glass mask prediction, we choose the lovász-hinge loss [3] for our glass surface detection to directly optimize the mean intersection-over-union (mIoU). In addition, deep supervision [32] is also introduced in our training process as:

$$Loss = \lambda \left\| R_{global} \otimes G - R_{glass} \otimes G \right\|^2 + \sum_{i=1}^{N} w_i L_i, \quad (2)$$

where R_{glass} denotes the predicted reflection map. R_{global} is the ground-truth global reflection map. G is the ground-truth glass surface mask. L_i is the lovász-hinge loss on the i-th stage of the predicted glass mask and N refers to the five stages of the backbones. λ and w_i are weight parameters. $\otimes$ is an element-wise multiplication operation.

5. Experiments

Implementation Details. We use the ResNeXt-101 [31] pre-trained on ImageNet as our backbone, and the remaining layers are initialized randomly with the default setting in PyTorch. We use stochastic gradient descent (SGD) as the optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. We adopt the "Poly" decay strategy [18], where the current learning rate is the base learning rate multiplied by $(1 - \frac{current_{iter}}{max_{iter}})^{Power}$. The base learning rate is 0.001 and *Power* is 0.9. The batch size is 6. We run 80 epochs for training. To prevent the over-fitting problem, input images are first resized to 400 × 400 and randomly cropped to 384 × 384 patches for training. Randomly horizontal flipping is also considered in our experiments. Our model takes about 10 hours to converge, and runs at $\sim 37 fps$ on a RTX 2080Ti GPU card. During inference, the test images are also first resized to 384 × 384 before feeding into the network. The outputs from our network are further refined using CRF [14], by exploring spatial pixel coherence.

Evaluation Datasets. We evaluate our proposed method on two glass surface detection datasets: GDD [20] with 936 test images and our proposed dataset GSD with 813 test images. We also select 2,047 training images and 1771 testing images with transparent surfaces from the transparent object detection dataset Trans10K [30] for evaluation. All methods are trained and tested on the training/testing splits from the same dataset.

Evaluation Metrics. We consider the intersection over union (IoU), Mean Absolute Error (MAE), maximum F-measure (F_β), and balance error rate (BER) as evaluation metrics, which are widely used in computer vision tasks. MAE is the average pixel-wise error between the predicted mask and ground truth as:

$$MAE = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} |P(i,j) - G(i,j)|, \quad (3)$$

where P is the predicted mask and G is ground truth. H and W are the width and height of the input image. F_β is used to evaluate the overall performance with a trade-off between precision and recall, and is defined as:

$$F_\beta = \frac{1 + \beta^2 (Precision \times Recall)}{\beta^2 Precision + Recall}, \quad (4)$$

where β is set to 0.3 as suggested in [1].

To quantitatively evaluate the performance of glass surface segmentation, we also report the BER score as:

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

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

5.1. Comparison to the State-of-the-art Methods

We compare our proposed methods with the state-of-the-art methods from relevant fields, including BASNet [23] and MINet [22] for saliency object detection, BDRAR [41] for shadow detection; PSPNet [40] from semantic segmentation, SINet [7] for camouflaged object detection, TransLab [30] for transparent object segmentation, and GDNet [20] for glass surface detection. For PSPNet [40], we restrict it to output binary classification (*i.e.*, glass or non-glass regions). We use the publicly available codes of these models with default configurations. Table 2 shows the comparison on four metrics: intersection over union (IoU), F-measure (F_β), mean absolute error (MAE), and balance error rate (BER). We can see that our method outperforms all existing state-of-the-art methods by a large margin.

Table 2: Quantitative results. We compare our model with relevant state-of-the-art models: BASNet [23] and MINet [22] for salient object detection, BDRAR [41] for shadow detection, PSPNet [40] for semantic segmentation, SINet [7] for camouflaged object detection, TransLab [30] for transparent object segmentation, and GDNet [20] for glass surface detection. We use their publicly available codes with default configurations. All methods are trained and tested on the training/testing splits from the same dataset. We can see that our dataset is more challenging than GDD. Best results are shown in bold.

Methods	Trans10K (ECCV 20')				GDD (CVPR 20')				GSD (Ours)			
	IoU↑	F_β ↑	MAE↓	BER↓	IoU↑	F_β ↑	MAE↓	BER↓	IoU↑	F_β ↑	MAE↓	BER↓
PSPNet [39]	67.83	0.904	0.084	15.40	79.16	0.875	0.132	11.51	70.26	0.834	0.110	10.66
BDRAR [41]	57.34	0.640	0.188	16.71	80.01	0.908	0.098	9.87	75.91	0.860	0.081	8.61
BASNet [23]	80.63	0.899	0.087	10.07	80.78	0.891	0.106	9.37	69.79	0.808	0.106	13.54
MINet [22]	75.84	0.924	0.064	10.11	84.35	0.919	0.077	7.40	77.29	0.879	0.077	9.54
SINet [7]	83.53	0.906	0.066	8.15	83.27	0.912	0.101	8.35	77.04	0.875	0.077	9.25
GDNet [20]	84.46	0.916	0.068	6.50	81.42	0.909	0.097	8.83	79.01	0.869	0.069	7.72
TransLab [30]	86.13	0.916	0.055	5.86	82.93	0.891	0.091	8.87	74.05	0.837	0.088	11.35
Ours	**89.16**	**0.937**	**0.043**	**4.50**	**88.07**	**0.932**	**0.059**	**5.71**	**83.64**	**0.903**	**0.055**	**6.12**

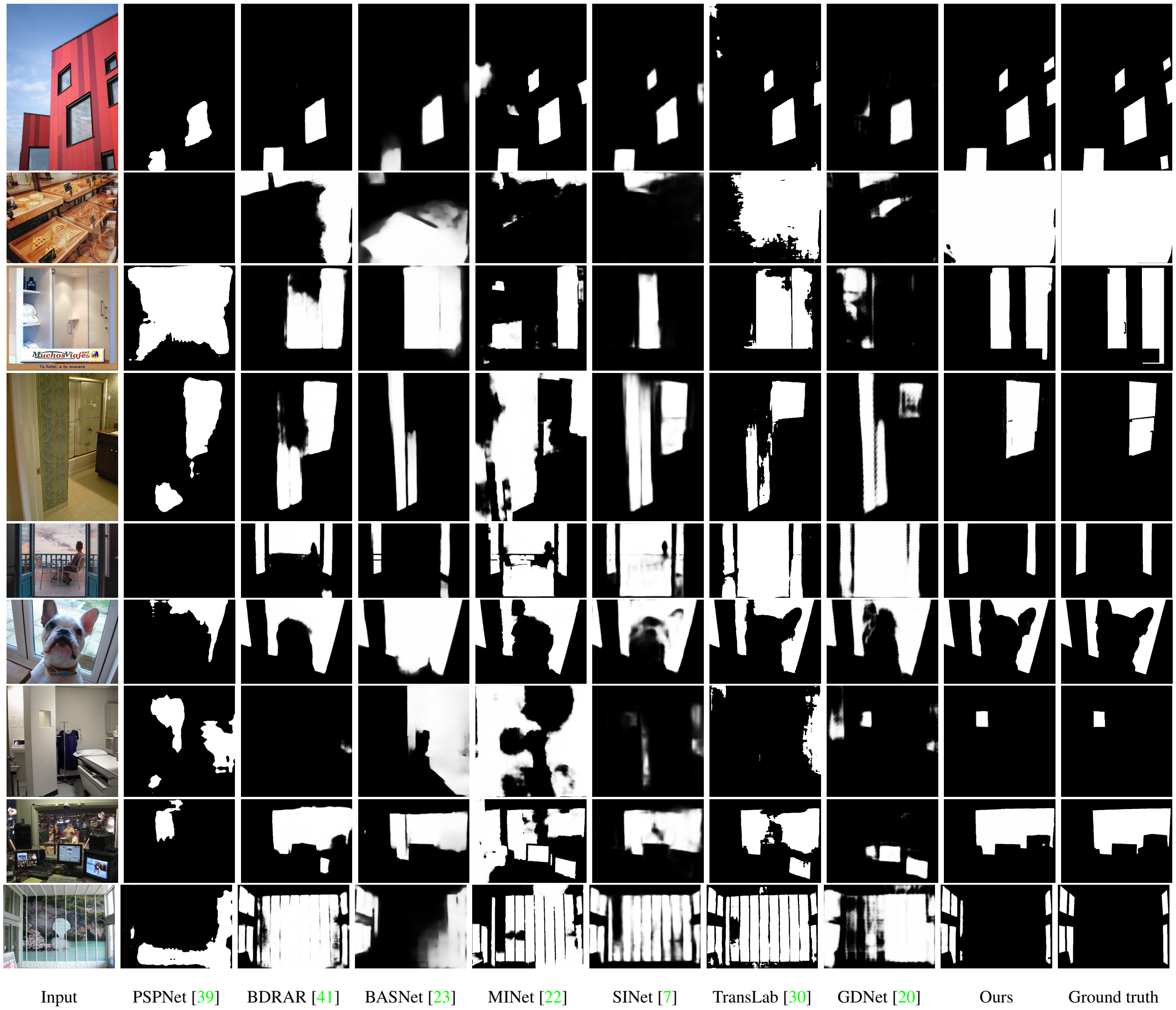


Figure 6: Visual comparisons of our method to state-of-the-art methods on some example images.

Figure 6 shows visual comparisons on some challenging images. We can see that the state-of-the-art methods may under-detect the glass regions in some images (*e.g.*, some windows in the $1st$ row and various parts of the window in the $2rd$ row) but over-detect the glass regions in other images (*e.g.*, the non-glass region of the shower room in the $4th$ row, the green wall in the $5th$ row, and the center region of the last row). In contrast, our method can effectively detect the glass regions of different sizes and shapes, and can differentiate some ambiguous non-glass regions in some challenging images (*e.g.*, $3rd$ to $9th$ rows). We attribute the superior performances of our method on these challenging examples to the use of reflection by the RRM.

5.2. Ablation Study

To verify the effectiveness of RCAM and RRM, we first implement a baseline with only the original backbone network [31] ("basic"), and then three alternative approaches built on the baseline. The first one includes RCAM but not the proposed two-level attention mechanism ("basic + PCM"). The second one includes RCAM ("basic + RCAM"). Note that these two alternatives do not include the RRM. The third one includes the RRM but not the RCAM ("basic + RRM").

Table 3 shows the experimental results. We observe that using only the backbone network for glass surface detection performs the worst among all ablated models. We may also observe that adding the two-level attention mechanism (*i.e.*, "basic + RCAM") performs better than the other two alternatives (*i.e.*, "basic + PCM" and "basic + RRM"). While "basic + RCAM" performing better than "basic + PCM" demonstrates the importance of the two-level attention mechanism, "basic + RCAM" performing better than "basic + RRM" is mainly because some of the glass surfaces in our GSD dataset may not contain obvious reflections. The full model, on the other hand, performs the best among all the ablated models, demonstrating that both RCAM and RRM can benefit each other in the glass surface detection task. Figure 7 shows a visual example where the RRM can play an important role in resolving an ambiguous scene.

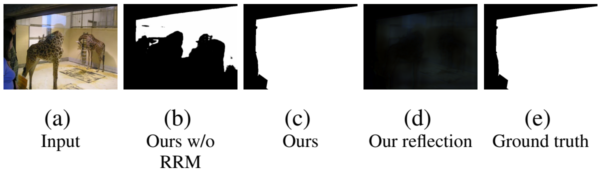

(a) Input
(b) Ours w/o RRM
(c) Ours
(d) Our reflection
(e) Ground truth

Figure 7: A visual example of the ablation study. Without RRM, the model under-detects the glass region in (b). RRM helps detect various degrees of reflection in the glass in (c). By including the RRM, our proposed model can correctly detect the glass region.

Table 3: Component analysis, trained and tested on GSD. "Basic" denotes the original backbone network, with RCAM and RRM removed. "RCAM" is the rich context aggregation module. "PCM" is the "RCAM" without the two-level attention mechanism. "RRM" is the reflection-based refinement module. Our final model includes both RCAM and RRM. Best results are shown in bold.

Methods	IoU ↑	F_β ↑	MAE ↓	BER ↓
basic	80.12	0.881	0.073	8.11
basic + PCM	81.56	0.878	0.073	7.93
basic + RCAM	82.36	0.898	0.062	7.19
basic + RRM	81.11	0.889	0.071	7.84
Ours	**83.64**	**0.903**	**0.055**	**6.12**

6. Conclusion and Future Work

In this paper, we have investigated the glass surface detection problem. To this end, we have made the following contributions. First, we have constructed a large-scale glass dataset, which contains 4,012 glass images. This dataset covers a wide range of daily scenes and includes manually annotated glass surface masks. Second, we have proposed a model for glass surface detection, which includes two novel modules: the Rich Context Aggregation Module (RCAM) for extracting multi-scale boundary features to locate glass surface boundaries of different scales and shapes, and the reflection-based refinement module (RRM) to detect reflection for helping differentiate glass from non-glass surfaces. Finally, we have conducted extensive experiments to evaluate the performance of the proposed model against the state-of-the-art methods. Our results demonstrate the superiority of the proposed model on the glass surface detection task.

Our work also has limitations. If the reflection on the glass surface is too weak to be detected by our RRM, our model may not be able to detect the glass surface correctly, as shown in Figure 8. We note that in this case, it is difficult even for humans to correctly identify where the glass surfaces are. As a future work, we are currently investigating ways to learn a stronger reflection detector to help detect very weak reflections.

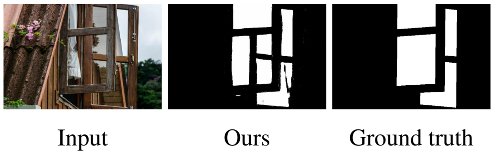

Input
Ours
Ground truth

Figure 8: Failure cases. Our method may fail to detect glass in some very challenging scenes, where there is a lack of reflection for the model to correctly detect the glass surfaces.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009. 6

[2] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *CVPR*, 2015. 3

[3] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, 2018. 6

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 2017. 4

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 3

[6] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *CVPR*, 2018. 5

[7] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, 2020. 6, 7

[8] Mario Fritz, Gary Bradski, Sergey Karayev, Trevor Darrell, and Michael J Black. An additive latent feature model for transparent object recognition. In *NeurIPS*, 2009. 2

[9] Jiaqi Guo. Transparent object recognition using gradient grids. 2

[10] Chen Guo-Hua, Wang Jun-Yi, and Zhang Ai-Jun. Transparent object detection and location based on rgb-d camera. In *JPCS*, 2019. 2

[11] Shengfeng He, Rynson Lau, Wenxi Liu, Zhe Huang, and Q. Yang. Supercnn: A superpixelwise convolutional neural network for salient object detection. *IJCV*, December 2015. 2

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 4

[13] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *CVPR*, June 2020. 1, 2

[14] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, 2011. 6

[15] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, 2014. 3

[16] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau. Progressive mirror detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020. 3

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[18] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. *arXiv:1506.04579*, 2015. 6

[19] Tracy McLellan and John A Endler. The relative success of some methods for measuring and describing the shape of complex objects. *Systematic Biology*, 47(2):264–281, 1998. 3

[20] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020. 1, 2, 3, 4, 6, 7

[21] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 3

[22] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, June 2020. 2, 6, 7

[23] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019. 2, 6, 7

[24] Shreeyak Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *ICRA*, pages 3634–3642. IEEE, 2020. 1

[25] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In *ICCV*, 2017. 2

[26] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven CH Hoi, and Ali Borji. Salient object detection with pyramid attention and salient edges. In *CVPR*, 2019. 2, 3

[27] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *CVPR*, 2019. 6

[28] Changqun Xia, Jia Li, Xiaowu Chen, Anlin Zheng, and Yu Zhang. What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors. In *CVPR*, pages 4142–4150, 2017. 3

[29] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 3

[30] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In *ECCV*, 2020. 1, 2, 6, 7

[31] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 3, 5, 6, 8

[32] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015. 6

[33] Yichao Xu, Hajime Nagahara, Atsushi Shimada, and Rin-ichiro Taniguchi. Transcut: Transparent object segmentation from a light-field image. In *ICCV*, 2015. 2

[34] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 2

[35] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson W.H. Lau. Where is my mirror? In *ICCV*, 2019. 4, 5

[36] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3

[37] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *CVPR*, 2018. 5, 6

[38] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *CVPR*, 2018. 2

[39] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 7

[40] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 3, 6, 7

[41] 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, 7

[42] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *CVPR*, 2014. 2