Exploiting Semantic Relations for Glass Surface Detection

Anonymous Author(s)
Affiliation
Address
email

Abstract

Glass surfaces are omnipresent in our daily lives and often go unnoticed by the majority of us. While humans are generally able to infer their locations and thus avoid collisions, it can be difficult for current object detection systems to handle them due to the transparent nature of glass surfaces. Previous methods approached the problem by extracting global context information to obtain priors such as boundary and reflection. However, their performances cannot be guaranteed when these critical features are not available. We observe that humans often reason through the semantic context of the environment, which offers insights into the categories of and proximity between entities that are expected to appear in the surrounding. For example, the odds of co-occurrence of glass windows with walls and curtains is generally higher than that with other objects such as cars and trees, which have relatively less semantic relevance. Based on this observation, we propose a model that integrates the contextual relationship of the scene for glass surface detection with two novel modules: (1) Scene Aware Activation (SAA) Module to adaptively filter critical channels with respect to spatial and semantic features, and (2) Context Correlation Attention (CCA) Module to progressively learn the contextual correlations among objects both spatially and semantically. In addition, we propose a large-scale glass surface detection dataset named GSD-S, which contains 4,519 real-world RGB glass surface images from diverse real-world scenes with detailed annotations. Experimental results show that our model outperforms contemporary works, especially with 48.8% improvement on MAE from our proposed GSD-S dataset.

1 Introduction

Glass objects, in particular those with large specular surfaces, are becoming prevalent in daily lives, as can be seen on numerous occasions, including glass doors, windows, and walls of modern architecture. Autonomous systems of contemporary works typically lack the ability to identify glass objects due to the ambiguity of the transparency property. With such characteristics, the peripheral scene that the glass surface displays is merely the opaque objects and scenes from the surroundings. These result in a myriad of potential dangers due to the impairment of the models' object detection capability, as manifested in previous works [1, 2]. Consequently, it brings forth a pressing need for a better glass detection model. Existing methods have explored many physical characteristics of glass surface objects including boundary edge [3, 4, 5], polarization [6], surface normal [7] and reflection [8]. Although these methods perform satisfactorily well, under the circumstances that the assumed object properties do not exist, the detection ability of the model is greatly impeded. Additionally, the object segment prediction in terms of binary classification introduces a bottleneck for the model on knowledge learning, as opposed to other multi-class models that encourage knowledge acquisition through a more diverse pool of knowledge.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

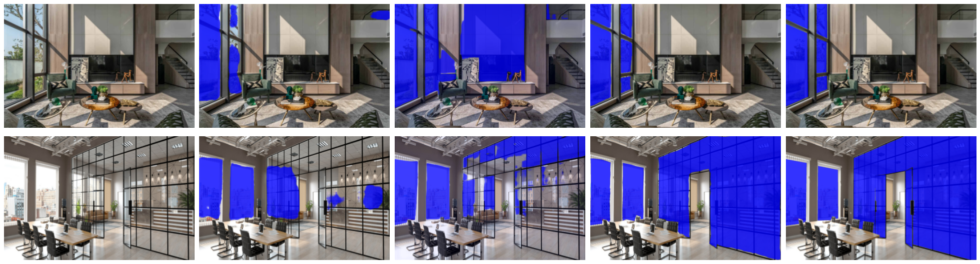


Figure 1: Visual comparisons of our method to state-of-the-art methods for glass surface detection [8, 18] on some example images.

In this project, we observed that there exist certain implicit relationships according to the surrounding scene context. For example, glass surfaces of ‘windows’ have a recurrent appearance along with ‘curtains’ or ‘blinds’. Moreover, ‘glass doors’ have a high likelihood of conjunctions with ‘wall’. It prompts us to reconsider the problem from a new perspective by incorporating knowledge on top of superficial features. We hereby propose to focus more on syntactic context learning. Studies in cognitive neuroscience [9, 10, 11, 12] demonstrate that surrounding objects can offer an effective derivation of contextual information. Furthermore, the application of contextual information has had many proven successful impacts. Examples include tasks for salient object detection [13], domain adaptation [14], network prunning [15], image style transfer [16], image-text retrieval [17] etc. Based on these and our findings, we initiate to incorporate the intrinsic expositions of identified objects and the underlying relationships between target entities within the environment.

Figure 1 illustrates the robustness of our model in the cases of obscure scenes wherein explicit physical cues such as edges and reflections are misleading. Existing methods [8, 18] are easily obfuscated and mistake seemingly likely areas to be false positive due to the superficial characteristics such as boundary and reflection cues. In the upper row, the shaded area on the wall near windows enclosed by frame shadows gives an illusion of a bounded area that resembles glass frames. The boundary information becomes a distraction rather than a cue that aids prediction. Furthermore, the window wall in the lower row has a considerably large size, with its boundary exceeding beyond the visible scene of the picture; boundary information would not be of use either in this case. Existing models instead bypassed the large ‘window wall’ and classified the windows behind in another room to be a candidate ground-truth label. Although this is not entirely unreasonable, the ‘window wall’ closer to the front should be detected considering its proximity.

To address these issues, we propose a novel model exploiting semantic relations for glass surface detection. Our model adopts the encoder-decoder architecture. The encoder component is supported by two backbones: 1) SegFormer network [19] for comprehensive spatial context learning and 2) ResNet50 network [20] for semantic relationship learning. The semantic submodule is first trained with regular segmentation modeling to grasp scene context reference and object relationships. Specifically, two modules are proposed to achieve the ontology learning objectives to assist glass surface detection: 1) Scene Aware Activation (SAA) Module to guide contextual feature modelling, and 2) Context Correlation Attention (CCA) Module to associate spatial context and semantic meanings of objects in the environment. Inspired by SENet [21], the SAA Module consists of two feature selection pathways that respectively assimilate the information concerning object locations and object categorical connotations. The CCA Module adopts Transformer block [22] to conduct attention modeling between the extracted backbone features. Figure 1 shows the superior performance of our method.

Besides, we notice that although Mei *et al*. [18] and Lin *et al*. [8] both propose datasets for glass surface detection, these datasets do not contain semantic data (e.g., semantic labels of different objects around glass surfaces) to model the spatial context and high-level scene context. To further the research on glass surface detection, we propose a new large-scale challenging semantic-aware glass surface dataset (GSD-S) with ground truth semantic labels, not only limited to the binary masks of glass surfaces. Our dataset consists of 4,519 images collected from various scenes. Our dataset is larger than those proposed by Mei *et al*. [18] (3,900 images) and Lin *et al*. [8] (4,102 images) and can largely facilitate the research on this area. Exhaustive experimentation on all three datasets was conducted to validate the enhanced performance of our model.

Our contributions can be summarized as follows:

- We initiate an under-explored strategy to apply semantic relationship modeling to infer the connections between glass objects and everyday objects for glass surface detection.
- We present two novel deep learning modules for cross relationship modeling to capture long-range spatial and implicit semantic dependencies, with results being substantiated by thorough studies.
- We built a large-scale dataset with complex and challenging scenes that consider semantic contexts and serve as a benchmark for performance validation on future models.

2 Related Work

Transparent Object Detection. Transparent Object Detection aims to identify glass-made objects, specifically with bounded shapes such as glasses and glass bottles; occasionally, the task also accommodates window panels. Existing works approached this task by leveraging the bounded shapes to localize the position of prospective transparent objects. The methods range from as simple as adopting an encoder-decoder structure to extract boundary [3, 4, 5] and surface normal [7]. Light polarization was utilized to capture the rotation of light waves from a Physics perspective [6] and multi-view stereo images to generate depth maps that further outline the object shape information [6]. However, glass panels with flat surfaces usually do not possess the boundary characteristic, which induces an even more challenging obstacle for detection models.

Context-Aware Detection. Context-aware methods tackle the limited receptive field bottlenecks of convolutional kernels. Most works employed auxiliary operations such as dilation [23, 24] and pooling [25] to enlarge the receptive field such that it gains more global contextual information. More specific solutions that are tailored for various types of surface detection also followed this fashion of contextual learning, such as for mirror surface [26, 27], and glass surface [8, 18]. Nevertheless, these methods are yet another feature aggregation strategy that leaves behind the implicit reasoning embedded in the network learning and rarely exploits the explicit semantic relationship.

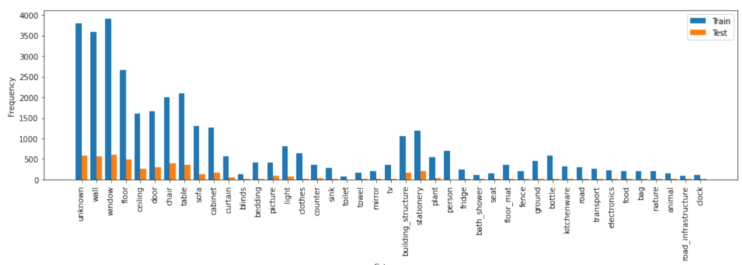
Attention-based Detection. Attention mechanism from [28] enables the modeling to be more specific and oriented towards meaningful context. Early works used matrix formulation to construct such attention purpose [25, 29, 30]. A boost in performance is nurtured by [31] which actualized the 'transformer' concept in Computer Vision tasks. Nonetheless, it is still in the form of spatial context, which relies on semantic feature extraction from each local patch. [32, 33] proposed novel strategies to instead focus on semantic context to study the relationships between objects and scenes for semantic meanings, which served as an inspiration for our work to utilize semantic dependency. Subsequent work [34] swiftly merged the 'transformer' concept and semantic category embedding in transparent object detection. However, the model is constrained to model relationships between a few types of transparent objects, e.g., eyeglasses, bowls, freezers, and windows. This method completely ignores and disposes of the potential of semantic meanings between object categories and scenic information; we aim to fill the gap by emphasizing contextual relationships.

3 Proposed Dataset

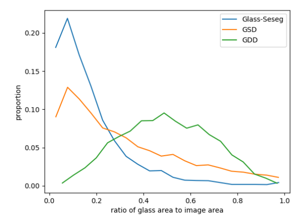
There exist numerous datasets [35, 36] that are dedicated for semantic segmentation tasks. Since most of them are augmented for common objects and thus have a general classification purpose, more refined labelings would be required. Consequently, a rectified dataset that specifically caters to 'glass surface' detection with respect to semantic context is still in need of. On the other hand, [18] and [8] are among the earliest teams who pioneered the 'glass detection' studies and contributed large-scale glass datasets. While half of [8]'s GSD dataset was assembled from existing semantic segmentation datasets (e.g. [35, 36]), [18]'s was constructed from manual collection and organization with only ground truth glass masks. We herewith propose a new Semantic-Aware Glass Surface Detection (GSD-S) dataset, which is accompanied by polished ground truth 'glass surface' masks, with the hope that this contributes for future extensions.

Table 1: Composition of our proposed RGB-D GSD dataset. We collect glass images from four existing RGB images datasets with semantic annotations. Note that these datasets originally lack refined annotations of ground truth glass surface masks. Therefore, we re-labelled the GT masks of glass surfaces during our dataset construction.

Dataset	Whole	Train	Test
SUN RGB-D [37]	1,203	920	283
2D-3D-Semantics [38]	600	488	112
Matterport3D [2]	1,206	992	214
COCO-Stuff [36]	1,511	N/A	1,511
Total	4,519	3,911	608



(a) Class Distribution



(b) Area Ratio

Figure 2: Dataset Statistics

Dataset Composition. The GSD-S dataset is scrupulously organized from existing semantic segmentation datasets [39] with the corresponding ground truth annotation carefully refined since the glass mask labelings in original versions were in general inconsistent. Examples of ambiguous mask labeling include glass areas being segmented into a different category of surrounding objects due to the glass surface transparency; and glass surfaces that were ignored and treated as part of the larger subject, such as cupboard (glass door), car (glass window), table (glass table). We processed 4,519 images with 3,911 training images and 608 testing images altogether. The split between training and testing sets strictly follows that of the original datasets whenever possible. Subset from Matterport3D [39] was randomly sampled. Table 1 gives an overview of the mentioned distribution. Figure 2 displays the object category distributions of our dataset on which the semantic backbone model was trained. Both training and testing sets follow a similar distribution. The area ratio of GSD-S is distinctively small compared to the other ones due to the consideration that more semantic context can be included which allows comprehensive scene context relationship modelling.

4 Proposed Method

Figure 3 illustrates the proposed model's architecture. The input image is first fed into two backbone networks for spatial and semantic feature extraction. The semantic backbone is based on PyTorch's pretrained DeeplabV3-ResNet50 model [24] to learn the intrinsic representations of each object category. The spatial backbone adopted the SegFormer network [19], with transformer blocks that aim to aggregate spatial-wise object location features without the concern of inefficiency from long-range dependency. These cross-disciplinary features are initially input into SAA Module for feature selection and activation according to spatial and channel features. The higher-level feature that contains more abstract information is preserved for CCA Module for feature correlation by mapping correlations between objects' location information and their category-specific semantic meanings. In our proposed model, high-level feature is collected from the last layer of the backbone networks i.e., layer 4, as recommended by the results after exhaustive testings (Section 5.3). The collection of enhanced feature maps output by our two novel modules are processed by the decoder network, for which we applied UperNet [40].

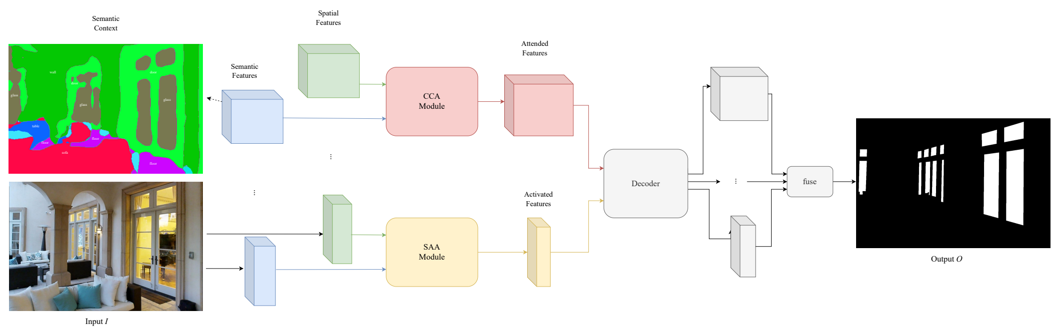


Figure 3: **Model Architecture**. We first feed input image into backbones to capture semantic knowledge and spatial location features. Low-level features will initially get selectively activated with respect to the bifurcated features. The CCA Module is positioned at higher level in the encoder component to inference the relationships between contextual meanings and locations of objects. Features from multiple stage are aggregated by decoder to produce output map.

The following subsections will go through the details on backbone networks (Section 4.1), Scene Aware Activation (SAA) Module (Section 4.2) and Context Correlation Attention (CCA) Module (Section 4.3).

4.1 Backbone Networks

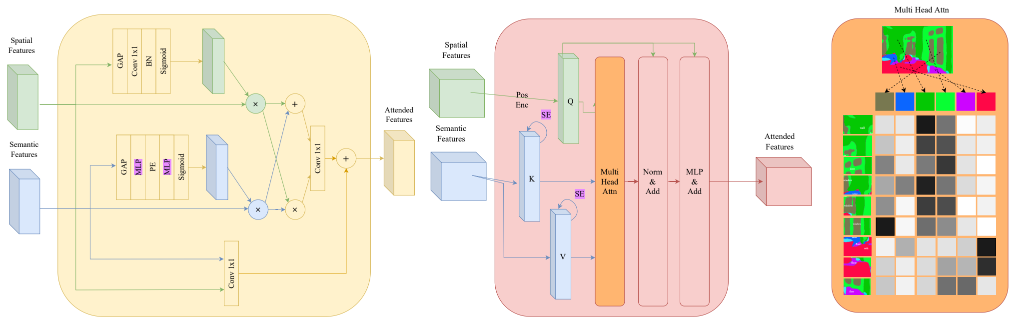
The backbone networks are complementary to each other in that we hope to explicitly leverage spatial and semantic features. The spatial-wise attention in the SegFormer backbone offers insight into each object's geographic information along with corresponding proximity. This takes into account the fact that objects can arise in different regions in the picture under various types of circumstances. For example, glass windows of commercial buildings that situate in different corners in the scene, although being spatially distant, both should belong to the same category and have a hidden dependency. On the other hand, glass surfaces can appear in the form of a 'glass window' on the left-hand side of the room, and 'glass door' on another; as well as 'glass table' somewhere else. Given the different forms of existence, on top of varying physical shapes, we will need to devise the semantic meanings such that the model can better differentiate the semantic categories and, at the same time, correlate the implicit relationships among objects. For instance, co-occurrence of objects due to semantic context such as 'vehicle' and 'glass_surface' (car window), 'wall' and 'glass_surface' (window), etc. The ResNet backbone comes into play by offering a richer representation of objects beyond locations and physical characteristics e.g., boundary and reflection cues. Concretely, the integration of both paths will enable the differentiation of object representations and correlation on object dependency.

After backbone feature extraction, we segregate the feature layers into low-level f_l for $l = \{1, 2, 3\}$ and high-level f_h for $h = 4$ by inputting the features respectively into SSA Module and CCA Module. Since low-level features retain high-resolution spatial context and thus is effective for fine-grained details, they are used for context and object differentiation. Those in high-level embedded with richer and abstract information are then used for semantic correlations.

4.2 Scene Aware Activation (SAA) Module

Inspired by [41], the information contained in feature maps from each layer can be further reinforced through selection and activation operations. Compared to [41], which only considers generic convolutional layers, we decouple the enhancement process into respective spatial and semantic paths to suit our contextual learning settings. Formally, the definition is:

$$
\begin{aligned}
A_{sp}(f_{sp}) &= f_{sp} * \sigma(BN(\phi_{1\times1}(\rho_s(f_{sp})))) \\
A_{se}(f_{se}) &= f_{se} * \sigma(l_{C\times nc}(l_{nc\times C}(\rho_c(f_{se})) + \gamma)) \\
f_{act} &= \phi([A_{sp}(f_{sp}) + A_{se}(f_{se}); A_{sp}(f_{sp}) \times A_{se}(f_{se})]) \\
A(f_{sp}, f_{se}) &= \phi([f_{sp}, f_{se}]) + f_{act}
\end{aligned}
\tag{1}
$$



(a) SAA Module (b) CCA Module

Figure 4: SAA Module takes low-level backbone features to activate and discriminate object meanings; higher-level feature is used for contextual information correlation.

where ϕ denotes the convolution operation with kernel size of 1×1. ρ is the average pooling operation with s being global spatial average and c channel average. $sigma$ is the activation function which in this case, sigmoid is chosen. l is a fully connected layer, for which we set the intermediate reduction dimension to be the number of categories in dataset (*i.e.*, $nc = 43$). γ refers to the positional encoding added to the reduced intermediate layer to ensure the consistent ordering of categorical information. This is a critical step to reinforce semantic class knowledge by abstracting the channel features into corresponding categories as well as preserving the order to align the learned knowledge dependency. This enhancement strategy of category-specific knowledge is shaded in purple, as shown in the network diagrams above (Figure 4). Through selective activations, the SAA Module serves to differentiate the semantic categories.

4.3 Context Correlation Attention (CCA) Module

Witnessing the success of ViT [31], the attention mechanism significantly promotes the modeling efficiency on long-range dependency, enabling objects in any region to be thoroughly analyzed. Existing methods operate by generating the (query, key, value) triplets from single input to achieve self-attention. We propose to bifurcate the attention procedure with respect to spatial and semantic features in order to model the correlation between objects of different categories and their corresponding locations. Formally, the procedure is defined as:

$$
\begin{gathered}
\mathcal{Q} = l(f_{sp}) \\
\mathcal{K}, \mathcal{V} = \xi(l(f_{se})) \\
Attention(Q, K, V) = softmax(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{d_k}})\mathcal{V}
\end{gathered}
\tag{2}
$$

where l denotes the fully connected layer and ξ is the Squeeze and Excitation step adopted from [42]. Before proceeding to the Attention operation, K and V from semantic backbone feature f_{se} is processed with an intentional enhancement operation according to class-specific knowledge as mentioned above. As displayed in Figure 4, the relationship between objects can be explicitly interpreted in terms of spatial information and semantic knowledge for more delicate reasoning.

5 Experiments

5.1 Implementations

Specifically, the SegFormer backbone adopted variation B5 of Mix Transformer encoders (MiT-B5) The model is coupled with pre-trained weights for the purpose of transfer learning. The ResNet backbone is based on PyTorch's DeeplabV3-ResNet50 model [24] pre-trained on COCO train2017 [43] with only 21 categories from Pascal VOC [44]. We then further fine-tuned the model using our GSD-S dataset to introduce a more diverse set of object categories for better semantic extraction

Table 2: Evaluation results on GDD and GSD.

Dataset		GDD				GSD			
Methods	Venue	IOU↑	F_β ↑	MAE↓	BER↓	IOU↑	F_β ↑	MAE↓	BER↓
PSPNet [47]	CVPR 2017	0.792	0.875	0.132	11.51	0.703	0.834	0.110	10.66
BDRAR [48]	ECCV 2018	0.800	0.908	0.098	9.87	0.759	0.860	0.081	8.61
BASNet [49]	ICCV 2020	0.808	0.891	0.106	9.37	0.698	0.808	0.106	13.54
MINet [50]	CVPR 2020	0.844	0.919	0.077	7.40	0.773	0.879	0.077	9.54
GateNet [51]	ECCV 2020	0.817	0.931	0.073	8.84	0.689	0.898	0.073	10.12
MirrorNet [26]	ICCV 2019	0.851	0.903	0.083	7.67	0.742	0.828	0.090	10.76
PMD [52]	CVPR 2020	0.870	0.930	0.067	6.17	0.817	0.890	0.061	6.74
GDNet [18]	CVPR 2020	0.876	0.937	0.063	5.62	0.790	0.869	0.069	7.72
GlassNet [8]	CVPR 2021	0.881	0.932	0.059	5.71	0.836	0.901	**0.055**	6.12
Ours		**0.902**	**0.942**	**0.059**	**4.67**	**0.854**	**0.903**	0.068	**5.69**

capacity. Note that the semantic backbone after fine-tuning is fixed and isolated from subsequent training for glass surface detection, lest additional information would distort the learned semantic representations. Kaiming uniform initialization [45] is used before the model was trained on a NVidia RTX 2080Ti GPU. The input data is first uniformly resized to the size of 384 $\times$ 384 before applying normalization, with Binary Cross Entropy with Logits loss being used to supervise the output feature map. The prediction evaluation is accompanied by Fully Connected Conditional random fields [46] technique for binarization refinement. The evaluation metrics include intersection over union (IoU), Mean Absolute Error (MAE), maximum F-measure (F_β), and balance error rate (BER).

5.2 Comparisons

We evaluate our model against 13 other state-of-the-art methods, including PSPNet [47], DeepLabV3+ [53], PSANet [54], DANet [55] for generic semantic segmentation, and SCA-SOD [13] for Salient Object Detection; recent avant-garde models that utilize transformer technique such as SETR [56], Segmenter [57], Swin [58], ViT [31], SegFormer [19], Twins [59]; and glass surface detection models, GDNet [18] and GlassNet [8]. All the methods are re-trained on GDD, GSD and GSD-S, according to the default training settings stated in the original papers. Table 2 and Table 3 outlines the quantitative performance on the three glass detection datasets with respect to the four evaluation metrics which shows that our model gives a major performance increase compared to most models. Comparing to the second best model i.e. GlassNet [8], our model surpasses with an improvement of 4.02%, 4.87%, 48.8% and 2.53% respectively for IOU, F_β, MAE and BER on GSD-S.

Table 3: Evaluation results on GSD-S.

Methods	Venue	IOU↑	F_β ↑	MAE↓	BER↓
PSPNet	CVPR 2017	0.560	0.679	0.093	13.40
DeepLabV3+	CVPR 2018	0.557	0.671	0.100	13.11
PSANet	ECCV 2018	0.550	0.656	0.104	12.61
DANet	CVPR 2019	0.543	0.673	0.098	14.78
SCA-SOD	ICCV 2021	0.558	0.689	0.087	15.03
SETR	CVPR 2021	0.567	0.679	0.086	13.25
Segmenter	ICCV 2021	0.536	0.645	0.101	14.02
Swin	ICCV 2021	0.596	0.702	0.082	11.34
ViT	ICLR 2021	0.562	0.693	0.087	14.72
SegFormer	NeurIPS 2021	0.547	0.683	0.094	15.15
Twins	NeurIPS 2021	0.590	0.703	0.084	12.43
GDNet	CVPR 2020	0.529	0.642	0.101	18.17
GlassNet	CVPR 2021	0.721	0.821	0.061	10.02
Ours		**0.754**	**0.861**	**0.041**	**9.77**

On the other hand, it is evident that the significant disparity illustrates the challenging nature of our dataset since GSD-S has a relatively small area ratio for glass surfaces, thus giving more room for a diversified set of additional objects that offer richer semantic context. While GDNet is not completely catered for such assorted scenarios, GlassNet stands out with its Rich Context Aggregation Module.

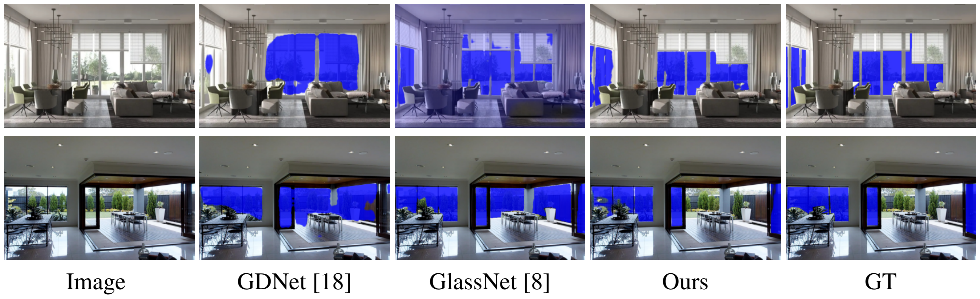


Figure 5: Existing methods that rely on physical boundary features [18] and contextual contrasts [8] tend to mistake non-target regions due to coincident distractions (e.g. wall region in 1^{st} row and window wall in $2^{nd}row$). Our model evaluates semantic correlations on top of superficial characteristics e.g. (wall, furniture and windows in the room to refine predictions).

However, with the assistance of an understanding of intricate scene context, the performance of our model is further elevated.

Figure 5 shows the qualitative comparisons of our method with state-of-the-arts. State-of-the-art methods tend to wrongly predict the regions associated with glass surfaces (*e.g.*, curtains, opened doors) as the glass regions due to the insufficient semantic modeling. In contrast, our method can handle these complex cases and correctly segment the glass regions.

5.3 Ablation Study

Ablation study was conducted to validate the contribution of each module, as detailed in Table 4.

Semantic Backbone. We start with the most fundamental baseline, which only consists of the pre-trained DeepLabV3-ResNet backbone readily provided by PyTorch with no fine-tuning and the CCA Module with transformer block (Version 1). After fine-tuning the model, a noticeable improvement can be seen (Version 2). By fixing the gradient update of the semantic backbone (Version 3), except for a slight drop in F_β, the performance gain is in accordance with the hypothesis that extra information from glass surface detection would deform the original semantic knowledge from the previous fine-tuning stage.

Attention Encoding Formulation. We compared the two options of embedding vectors (Q, K, V) assignment. The initial trial designates semantic backbone features to be ‘Query’ embedding and spatial features to be ‘Key’ and ‘Value’ embeddings (Version 3). The roles are then switched in the subsequent trial (Version 4), where a drastic performance gain was observed. We deduced that after the correlation mapping between Q and K, using semantic backbone features as V allows a more variegated query space than that by spatial features, which only has information in terms of separate patches over the scene.

Table 4: Ablation study.

Version	Split	Spatial	Semantic	SSA	CCA	IOU↑	F_β ↑	MAE↓	BER↓
1	12|34	KV	Q ; w/o tune	✗	✓	0.724	0.835	0.044	11.03
2	12|34	KV	Q ; w/o fix	✗	✓	0.726	0.841	0.046	10.77
3	12|34	KV	Q	✗	✓	0.728	0.839	0.044	10.56
4	12|34	Q	KV	✗	✓	0.743	0.852	0.048	10.04
5	12|34	Q ; SE	KV ; SE	✗	✓	0.745	0.856	0.042	10.06
6	123|4	Q ; SE	KV ; SE	✗	✓	0.750	0.853	0.051	9.76
7	123|4	Q ; SE	KV ; SE	✓	✗	0.749	0.855	0.042	9.88
8	123|4	Q ; SE	KV ; SE	✓	✓	0.753	0.857	0.042	9.70
9	123|4	Q ; SE + PE	KV ; SE	✓	✓	0.751	0.853	0.041	**9.32**
10*	123|4	Q ; SE + PE	KV ; SE + PE	✓	✓	**0.754**	**0.861**	**0.041**	9.67

Image GDNet [18] GlassNet [8] Ours GT

Figure 6: Limitations. Our method may fail to detect glass surfaces in some very challenging scenes with ambiguous visual semantics caused by mirrors.

Structural Enhancement. Category-specific Squeeze and Excitation operations are applied on both backbones (Version 5). The positive effect is backed by a marginal gain on all evaluation metrics.

Holding all settings constant, we then re-define the threshold of low- and high-level feature segregation (Version 6). In other words, we changed the module inputs from:

SSA(f_{low}) where $f_{low} = f_i$ for $i = \{1, 2\}$ and CCA(f_{high}) where $f_{high} = f_j$ for $j = \{3, 4\}$ to:

SSA(f_{low}) where $f_{low} = f_i$ for $i = \{1, 2, 3\}$ and CCA(f_{high}) where $f_{high} = f_j$ for $j = \{4\}$. A mild effect is observed from the improvement IOU and BER and meanwhile a degradation on F_β and MAE.

Proposed Modules. From the permutation (Version 6, 7, 8), the results demonstrate that the mutual presence of both SSA and CAA Modules can offer a breakthrough advancement on all metrics compared to the cases when either one was missing. This confirms the significance of the SSA Module in object characteristic differentiation, as well as that of the CAA Module on context correlation.

Although [19] explicitly mentioned that the hierarchical Transformer structure in the SegFormer network removes the need for 'Positional Encoding'. We still experimented with this variation (Version 9) in the ablation study out of curiosity. The addition indeed provided limited effect. The performance given by most metrics has trivial fluctuations except BER with an apparent gain. Considering such potential impact, we included 'Positional Encoding' in our final version (Version 10) and received a considerably favorable result, which is our conclusion of the proposed model.

5.4 Limitation

Our model would have constrained performance under particular circumstances. For example, in the presence of 'mirror' where there exist high-resolution reflections, scenery along with clear semantic context is mirrored. This leads to a wrong prediction of false-positive glass surface presence inside the mirror region (shown in Figure 6), which is admittedly unsatisfactory. However, this is also a challenging topic that requires state-of-the-art resolutions. In the future, it is hoped that we can integrate both detection strategies to ameliorate this bottleneck.

6 Conclusion

In this paper, we proposed to consider semantic knowledge in combination with spatial information to better capture the scene context as a strategical enhancement for tackling the glass surface detection problem. This comes with a meticulously processed large-scale dataset with refined ground truth masks. Thorough experimentation demonstrates the capability of the SAA Module on object characteristic differentiation and the effectiveness of the CCA Module on context correlation. In consequence, our state-of-the-model model sets new benchmark records on GDD and our GSD-D datasets.

References

[1] Oliver Zendel, Katrin Honauer, Markus Murschitz, Martin Humenberger, and Gustavo Fernandez Dominguez. Analyzing computer vision data - the good, the bad and the ugly. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

[3] Guanying Chen, Kai Han, and Kwan-Yee K. Wong. Tom-net: Learning transparent object matting from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII*, page 696–711, Berlin, Heidelberg, 2020. Springer-Verlag.

[5] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15859–15868, October 2021.

[6] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[7] Zhengqin Li, Yu-Ying Yeh, and Manmohan Chandraker. Through the looking glass: Neural 3d reconstruction of transparent shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[8] Jiaying Lin, Zebang He, and Rynson W.H. Lau. Rich context aggregation with reflection prior for glass surface detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13415–13424, June 2021.

[9] Daniel Kaiser, Timo Stein, and Marius V Peelen. Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. *Proceedings of the National Academy of Sciences*, 111(30):11217–11222, 2014.

[10] Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5:617–629, 2004.

[11] Moshe Bar and Elissa Aminoff. Cortical analysis of visual context. *Neuron*, 38:347–58, 2003.

[12] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527, 2007.

[13] Avishek Siris, Jianbo Jiao, Gary K.L. Tam, Xianghua Xie, and Rynson W.H. Lau. Scene context-aware salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4156–4166, October 2021.

[14] Jinyu Yang, Weizhi An, Chaochao Yan, Peilin Zhao, and Junzhou Huang. Context-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 514–524, January 2021.

[15] Wei He, Meiqing Wu, Mingfu Liang, and Siew-Kei Lam. Cap: Context-aware pruning for semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 960–969, January 2021.

[16] Yi-Sheng Liao and Chun-Rong Huang. Semantic context-aware image style transfer. *IEEE Transactions on Image Processing*, 31:1911–1923, 2022.

[17] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z. Li. Context-aware attention network for image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[18] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[19] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12077–12090. Curran Associates, Inc., 2021.

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

[21] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[23] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[24] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017.

[25] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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

[27] Jiaying Lin, Guodong Wang, and Rynson W. H. Lau. Progressive mirror detection. In *CVPR 2020*.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[29] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[30] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[31] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[32] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph E. Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Where do transformers really belong in vision models? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 599–609, October 2021.

[33] Mingyuan Liu, Dan Schonfeld, and Wei Tang. Exploit visual dependency relations for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, June 2021.

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

[35] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *IJCV*, 127(3):302–321, 2019.

[36] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR 2018*.

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

[38] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015.

[39] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *3DV 2017*.

[40] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[41] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

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

[43] 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.

[44] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2), 2010.

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

[46] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *NeurIPS*, pages 109–117, 2011.

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

[48] 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.

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

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

[51] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, 2020.

[52] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau. Progressive mirror detection. In *CVPR*, 2020.

[53] 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.

[54] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. PSANet: Point-wise spatial attention network for scene parsing. In *ECCV*, 2018.

[55] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019.

[56] 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.

[57] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[58] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[59] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *NeurIPS 2021*, 2021.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes]
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Yes
   (b) Did you describe the limitations of your work? [Yes] Yes
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] N/A
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] Yes
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] N/A
   (b) Did you include complete proofs of all theoretical results? [N/A] N/A
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] N/A
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Yes
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Yes
   (b) Did you mention the license of the assets? [Yes] Yes
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A] N/A
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Yes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] N/A
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] Yes
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] N/A
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] N/A