Exploiting Semantic Relations for Glass Surface Detection

Jiaying Lin*, Yuen-Hei Yeung*, Rynson Lau

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Background:

- Glass surfaces are prevalent in daily lives and often go unnoticed by us
- Humans are generally capable to infer them and avoid collisions, it is difficult for robotic systems due to the transparent nature [1]
- Previous methods attempted to extract priors e.g. object boundaries[1], reflections[2] and polarization[3], without which the methods would fail

Observation:

- Humans are able to reason through the semantic context of the environment[4, 5], which offers insights into the category of and proximity between entities that are expected to appear in the surroundings

Contributions:

- Strategy to apply semantic relationship modelling to cognitively infer correlations between transparent objects and everyday objects for glass surface detection
- Large-scale dataset with complex scenes containing semantic contexts, serving as a benchmark for performance validation on future methods
- Two novel deep learning modules to capture long-range spatial and implicit semantic dependencies, with STOA performance

Proposed Dataset

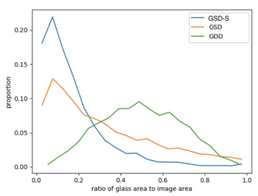
Glass Surface Detection – Semantics Dataset (‘GSD-S’)

- Existing datasets [1, 2] have mostly close-up shots that lack diverse semantic contexts
- GSD-S contains richer semantic content and refined ground truth masks for both semantic segmentations and glass surfaces.

Preview of GSD-S (RGB, Semantic Segmentation, Ground Truth Glass Masks)

Dataset	Whole	Train	Test
SUN RGB-D [38]	1,203	920	283
2D-3D-Semantics [39]	600	488	112
Matterport3D [2]	1,206	992	213
COCO-Stuff [37]	1,511	1,511	N/A
Total	4,519	3,911	608

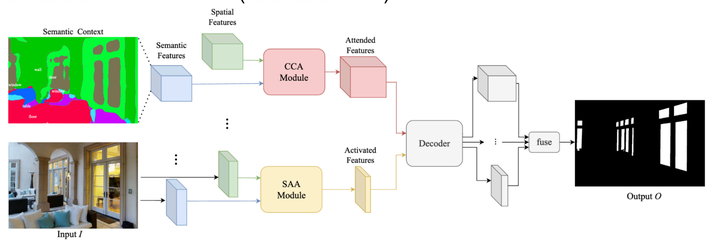
GSD-S Statistics



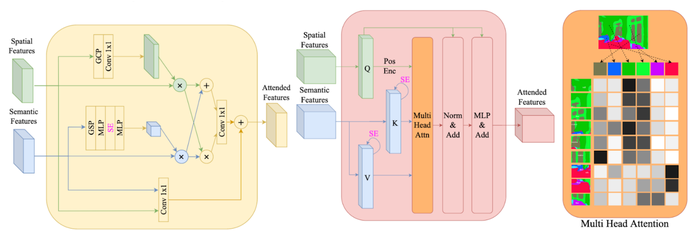
Area Ratio of GSD-S, GSD[2], GDD[1]

Proposed Method

Glass Semantic Network (‘GlassSemNet’)



- Associate pixel-level physical features with semantic meanings via cross-domain integration
- Image input first fed into two backbone networks for spatial and semantic feature extraction
- Subsequent SAA and CCA modules infer semantic contexts before decoder outputs final prediction.



(a) SAA Module (b) CCA Module

SAA Module:

- Capture long-range spatial dependencies
- Decouple enhancement process into respective spatial and semantic paths to suit contextual learning settings

$$f_{sp} \in \mathbb{R}^{H\times W\times C} \rightarrow f'_{sp} \in \mathbb{R}^{H\times W\times 1}$$
$$f_{se} \in \mathbb{R}^{H\times W\times C} \rightarrow f'_{se} \in \mathbb{R}^{nc\times C}$$
$$Activation(f_{sp}, f_{se}; f_{sem\_encod}) = (f_{sp} \times f'_{sp}) \circledast (f_{se} \times (f'_{se} + f_{sem\_encod})')$$

CCA Module:

- Correlate semantic categorical contexts
- Attends to features with strong correlations between objects, and that between objects and the surroundings (spatially and semantically) with reinforcement from semantic encodings by semantic backbone

$$f_{sp} \in \mathbb{R}^{H\times W\times C} \rightarrow Q \in \mathbb{R}^{HW\times C}$$
$$f_{se} \in \mathbb{R}^{H\times W\times C} \rightarrow [K \in \mathbb{R}^{nc\times C}; V \in \mathbb{R}^{nc\times C}]$$
$$Attention(Q, K, V; f_{sem\_encod}) = \text{softmax}(\frac{Q(K + f_{sem\_encod})^\top}{\sqrt{d_k}})(V + f_{sem\_encod})$$

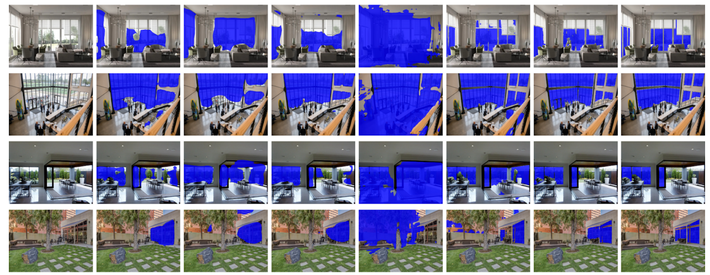


Table 3: Evaluation results on GSD-S.

Methods	Venue	IOU↑	F_β ↑	MAE↓	BER↓
PSPNet [48]	CVPR 2017	0.560	0.679	0.093	13.40
DeepLabV3+ [54]	CVPR 2018	0.557	0.671	0.100	13.11
PSANet [55]	ECCV 2018	0.550	0.656	0.104	12.61
DANet [56]	CVPR 2019	0.543	0.673	0.098	14.78
SCA-SOD [57]	ICCV 2021	0.558	0.689	0.087	15.03
SETR [58]	CVPR 2021	0.567	0.679	0.086	13.25
Segmenter [59]	ICCV 2021	0.536	0.645	0.101	14.02
Swin [60]	ICCV 2021	0.596	0.702	0.082	11.34
ViT [61]	ICLR 2021	0.562	0.693	0.087	14.72
SegFormer [15]	NeurIPS 2021	0.547	0.683	0.094	15.15
Twins [62]	NeurIPS 2021	0.590	0.703	0.084	12.43
GDNet [3]	CVPR 2020	0.529	0.642	0.101	18.17
GlassNet [4]	CVPR 2021	0.721	0.821	0.061	10.02
Ours		**0.753**	**0.860**	**0.035**	**9.26**

Table 8: Ablation study (Cross Dataset).

Train	Test	IOU↑	F_β ↑	MAE↓	BER↓
GDD	GSD	0.701	0.782	0.129	11.20
Glass-Seg	GSD	**0.774**	**0.882**	**0.0857**	**9.44**

Comparisons

- Trainings and evaluations were conducted on GSD-S, GSD[2] and GDD[1] with STOA performance achieved in all comparison metrics

References

[1] H. Mei, ‘Don’t Hit Me! Glass Detection in Real-World Scenes’, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[2] J. Lin, Z. He, R. W. H. Lau, ‘Rich Context Aggregation With Reflection Prior for Glass Surface Detection’, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 13415–13424.

[3] H. Mei, ‘Glass Segmentation Using Intensity and Spectral Polarization Cues’, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 12622–12631.

[4] D. Kaiser, T. Stein, M. 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.

[5] M. Bar, ‘Visual objects in context’, *Nature Reviews Neuroscience*. 5. 8. 617–629, 2004.