Multi-Semantic Modeling for Glass Surface Detection in the Wild

Qianyu Cheng, Huankang Guan*, Rynson W.H. Lau*†

Department of Computer Science, City University of Hong Kong
qiancheng6-c@my.cityu.edu.hk, Huankang.Guan@my.cityu.edu.hk, Rynson.Lau@cityu.edu.hk

Abstract

Glass surfaces challenge object detection models as they mix the transmitted background with the reflected surrounding, creating confusing visual patterns. Previous methods relying on low-level cues (*e.g.*, reflections and boundaries) or surrounding semantics are often unreliable in complex real-world scenarios. A glass image inherently comprises three distinct semantic components: semantics of the transmitted content, semantics of the reflected content, and semantics of the surrounding content. In this work, we observe that there is a relationship among these three types of semantics, where reflection semantics closely resembles surrounding semantics, while these two types of semantics tend to be different from the transmission semantics. For example, when on a street, we may see into a cafeteria through a glass wall, intermixed with reflection of the street, while the glass is surrounded by other street contents like shops and pedestrians, thereby creating a unique multi-semantic signature. Based on this observation, we propose the Multi-Semantic Net, *MSNet*, which identifies transmission, reflection, and surrounding semantics from glass images and exploits their relationships for glass surface detection. MSNet consists of two novel modules: (1) A *Semantic Decomposition Module* (SDM) containing Dual-Semantics Extraction Block to extract original image and reflection semantics and Semantic Elimination Block to progressively derive transmission and surrounding semantics, and (2) An *Adaptive Semantic Fusion Module* (ASFM) to fuse these semantic components and adaptively learn their relationships to handle varying reflection conditions. Extensive experiments demonstrate that MSNet surpasses SOTA methods on public glass detection benchmarks. Code will be available at https://github.com/chengqianyu03/MSNet.

Introduction

Glass surface detection (GSD) poses a critical challenge in computer vision due to glass's unique physical property of being both transparent and reflective, resulting in complex visual patterns that confound conventional object detection algorithms. Accurate detection of glass surfaces can benefit applications such as autonomous navigation and robotic systems (for safety navigation), as well as comprehensive scene understanding (Weibel et al. 2023; Sajjan et al. 2020). Existing approaches for GSD suffer from various distinct limitations. Methods relying primarily on low-level visual cues like reflections or boundaries (Mei et al. 2020; Lin, He, and Lau 2021; Mei et al. 2023; He et al. 2021; Yu et al. 2022) can be less effective when these specific cues are subtle or unavailable. Another strategy infers glass presence by analyzing the surrounding content (Lin, Yeung, and Lau 2022). Recent techniques leveraging physical effects, such as ghosting effect (Yan et al. 2025), provide valuable insights, but their effectiveness often depends on acquiring high-quality, close-up images in order for these subtle physical artifacts to be detectable. These limitations highlight the need for a more robust framework that focuses on intrinsic high-level properties of glass surfaces.

In this paper, we introduce a novel approach to glass detection by leveraging the inherent relationships among three distinctive types of semantics present in glass scenes: *transmission semantics* (semantics of the transmitted content of glass regions), *reflection semantics* (semantics of the re-

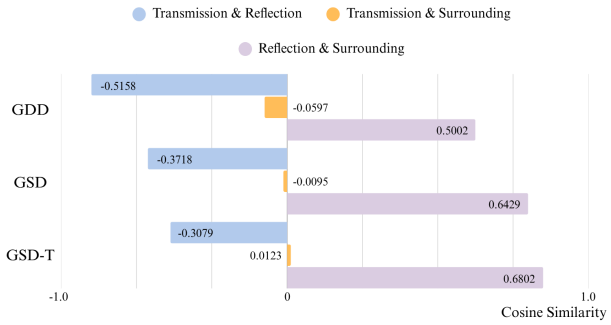


Figure 1: **Semantic Similarity Analysis.** We quantify similarity among *transmission semantics* (semantics of glass transmitted content), *reflection semantics* (semantics of glass reflected content), and *surrounding semantics* (semantics of non-glass regions) based on CLIP, evaluated on GDD (top), GSD (middle) and DSD-T (bottom) datasets. Results consistently indicate a higher similarity between reflection and surrounding semantics, while both differ from transmission semantics. (Positive numbers indicate similarity scores, and negative numbers indicate dissimilarity scores.)

*Corresponding authors: Huankang Guan, Rynson W.H. Lau
†Rynson W.H. Lau leads this project.

flected content of glass regions), and *surrounding semantics* (semantics of non-glass regions). This is based on our observation that reflection semantics typically shares high similarity with surrounding semantics, while both of these semantics are distinctly different from the transmission semantics. For example, when viewing into a cafe with glass walls from the street, the glass wall creates a composite visual experience: it transmits the interior scene of the cafe (*e.g.*, furniture), reflects the street scene (*e.g.*, traffic lights, passing cars), while the glass wall is framed by the surrounding street environment (*e.g.* sidewalks, trees). Together, these three layers form a cohesive yet multi-semantic signature.

Our experimental analysis in Fig. 1 validates this semantic relationship hypothesis. By separately extracting the three types of semantics, a consistent pattern emerges across diverse glass datasets, GDD (Mei et al. 2020), GSD (Lin, He, and Lau 2021) and GSD-S (Lin, Yeung, and Lau 2022) that reflection semantic features exhibit strong cosine similarity with surrounding semantic features, while both have significantly lower similarity when compared to transmission semantic features. This characteristic relationship provides a distinctive signature for glass surfaces, establishing a robust foundation for GSD across diverse scene conditions.

Inspired by this observation, we present the Multi-Semantic Net, MSNet, to exploit multi-view semantics for GSD. It consists of two novel modules: Semantic Decomposition Module (SDM) and Adaptive Semantic Fusion Module (ASFM), along with a base model: Glass-Specific SAM (GSSAM). The SDM employs a Dual-Semantics Extraction Block (DSEB) to identify reflections present in images and obtain reflection and original image semantics, followed by a Semantic Elimination Block (SEB) to effectively separate transmission semantics from the input image using the reflection semantics, and further disentangles surrounding semantics by analyzing the differences between the input image and the transmission semantics. Subsequently, the transmission, reflection, and surrounding semantics are fed into the ASFM, which adaptively learns the relationships among these three semantic components to handle scenes under diverse reflection conditions. The ASFM encodes the relationships among these separated semantics into prompts compatible with GSSAM, thereby guiding it to detect glass surfaces. GSSAM is a SAM model fine-tuned through Low-Rank Adaptation (LoRA) (Hu et al. 2022), which, compared to the original SAM, places greater emphasis on glass-related features during the encoding phase.

Our main contributions are summarized as follows:

- We propose a novel perspective for glass detection by explicitly modeling the interplay among transmission, reflection, and surrounding semantics, framing glass as a semantic boundary that partitions distinct semantic regions in everyday scenes.
- We propose MSNet, a novel glass surface detection framework with a Semantic Decomposition Module (SDM) and an Adaptive Semantic Fusion Module (ASFM) to extract, separate, and adaptively fuse transmission, reflection, and surrounding semantics, enabling robust glass surface detection in challenging scenarios.
- We conduct extensive experiments to demonstrate that our MSNet achieves state-of-the-art performances on publicly available glass detection benchmarks.

Related Work

Glass Detection aims to identify glass surfaces in images. Existing methods fall into three main categories: (1) those using low-level cues like reflections and boundaries (Mei et al. 2020; Lin, He, and Lau 2021; Mei et al. 2023; He et al. 2021; Yu et al. 2022), (2) those analyzing contextual information of the surrounding scene (Lin, Yeung, and Lau 2022), and (3) those exploiting physical phenomena like the ghosting effect (Yan et al. 2025).

GDNet (Mei et al. 2020) pioneered computational glass detection, introducing a large-scale benchmark and a module for contextual feature exploration. GSDNet (Lin, He, and Lau 2021) later improved upon this by incorporating both contextual information and reflection maps. Concurrently, several boundary-based strategies emerged. EBLNet (He et al. 2021) focused on boundary cues to address glass transparency, while GDNet-B (Mei et al. 2023) enhanced GDNet with boundary supervision. However, while reflection-based methods can fail with weak reflections, boundary-based methods may be unreliable for scenes with non-glass boundaries, such as an open window.

Moving beyond low-level cues, PGSNet (Yu et al. 2022) fused high- and low-level features but struggled when complex reflected content obscured semantics on the glass. GlassSemNet (Lin, Yeung, and Lau 2022) leveraged surrounding semantic relationships but faltered when such context was insufficient, for example, when glass dominated the image. Recently, Yan et al. (Yan et al. 2025) utilized the ghosting artifacts, a physical effect of glass, but generally worked on close-up images for the artifacts to be visible.

Video Glass Detection, a more recent and challenging extension, addresses the limitations of static methods in dynamic scenes. Liu et al. (Liu et al. 2024) proposed VGSD-Net, which integrates multi-view dynamic reflection cues as a temporal prior by leveraging the observation that reflections on glass change with camera motion. However, as VGSD-Net requires multiple frames to detect glass in dynamic scenes, a direct comparison with our single-image based method is not applicable.

Mirror Detection has evolved significantly over the years, starting with MirrorNet (Yang et al. 2019), which pioneered the field by modeling semantical and low-level color/texture discontinuities. PMDNet (Lin, Wang, and Lau 2020) progressively learns content similarity between mirror interiors and exteriors using a relational contextual contrasted local (RCCL) module. Guan et al. (Guan, Lin, and Lau 2022) leveraged semantic associations between mirrors and surrounding objects, while HetNet (He, Lin, and Lau 2023) combined low-level and high-level understandings to mimic human behavior. SAT-Net (Huang et al. 2023) focused on symmetry relationships between objects and their reflections, and VCNet (Tan et al. 2023) used visual chirality as a pixel-level cue. Lin et al. (Lin and Lau 2023) introduced the first self-supervised approach, capturing mid-level

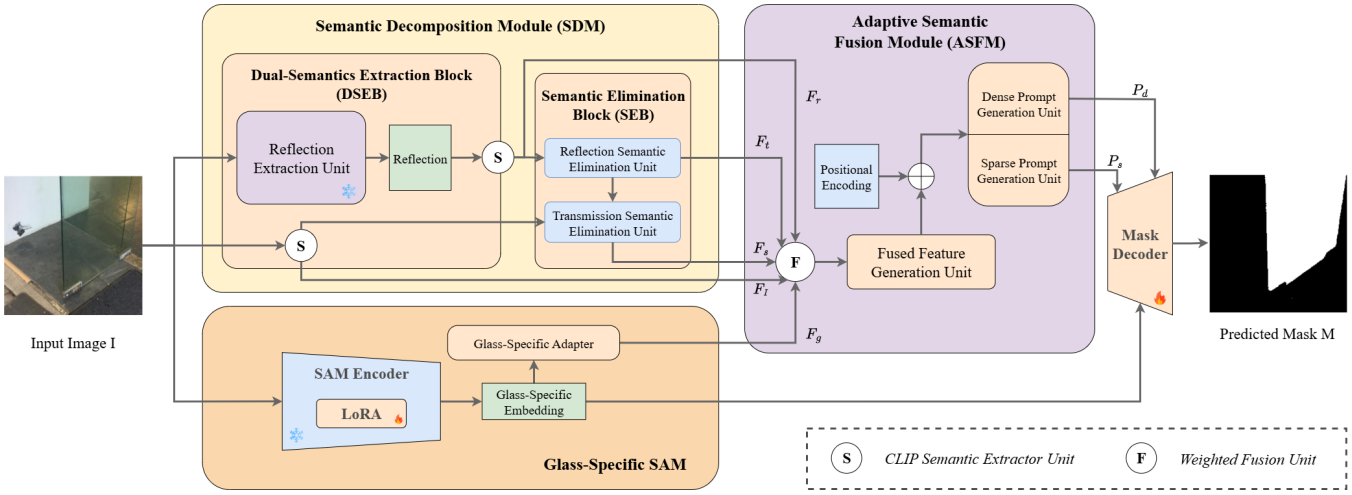


Figure 2: **The MSNet Pipeline.** MSNet exploits the similarity between reflection and surrounding semantics, while contrasting them with transmission semantics for glass detection. The Semantic Decomposition Module (SDM) disentangles these semantic layers, and the Glass-Specific SAM extracts complementary glass-specific features. The Adaptive Semantic Fusion Module (ASFM) then integrates these semantic features to generate prompts for the mask decoder, to produce the final glass mask.

features without supervised ImageNet pre-training.

However, transferring mirror detection techniques to glass detection faces fundamental challenges. While mirrors only reflect surroundings, glass surfaces have dual optical properties, simultaneously reflecting and transmitting content. This superimposition creates complex visual patterns where foreground and background elements overlap. This optical complexity makes glass detection substantially more challenging, rendering mirror-specific methods inadequate for effective glass detection tasks.

Method

Overview

We observe that glass inherently acts as a semantic boundary separating three distinct semantic contexts, transmission semantics, reflection semantics and surrounding semantics. While the reflection semantics often corresponds closely to the surrounding semantics, both differ remarkably from the transmission semantics. Motivated by this observation, we develop *MSNet* for glass surface detection (GSD) by explicitly modeling these semantic relationships and leveraging the similarity between reflection and surrounding semantics as well as their contrast with transmission semantics to localize glass regions effectively.

Fig. 2 shows the pipeline of our proposed MSNet. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we begin by disentangling its semantic layers using the proposed Semantic Decomposition Module (SDM). The SDM first applies a Dual-Semantics Extraction Block (DSEB) to derive a reflection map $\mathbf{R} \in \mathbb{R}^{H\times W\times 3}$ from the input image. It then encodes $\mathbf{I}$ and $\mathbf{R}$ with separate CLIP Surgery (Li et al. 2025) as semantic encoders, producing semantic feature tensors, $\mathbf{F}_I \in \mathbb{R}^{768\times 7\times 7}$ and $\mathbf{F}_r \in \mathbb{R}^{768\times 7\times 7}$. These semantic features are then processed by our Semantic Elimination Block (SEB) to obtain transmission features $\mathbf{F}_t \in \mathbb{R}^{768\times 7\times 7}$ by removing reflection semantics $\mathbf{F}_r$ from the original image semantics $\mathbf{F}_I$. It derives surrounding semantic features $\mathbf{F}_s \in \mathbb{R}^{768\times 7\times 7}$ by separating $\mathbf{F}_t$ from $\mathbf{F}_I$. We also extract the glass-specific features $\mathbf{F_g} \in \mathbb{R}^{256\times 64\times 64}$ using our Glass-Specific SAM (GSSAM), a LoRA-tuned (Hu et al. 2022) variant of SAM (Kirillov et al. 2023), to complement the semantic cues above. All semantic features and glass-specific features, *i.e.*, $\mathbf{F}_I$, $\mathbf{F}_r$, $\mathbf{F}_t$, $\mathbf{F}_s$ and $\mathbf{F_g}$, are fed into our Adaptive Semantic Fusion Module (ASFM). The ASFM employs a dynamic weighting mechanism to adaptively learn the relationships among these multi-semantic features under varying reflection conditions, and outputs both sparse and dense glass-specific prompts $\mathbf{P}_s \in \mathbb{R}^{p_{num}\times 256}$ and $\mathbf{P}_d \in \mathbb{R}^{256\times H\times W}$. Finally, we employ a mask decoder to transform the glass-specific prompts into the fine-grained glass mask $M \in \mathbb{R}^{H\times W\times 1}$.

Semantic Decomposition Module (SDM)

Due to glass's inherent property of simultaneously reflecting surrounding environments and transmitting background scenes, glass surfaces typically present superimposed semantics, adversely affecting object detection models. To tackle this challenge, we propose the Semantic Decomposition Module (SDM) to first employ a Dual-Semantics Extraction Block (DSEB) to isolate surface reflections and obtain image and reflection semantics, and then a Semantic Elimination Block (SEB) to remove the reflection semantics from the glass surface features to obtain the underlying transmission semantics. We further separate the surrounding semantics by excluding the transmission semantics from the entire scene semantics. This decomposition enables the exploitation of relationships between various semantic layers.

Dual-Semantics Extraction Block (DSEB). Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, semantic extraction is performed in two stages, targeting both visual and semantic represen-

tations. First, we employ a reflection extraction network to detect visual reflections:

$$\mathbf{R} = \mathcal{F}_{\text{reflection}}(\mathbf{I}), \tag{1}$$

where $\mathbf{R} \in \mathbb{R}^{H\times W\times 3}$ is the extracted reflection map. The reflection extraction network $\mathcal{F}_{\text{reflection}}(\cdot)$ is based on an existing reflection removal model (Dong et al. 2021), repurposed to extract reflection content rather than removing it.

To capture deeper semantic cues, we then use LoRA fine-tuned CLIP visual encoders to extract features from both the original image and the reflection map:

$$\mathbf{F}_I = \mathcal{E}_{\text{CLIP}}(\mathbf{I}) \in \mathbb{R}^{768\times 7\times 7}, \tag{2}$$

$$\mathbf{F}_r = \mathcal{E}_{\text{CLIP}}(\mathbf{R}) \in \mathbb{R}^{768\times 7\times 7}, \tag{3}$$

where $\mathcal{E}_{\text{CLIP}}(\cdot)$ denotes the LoRA fine-tuned CLIP feature extractor, serving as the semantic encoder in our framework. We utilize the intermediate transformer layer (layer 10 for ViT-B) as the source of semantic features, since intermediate features can preserve spatial information better, which is crucial for the detection task.

Semantic Elimination Block (SEB). After obtaining F_I and F_r, we then introduce the Semantic Elimination Block (SEB) to disentangle transmission and surrounding semantics from the original image features, upon which their interrelations can be modeled. We first obtain the transmission features by removing the reflection semantics from the original image semantics as:

$$\mathbf{F}_t = \mathbf{F}_I \cdot \text{Attn}(\mathbf{F}_I, \mathbf{F}_r) - \mathbf{F}_r, \tag{4}$$

where F_t is the transmission features, and $\text{Attn}(\cdot,\cdot)$ computes an attention map highlighting regions where reflection features dominate. This mechanism utilizes the correlation between reflections and glass surfaces to coarsely localize glass regions. By focusing on these regions and subsequently suppressing reflection semantics, we isolate the transmission semantics corresponding to objects visible through the glass regions.

We then extract the surrounding semantics by removing the transmission semantics from the full-image semantics as:

$$\mathbf{F}_s = \mathbf{F}_I - \text{Attn}(\mathbf{F}_I, \mathbf{F}_t) \cdot \mathbf{F}_t, \tag{5}$$

where F_t denotes the transmission features. Therefore, the SDM breaks down complex glass semantics into four components: original image, reflection, transmission, and surrounding semantics. This multi-semantic decomposition equips the model with rich semantic priors, particularly emphasizing the boundaries and transitions that typically emerge at glass surfaces.

Glass-Specific SAM. We propose a glass-specific SAM by utilizing SAM (Kirillov et al. 2023)'s image encoder as the backbone and applying LoRA adaptation (Cheng et al. 2024) for parameter-efficient fine-tuning. These LoRA modules project high-dimensional features into a compressed latent space and then reconstruct them to their original dimensionality, enabling effective learning of glass-specific representations while updating only a small subset of parameters (Aghajanyan, Gupta, and Zettlemoyer 2021). The glass-specific SAM retains SAM's decoder as a trainable component and replaces the original prompt encoder with our glass prompting mechanism SDM and ASFM.

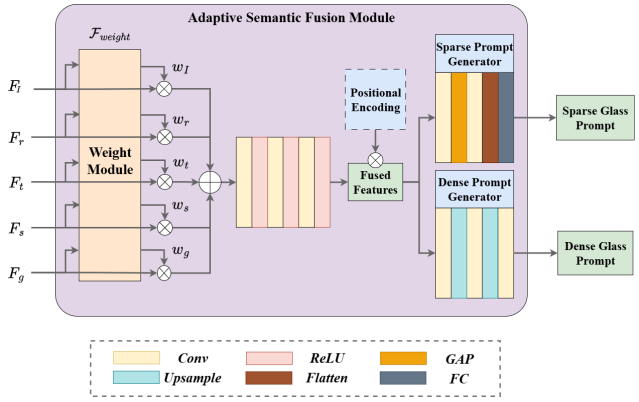


Figure 3: **The ASFM Architecture.** We design the Adaptive Semantic Fusion Module to effectively integrate all semantic features, *i.e.*, $\mathbf{F}_I$, $\mathbf{F}_r$, $\mathbf{F}_t$, $\mathbf{F}_s$, and $\mathbf{F}_g$, which correspond to the image semantics, reflection semantics, transmission semantics, surrounding semantics, and glass-specific features. By fusing these diverse representations, ASFM learns to generate sparse and dense glass prompts that guide the mask decoder to output fine-grained masks.

Adaptive Semantic Fusion Module (ASFM)

As shown in Fig. 3, the ASFM takes multi-semantic features as input and learns to adaptively model their relationships to generate prompts for GSD. The motivation behind our fusion design is to address the diverse and complex nature of reflection conditions and handle potential unreliability of the off-the-shelf reflection removal model. Through adaptive feature weighting, when reflection extraction is less reliable due to model limitations, ASFM dynamically emphasizes other semantic cues (transmission and surrounding semantics) to maintain robust detection. Under strong reflections, the module learns to separate the transmission semantics from the reflection semantics. In contrast, under weak reflections or when reflection features are unreliable, it focuses on distinguishing between surrounding semantics and transmission semantics to effectively localize glass surfaces. Specifically, we first project each input semantic feature map into a unified feature space as:

$$\hat{\mathbf{F}}_j = \mathcal{W}_j(\mathbf{F}_j), \quad j \in \{I, r, t, s, g\}, \tag{6}$$

where $\mathcal{W}_j(\cdot)$ denotes a 1×1 convolutional layer that projects each feature map to a unified dimensionality, enabling semantic fusion. The index j refers to the semantic features extracted from the SDM and the Glass-Specific SAM. Collectively, these projected features offer complementary cues that enhance the robustness of glass detection.

We implement the adaptive weighting mechanism by computing fusion weights dynamically based on the input features to capture diverse semantic relationships across complex glass-containing scenes, as:

$$w_j = \mathcal{F}_{\text{weight}}(\hat{\mathbf{F}}_j), \quad j \in \{I, r, t, s, g\}, \tag{7}$$

$$\mathbf{F}_{\text{fused}} = \frac{\sum_{i\in\{I,r,t,s,g\}} e^{w_i} * \hat{\mathbf{F}}_i}{\sum_{j\in\{I,r,t,s,g\}} e^{w_j}}, \tag{8}$$

Table 1: Quantitative comparison with SOTA methods on **GSD-S**. The best results are in **bold**, while the second best results are underline.

Method	Pub.	IoU ↑	MAE ↓	F_β ↑	BER ↓
GSD-S (Lin, Yeung, and Lau 2022)					
SCA-SOD	ICCV'21	0.558	0.087	0.689	15.03
SETR	CVPR'21	0.567	0.086	0.679	13.25
Swin	ICCV'21	0.596	0.082	0.702	11.34
ViT	ICLR'21	0.562	0.087	0.693	14.72
SegFormer	NeurIPS'21	0.547	0.094	0.683	15.15
Twins	NeurIPS'21	0.590	0.084	0.703	12.43
Mask2Forme	CVPR'22	0.732	0.043	0.838	8.93
MaskDINO	CVPR'23	0.687	0.049	0.816	11.67
FASeg	CVPR'23	0.725	0.048	0.843	10.26
MP-Former	CVPR'23	0.734	0.042	0.827	8.67
NAT	CVPR'23	0.730	0.041	0.846	10.16
X-Decoder	CVPR'23	0.320	0.218	0.452	26.82
SAM	ICCV'23	0.502	0.110	0.618	18.75
SEEN	NeurIPS'23	0.318	0.209	0.474	26.95
SEEN(Fine-tune)	NeurIPS'23	0.751	0.039	0.856	8.98
GDNet	CVPR'20	0.529	0.101	0.642	18.17
GSDNet	CVPR'21	0.721	0.061	0.821	10.02
GlassSemNet	NeurIPS'22	0.754	0.041	0.861	9.77
GhostingNet	TPAMI'25	0.560	0.099	0.703	16.30
Ours		**0.817**	**0.027**	**0.892**	**6.09**

Table 2: Quantitative comparison with SOTA methods on **GDD** and **GSD**. The best results are in **bold**, while the second best results are underline.

Method	Pub.	IoU ↑	MAE ↓	F_β ↑	BER ↓
GDD (Mei et al. 2020)					
MirrorNet	ICCV'19	0.851	0.083	0.903	7.67
PMD	CVPR'20	0.870	0.067	0.930	6.17
GDNet	CVPR'20	0.876	0.063	0.937	5.62
GSDNet	CVPR'21	0.881	0.059	0.932	5.71
EBLNet	ICCV'21	0.887	0.055	0.940	5.36
PGSNet	TIP'22	0.878	0.062	0.901	5.56
GlassSemNet	NeurIPS'22	0.902	0.059	0.942	4.67
RFENet	IJCAI'23	0.874	0.062	0.929	5.79
GhostingNet	TPAMI'25	0.893	0.054	0.943	5.13
Ours		**0.915**	**0.043**	**0.955**	**4.17**
GSD (Lin, He, and Lau 2021)					
MirrorNet	ICCV'19	0.742	0.090	0.828	10.76
PMD	CVPR'20	0.817	0.061	0.890	6.74
GDNet	CVPR'20	0.790	0.069	0.869	7.72
GSDNet	CVPR'21	0.836	0.055	0.901	6.12
EBLNet	ICCV'21	0.817	0.059	0.878	6.75
PGSNet	TIP'22	0.837	0.054	0.868	6.25
GlassSemNet	NeurIPS'22	0.854	0.068	0.903	5.69
RFENet	IJCAI'23	0.836	0.049	0.904	6.24
GhostingNet	TPAMI'25	0.838	0.055	0.904	6.06
Ours		**0.878**	**0.042**	**0.916**	**4.69**

where $\mathcal{F}_{\text{weight}}(\cdot)$ is a lightweight module that assesses the importance of each feature stream based on the current input, outputting a scalar weight. This mechanism allows the model to dynamically modulate the contribution of each semantic component in response to the reflection characteristics specific to each image. The fused features are subsequently refined by a network composed of three convolutional layers followed by ReLU activations, which serves to enhance the semantic relationships:

$$\mathbf{F}_{\text{refined}} = \mathcal{F}_{\text{refine}}(\mathbf{F}_{\text{fused}}). \tag{9}$$

To incorporate spatial structure information and improve positional awareness, we add a learnable positional encoding to the fused features:

$$\mathbf{F}_{\text{spatial}} = \mathbf{F}_{\text{refined}} + \mathcal{P} \in \mathbb{R}^{256\times H'\times W'}, \tag{10}$$

where $\mathcal{P}$ is a learnable positional encoding tensor that is bilinearly interpolated to match the spatial dimension of the refined features, helping the model understand spatial relationships between all types of semantics.

Finally, the enhanced features are used to generate both sparse and dense prompts for the mask decoder:

$$\mathbf{P}_{\text{sparse}} = \mathcal{F}_{\text{sparse}}(\mathbf{F}_{\text{spatial}}) \in \mathbb{R}^{p\_num\times 256}, \tag{11}$$

$$\mathbf{P}_{\text{dense}} = \mathcal{F}_{\text{dense}}(\mathbf{F}_{\text{spatial}}) \in \mathbb{R}^{256\times H\times W}, \tag{12}$$

where $\mathcal{F}_{\text{sparse}}(\cdot)$ consists of fully-connected layers to generate point-based prompts, while $\mathcal{F}_{\text{dense}}(\cdot)$ comprises convolutional layers and upsampling operations to produce dense prompts. These prompts guide the mask decoder in predicting glass masks by decoding the glass-specific features.

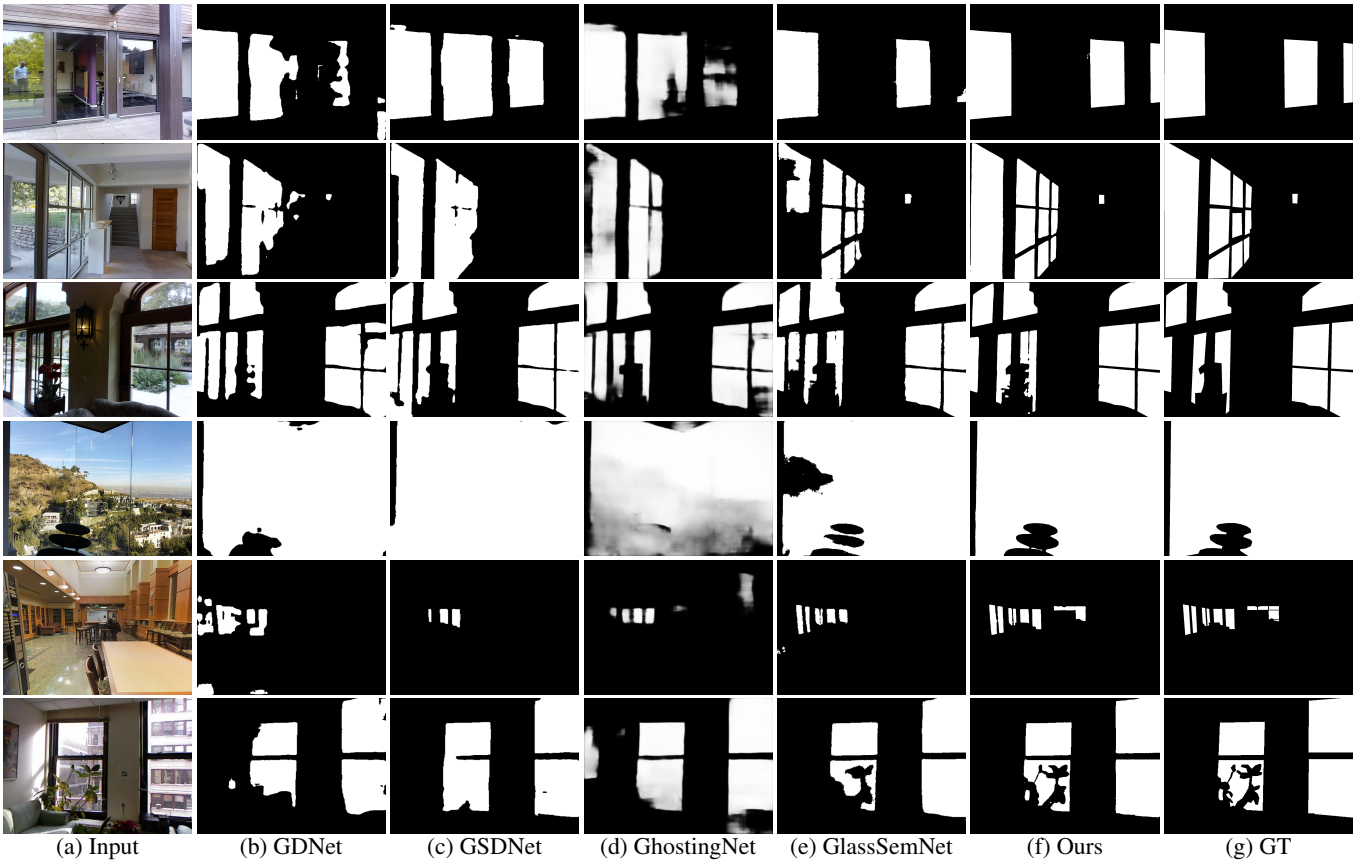
Experiments

Implementation

We implement our method using PyTorch and conduct all experiments on a RTX4090 GPU. Our framework integrates a multi-semantic prompt generation mechanism into the pre-trained SAM-ViT-H model (Kirillov et al. 2023), which is fine-tuned efficiently by applying Low-Rank Adaptation (LoRA) (Hu et al. 2022) to its attention layers with empirically selected rank $r = 512$ and scaling factor $\alpha = 512$. For semantic decomposition, we extract features from the CLIP Surgery model (Li et al. 2025), applying LoRA ($r = 128$, $\alpha = 256$) to its transformer blocks. We use the AdamW optimizer with a weight decay of 5×10^{-4} and a base learning rate of 1×10^{-5}, with a $10\times$ higher rate applied to the SDM and ASFM, as they are more lightweight. Training images undergo data augmentation including random horizontal flipping, scaling ($\pm 20\%$), rotation ($\pm 15°$), and mixup ($\alpha = 0.2$). The loss function is a weighted combination of binary cross-entropy, Dice loss, and focal loss (Lin et al. 2017), with an equal weight of 1.0 for each component. We train MSNet for 50 epochs with a batch size of 1. We adopt IoU, MAE, F-measure (F_β), and BER to evaluate the model.

Main Results

Quantitative Comparison. We evaluate our method on the most recent and challenging GSD-S (Lin, Yeung, and Lau 2022) benchmark, which features complex scene semantics and contains glass surfaces of relatively small size.



(a) Input (b) GDNet (c) GSDNet (d) GhostingNet (e) GlassSemNet (f) Ours (g) GT

Figure 4: **Qualitative Comparison.** Our method (f) consistently yields more accurate results than existing methods (b-e).

As shown in Table 1, we compare our method against a broad set of state-of-the-art methods, including SOD methods (SCA-SOD (Siris et al. 2021)), transformer-based methods (SETR (Zheng et al. 2021), Swin (Liu et al. 2021), ViT (Dosovitskiy et al. 2021), SegFormer (Xie et al. 2021), Twins (Chu et al. 2021)), recent advanced segmentation models (Mask2Former (Cheng et al. 2022), MaskDINO (Li et al. 2023), FASeg (He et al. 2023), MP-Former (Zhang et al. 2023), NAT (Hassani et al. 2023)), universal segmentation approaches (X-Decoder (Zou et al. 2023a), SEEN (Zou et al. 2023b), SAM (Kirillov et al. 2023)), and glass surface detection methods (GDNet (Mei et al. 2020), GSDNet (Lin, He, and Lau 2021), GlassSemNet (Lin, Yeung, and Lau 2022), GhostingNet (Yan et al. 2025)).

Due to the rich semantic information in this dataset, GlassSemNet (Lin, Yeung, and Lau 2022), which leverages surrounding semantic context to localize glass regions, performs strongly. However, our method goes further by decomposing the complex semantics of glass surfaces and explicitly modeling multiple semantic cues, enabling more robust GSD. It surpasses GlassSemNet by 8.4% (IoU), 34.1% (MAE), 3.6% (F_β), and 37.7% (BER), highlighting its effectiveness in tackling semantically complex GSD tasks.

We also compare our method with glass and mirror detection models on the GDD (Mei et al. 2020) and GSD (Lin, He, and Lau 2021) benchmarks (Table 2). Compared to GSDS, the GSD dataset contains larger glass regions, whereas GDD consists of simpler scenes with less semantic information. Our method improves over GhostingNet by 2.5% in IoU, 1.3% in F_β, while reducing MAE and BER by 20.4% and 18.7% on GDD. On GSD, we achieve improvements of 4.8% (IoU), 1.3% (F_β), and reductions of 23.6% (MAE) and 22.6% (BER) when compared with GhostingNet. In contrast to GlassSemNet (Lin, Yeung, and Lau 2022), which focuses on surrounding semantic context, and GhostingNet (Yan et al. 2025), which relies on detecting double reflections, our method models relationships among multiple semantics for a comprehensive semantic understanding, leading to consistently strong performances across diverse scenarios. Our model (907M parameters) achieves an inference time of $\sim$120ms (512$\times$512 image) with 6.78GB GPU memory cost.

Qualitative Comparison. Fig. 4 presents qualitative comparisons between MSNet and previous SOTA GSD methods. As shown in the 1^{st} and 2^{nd} rows, previous methods that rely only on surrounding context or reflections often miss some glass surfaces in scenes with multiple glass instances. In contrast, our model leveraging multiple semantic cues effectively captures all glass instances, including those that reflect outdoor vegetation or reveal indoor environments. In the 3^{rd} row, other methods mistakenly detect the open door as a glass region (false positive), whereas our model accurately excludes it, even though it is surrounded

Table 3: Ablation study of the key components in MSNet. DSEB: Dual-Semantics Extraction Block in SDM, SEB: Semantic Elimination Block in SDM. R: Reflection semantics, T: Transmission semantics, S: Surrounding semantics.

ID	Backbone	SDM	Semantics	ASFM	IoU↑	MAE↓	F_β ↑	BER↓
1	SAM	-	-	-	0.651	0.212	0.738	20.9
2	GSSAM	-	-	-	0.897	0.052	0.944	4.93
3	GSSAM	DSEB	R	-	0.905	0.048	0.940	4.60
4	GSSAM	DSEB+SEB	R+T	-	0.899	0.052	0.940	4.76
5	GSSAM	DSEB+SEB	R+T+S	-	0.907	0.048	0.945	4.58
6	GSSAM	DSEB	R	✓	0.903	0.050	0.943	4.70
7	GSSAM	DSEB+SEB	R+T	✓	0.910	0.045	0.945	4.48
8	GSSAM	DSEB+SEB	R+T+S	✓	**0.915**	**0.043**	**0.955**	**4.17**

by glass surfaces. Our method also performs well across a wide range of glass surface sizes. In contrast, other methods often over- or under-detect extremely large or small glass instances, *e.g.*, the large viewing glass in the 4^{th} row and the small bookshelf glass window in the 5^{th} row. Our method is also able to generate accurate masks even for glass surfaces with irregular shapes (6^{th} row).

Ablation Study

To evaluate the contribution of each major component in our proposed MSNet framework, we conduct a series of ablation studies, as summarized in Table 3. Our analysis focuses on three key aspects: the adaptation of SAM for GSD (Glass-Specific SAM), the effect of Semantic Decomposition Module (SDM), and the role of Adaptive Semantic Fusion Module (ASFM) in integrating multi-semantic cues.

Glass-Specific SAM (GSSAM). We begin with the baseline SAM (Kirillov et al. 2023) model (ID 1), which offers robust general-purpose segmentation performance. However, its effectiveness in detecting glass surfaces is limited due to glass's unique visual characteristics. By introducing a lightweight adaptation using Low-Rank Adaptation (LoRA) (Hu et al. 2022), we construct a Glass-Specific SAM (GSSAM, ID 2), which significantly improves performance across all metrics: IoU rises from 0.651 to 0.897, MAE drops from 0.212 to 0.052, F_β increases from 0.738 to 0.944, and BER falls from 20.9 to 4.93. These improvements show that our GSSAM is a suitable backbone for GSD.

Semantic Decomposition Module (SDM). We evaluate the effectiveness of SDM by integrating DSEB into GSSAM (ID 3), which incorporates reflection semantics. This modification yields a modest performance improvement, *i.e.*, IoU increases by 0.9% and MAE decreases by 7.7%, indicating that reflection cues aid in the GSD task. However, extending SDM with SEB to include transmission semantics (ID 4) results in comparable performance (IoU: 0.899). The minor performance differences from ID 3 to ID 5 suggest that simply adding semantic components without considering their relationships offers limited benefit. These findings underscore our idea that a mechanism is needed to dynamically integrate and balance multiple semantic cues.

Adaptive Semantic Fusion Module (ASFM). ASFM shows clear benefits when fusing multiple semantics but offers limited improvement with only reflection semantics (ID 6 - IoU: 0.903 vs. ID 3 IoU: 0.905). As more semantic cues are introduced, the effectiveness of ASFM becomes more apparent. With reflection/transmission semantics (ID 7), ASFM brings a clear performance gain over ID 4, improving IoU from 0.899 to 0.910 and reducing BER from 4.76 to 4.48. This demonstrates its ability to manage complex semantic interactions and enhance feature integration. The full model (ID 8), by combining reflection/transmission/surrounding semantics through ASFM, achieves the best results across all metrics, reaching an IoU of 0.915, MAE of 0.043, F_β of 0.955, and BER of 4.17. Compared to ID 2, our multi-semantic modeling achieves a 2.0% increase in IoU and a 17.3% decrease in MAE, confirming the importance of adaptive fusion for integrating multiple semantic cues. These ablation results validate that both multi-semantic extraction and adaptive fusion are crucial for robust glass surface detection across diverse scenarios.

Figure 5: **MSNet limitations in extreme lighting.**

Conclusion

In this paper, we have proposed MSNet, a multi-semantic framework for GSD. MSNet explicitly models the interplay among reflection/transmission/surrounding semantic cues. It builds upon a Glass-Specific SAM enhanced with a Semantic Decomposition Module (SDM) to disentangle these semantic components. It then uses an Adaptive Semantic Fusion Module (ASFM) to learn the relationships between semantic cues and fuse them effectively. Extensive experiments across three benchmarks show that MSNet achieves SOTA performance, particularly excelling in challenging scenes involving complex reflections and diverse glass sizes.

Despite its success, MSNet does have its limitations. For example, it struggles under extreme lighting conditions that affect the detected semantic cues. In Fig. 5, under-detection occurs when both transmission and reflection semantics are weak: (top row) the glass panels with dark backgrounds and minimal reflections, and (bottom row) dim lighting on the glass table surface preventing the model from distinguishing between different semantic components.

Acknowledgements: This work is in part supported by two GRFs from the Research Grants Council of Hong Kong (Ref: 11211223 and 11220724).

References

Aghajanyan, A.; Gupta, S.; and Zettlemoyer, L. 2021. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. In *IJCNLP*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *CVPR*.

Cheng, Z.; Wei, Q.; Zhu, H.; Wang, Y.; Qu, L.; Shao, W.; and Zhou, Y. 2024. Unleashing the potential of SAM for medical adaptation via hierarchical decoding. In *CVPR*.

Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; and Shen, C. 2021. Twins: Revisiting the design of spatial attention in vision transformers. *NeurIPS*.

Dong, Z.; Xu, K.; Yang, Y.; Bao, H.; Xu, W.; and Lau, R. W. 2021. Location-aware single image reflection removal. In *ICCV*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Guan, H.; Lin, J.; and Lau, R. W. 2022. Learning semantic associations for mirror detection. In *CVPR*.

Hassani, A.; Walton, S.; Li, J.; Li, S.; and Shi, H. 2023. Neighborhood attention transformer. In *CVPR*.

He, H.; Cai, J.; Pan, Z.; Liu, J.; Zhang, J.; Tao, D.; and Zhuang, B. 2023. Dynamic focus-aware positional queries for semantic segmentation. In *CVPR*.

He, H.; Li, X.; Cheng, G.; Shi, J.; Tong, Y.; Meng, G.; Prinet, V.; and Weng, L. 2021. Enhanced boundary learning for glass-like object segmentation. In *ICCV*.

He, R.; Lin, J.; and Lau, R. W. H. 2023. Efficient mirror detection via Multi-Level Heterogeneous learning. In *AAAI*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Huang, T.; Dong, B.; Lin, J.; Liu, X.; Lau, R. W.; and Zuo, W. 2023. Symmetry-Aware Transformer-Based mirror detection. In *AAAI*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *ICCV*.

Li, F.; Zhang, H.; Xu, H.; Liu, S.; Zhang, L.; Ni, L. M.; and Shum, H.-Y. 2023. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *CVPR*.

Li, Y.; Wang, H.; Duan, Y.; Zhang, J.; and Li, X. 2025. A closer look at the explainability of Contrastive language-image pre-training. *Pattern Recognition*.

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

Lin, J.; and Lau, R. W. H. 2023. Self-supervised pre-training for mirror detection. In *ICCV*.

Lin, J.; Wang, G.; and Lau, R. W. 2020. Progressive mirror detection. In *CVPR*.

Lin, J.; Yeung, Y.-H.; and Lau, R. 2022. Exploiting semantic relations for glass surface detection. *NeurIPS*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *ICCV*.

Liu, F.; Liu, Y.; Lin, J.; Xu, K.; and Lau, R. W. 2024. Multi-View Dynamic reflection prior for video glass surface detection. In *AAAI*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*.

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

Mei, H.; Yang, X.; Yu, L.; Zhang, Q.; Wei, X.; and Lau, R. W. H. 2023. Large-Field contextual feature learning for glass detection. *IEEE TPAMI*.

Sajjan, S.; Moore, M.; Pan, M.; Nagaraja, G.; Lee, J.; Zeng, A.; and Song, S. 2020. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *ICRA*.

Siris, A.; Jiao, J.; Tam, G. K.; Xie, X.; and Lau, R. W. 2021. Scene context-aware salient object detection. In *ICCV*.

Tan, X.; Lin, J.; Xu, K.; Chen, P.; Ma, L.; and Lau, R. W. 2023. Mirror detection with the visual chirality cue. *IEEE TPAMI*.

Weibel, J.-B.; Sebeto, P.; Thalhammer, S.; and Vincze, M. 2023. Challenges of depth estimation for transparent objects. In *International Symposium on Visual Computing*.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*.

Yan, T.; Gao, J.; Xu, K.; Zhu, X.; Huang, H.; Li, H.; Wah, B.; and Lau, R. W. H. 2025. GhostingNet: A Novel Approach for Glass Surface Detection With Ghosting Cues. *IEEE TPAMI*.

Yang, X.; Mei, H.; Xu, K.; Wei, X.; Yin, B.; and Lau, R. 2019. Where is my mirror? In *ICCV*.

Yu, L.; Mei, H.; Dong, W.; Wei, Z.; Zhu, L.; Wang, Y.; and Yang, X. 2022. Progressive glass segmentation. *IEEE TIP*.

Zhang, H.; Li, F.; Xu, H.; Huang, S.; Liu, S.; Ni, L. M.; and Zhang, L. 2023. Mp-former: Mask-piloted transformer for image segmentation. In *CVPR*.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*.

Zou, X.; Dou, Z.-Y.; Yang, J.; Gan, Z.; Li, L.; Li, C.; Dai, X.; Behl, H.; Wang, J.; Yuan, L.; et al. 2023a. Generalized decoding for pixel, image, and language. In *CVPR*.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2023b. Segment everything everywhere all at once. *NeurIPS*.