Multi-View Dynamic Reflection Prior for Video Glass Surface Detection

Fang Liu*, Yuhao Liu*, Jiaying Lin†, Ke Xu†, Rynson W.H. Lau‡

Department of Computer Science, City University of Hong Kong
{fawnliu2333, yuhaoliu7456, csjylin, kkangwing}@gmail.com, Rynson.Lau@cityu.edu.hk

Abstract

Recent research has shown significant interest in image-based glass surface detection (GSD). However, detecting glass surfaces in dynamic scenes remains largely unexplored due to the lack of a high-quality dataset and an effective video glass surface detection (VGSD) method. In this paper, we propose the first VGSD approach. Our key observation is that reflections frequently appear on glass surfaces, but they change dynamically as the camera moves. Based on this observation, we propose to offset the excessive dependence on a single uncertainty reflection via joint modeling of temporal and spatial reflection cues. To this end, we propose the VGSD-Net with two novel modules: a Location-aware Reflection Extraction (LRE) module and a Context-enhanced Reflection Integration (CRI) module, for the position-aware reflection feature extraction and the spatial-temporal reflection cues integration, respectively. We have also created the first large-scale video glass surface dataset (VGSD-D), consisting of 19,166 image frames with accurately-annotated glass masks extracted from 297 videos. Extensive experiments demonstrate that VGSD-Net outperforms state-of-the-art approaches adapted from related fields. Code and dataset will be available at https://github.com/fawnliu/VGSD.

Figure 1: Comparison of our VGSD-Net with the state-of-the-art image-based glass detection methods, GlassNet (Lin, He, and Lau 2021) and PGSNet (Yu et al. 2022). They produce temporal-inconsistent results when applied to the VGSD task, as they do not exploit any temporal information. In contrast, our method learns dynamic reflection cues from the video, yielding more accurate and robust results.

Introduction

Glass surfaces, including glass windows / walls / doors, pervade our everyday lives. Their existence significantly impacts various computer vision tasks, such as depth estimation (Bhat, Alhashim, and Wonka 2021), 3D scene understanding (Ye et al. 2021, 2022b,a), and vision-language navigation (Anderson et al. 2018; Liu et al. 2023b,a). For example, undetected glass surfaces could lead to mishaps like the crashing of drones and robots onto them. Thus, detecting glass surfaces is an essential prerequisite for enhancing the scene-understanding capabilities of vision systems.

Mei *et al.* (Mei et al. 2020) proposes the first image-based glass surface detection method and utilizes the contrasted features to localize the glass regions. Subsequent works leverage various priors for glass surface detection, including boundary (He et al. 2021), reflection (Lin, He, and Lau 2021), and context (Yu et al. 2022). Despite their success, none of them are tailored for video-based glass surface detection. On the other hand, real-world computer vision applications such as autonomous driving and robotic navigation are video-centric rather than image-centric. Effectively addressing the Video Glass Surface Detection (VGSD) problem can offer substantial benefits to them.

There are two main challenges for handling the VGSD problem. First, as existing glass detection methods are predominantly designed for single-image input, their priors/assumptions may not hold true in dynamic scenes. As shown in Fig. 1, GlassNet (Lin, He, and Lau 2021) fails to detect glass surfaces in the third frame as they do not explore temporal reflections, and the insufficient contexts (*e.g.*, the top-left region of the bottom image) fail PGSNet (Yu et al. 2022) to produce complete glass maps. Second, there are currently no datasets available for the VGSD problem.

In this paper, we aim to address the above two challenges. First, we observe that glass surfaces often contain reflections that exhibit a dynamic behavior across multiple frames of the input video (*i.e.*, the location and appearance of reflections on the glass surface change dynamically, as the camera moves). Inspired by this observation, we propose a novel approach, named *VGSD-Net*, which integrates multi-view dynamic reflections across multiple frames for VGSD. VGSD-Net contains two novel modules: a Location-aware Reflec-

*These authors contributed equally.

†Jiaying Lin and Ke Xu are joint corresponding authors.

‡Rynson W.H. Lau leads this project.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tion Extraction (LRE) module and a Context-enhanced Reflection Integration (CRI) module. The LRE module utilizes a masked deformable attention mechanism to extract localized reflection features, which can prompt deformable attention with position awareness. We then feed the reflection features from different frames to the CRI module to exploit multi-view dynamic reflection cues spatially and temporally. While the reflection may vary in intensity in some regions of single frames (*e.g.*, the region inside the red box in Fig. 1), our approach can still identify the whole glass surfaces accurately due to the effective incorporation of spatial and temporal reflection cues across frames.

Second, we build the first large-scale video glass surface detection dataset (VGSD-D). It contains 297 videos (lasting 575 seconds in total) with 19,166 image frames, all of which are carefully annotated with corresponding glass surface masks. We have conducted extensive experiments on our dataset to evaluate the performance of the proposed approach. Results show that our method outperforms 17 state-of-the-art methods in related tasks.

The key contributions of this work can be summarized as:

- We propose the first video glass surface detection method, VGSD-Net, to exploit dynamic reflection cues for video glass surface detection. It has a novel Location-aware Reflection Extraction (LRE) module to restrict the deformable attention to localized features surrounding the predicted glass regions, and a novel Context-enhanced Reflection Integration (CRI) module to incorporate spatial-temporal reflection cues across frames.
- We construct the first large-scale video glass surface detection dataset, VGSD-D, which contains 19,166 image frames from 297 videos with corresponding manually annotated glass surface masks.
- Extensive evaluations show that our method outperforms 17 existing state-of-the-art methods from relevant tasks on our proposed VGSD-D dataset.

Related Work

Glass Detection. Mei *et al.* (Mei et al. 2020) propose the first glass detection dataset for glass surface detection. Lin *et al.* (Lin, He, and Lau 2021) further introduce a more challenging glass surface dataset, and exploit glass reflections to refine the glass regions. Later, Lin *et al.* (Lin, Yeung, and Lau 2022b) propose the first large-scale RGB-D dataset for glass surface detection. Mei *et al.* (Mei et al. 2022) integrate polarization cues for glass segmentation and create a new RGB-Polarization dataset. Lin *et al.* (Lin, Yeung, and Lau 2022a) utilize the semantic feature extractor to exploit the semantic relationship between glass and non-glass regions, enhancing the glass detection in single-image.

LiDAR-based methodologies for detecting glass surfaces have been explored, albeit with inherent challenges. In general, LiDAR alone cannot be used to detect glass surfaces, as laser beams will pass through glass. To address this limitation, Yang *et al.* (Yang and Wang 2008) propose to integrate LiDAR with ultrasonic sensors. Glass surfaces can be detected by comparing the returned signals from the two sensors. However, as ultrasonic sensors have low sampling rates, they cannot be used to handle 3D scenes at video rates. Tibebu *et al.* (Tibebu et al. 2021) propose an alternative approach, using variations in range measurements across neighboring point clouds to identify glass surfaces.

A related task of transparent object detection also raises considerable attention. Several methods explore diverse techniques such as quantized local features (Fritz et al. 2009), depth cues (Guo-Hua, Jun-Yi, and Ai-Jun 2019), polarization information (Kalra et al. 2020), trimap cues (Liu et al. 2021a,b). Instead of relying on additional inputs, Xie *et al.* (Xie et al. 2020) propose a boundary-aware segmentation network that directly operates on the RGB image to detect transparent objects, and build a new transparent object dataset. He *et al.* (He et al. 2021) further propose to learn enhanced boundary cues via a refined differential module.

In contrast to existing methods that primarily address image-based glass detection, our work tackles the more challenging VGSD problem. The concurrent work (Qiao et al. 2023) needs polarization information as additional input, limiting its application in new scenes (*e.g.*, polarization cues are unavailable). In our work, we propose to leverage multi-view dynamic reflection cues extracted from the video to detect glass surfaces. To facilitate our research, we also construct a comprehensive and large-scale VGSD dataset.

Mirror Detection. Yang *et al.* (Yang et al. 2019) propose first mirror detection dataset and detect mirrors by modeling contrasted information. Subsequent approaches expand on this by exploiting various information for mirror detection, such as correspondence (Lin, Wang, and Lau 2020), depths (Mei et al. 2021; Tan et al. 2021), semantics (Guan, Lin, and Lau 2022), chirality cue (Tan et al. 2022), context (Yu et al. 2022) and symmetry (Huang et al. 2023). Recent works further address this problem from a learning-based perspective, such as efficiency (He, Lin, and Lau 2023) and self-supervised learning (Lin and Lau 2023). Most recently, Lin *et al.* (Lin, Tan, and Lau 2023) build the first video mirror detection dataset, and also develop the first video mirror detection network.

Unlike mirrors, which only reflect the scene, glass surfaces produce dual images (*i.e.*, a reflected image from the scene in front of the glass and a transmitted image from the scene behind the glass). This complexity can cause existing mirror detection methods to misinterpret the visual information from glass surfaces.

Video-based Salient Object Detection (VSOD) aims to segment the salient foreground objects from the background in the entire video. Early methods (Wei et al. 2012; Wang, Shen, and Porikli 2015) rely on hand-crafted features to detect and segment salient objects in the video. Wang *et al.* (Wang, Shen, and Shao 2017) pioneer the application of deep learning to VSOD. Gu *et al.* (Gu et al. 2020) further design a pyramid-constrained self-attention module for direct temporal information modeling. Zhang *et al.* (Zhang et al. 2021) use dynamic filters to model interactions between consecutive frames. Optical flow, as a time-continuous prior, is also introduced in several optical flow-centric VSOD methods (Li et al. 2019; Su et al. 2023).

However, as glass may not necessarily be salient, these

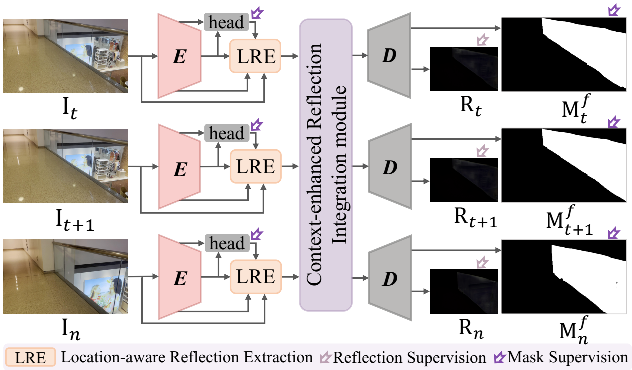


Figure 2: The schematic illustration of our method. With two consecutive frames, $\mathbf{I}_t, \mathbf{I}_{t+1}$, with $t \in \{1, 2, \cdots, N\}$ and one randomly selected frame $\mathbf{I}_n$, the proposed method first extract their multi-scale visual features via a shared encoder. Then, we employ the Location-aware Reflection Extraction (LRE) module to extract localized reflection features for each frame, which are then spatially and temporally integrated via the Context-enhanced Reflection Integration (CRI) module, to form the enhanced features for each frame. One shared decoder is finally utilized to output the predicted glass masks, $\mathbf{M}_t^f$, $\mathbf{M}_{t+1}^f$, $\mathbf{M}_n^f$. We leverage reflections as the auxiliary output (*i.e.*, $\mathbf{R}_t$, $\mathbf{R}_{t+1}$, $\mathbf{R}_n$) to enhance the overall efficiency of the glass detection process.

VSOD methods cannot be used to solve the VGSD problem.

Method

Overview

Our key observation is that while reflections frequently appear on glass surfaces, they display dynamic behavior in the video, with the location and appearance of the reflections changing as the camera position shifts. This observation motivates us to exploit multi-view dynamic reflections for learning a more robust reflection-based glass surface representation for video glass surface detection.

Fig. 2 illustrates the overall structure of the proposed VGSD-Net. Given three glass images as inputs, with the first two images ($\mathbf{I}_t$ and $\mathbf{I}_{t+1}$) taken from adjacent frames, and the third image $\mathbf{I}_n$ randomly selected from other frames in the same video, we first extract their multi-scale visual features $\{\mathbf{F}_i^1, ..., \mathbf{F}_i^5\}$, $i \in \{t, t+1, n\}$ via a shared encoder. Subsequently, low-level features $\mathbf{F}_i^1$ and high-level features $\mathbf{F}_i^5$ are fed into the mask head for intermediate mask prediction $\mathbf{M}_i^c$. $\mathbf{F}_i^1$, $\mathbf{F}_i^5$ and $\mathbf{M}_i^c$ are combined with the input frame $\mathbf{I}_i$ to extract localized reflection features through the Location-aware Reflection Extraction (LRE) module. The extracted reflection features across all frames are then sent into the proposed Context-enhanced Reflection Integration (CRI) module, to facilitate the cross-frame information exchange and integration. Finally, these refined features, combined with multi-level visual features, are sent to the shared decoder to predict both reflection and glass surface mask for each frame.

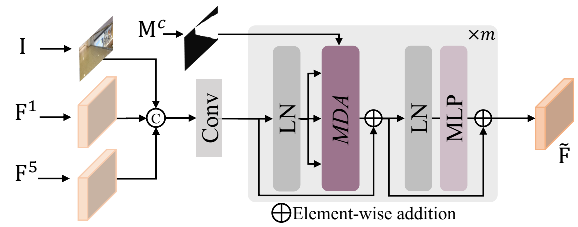


Figure 3: Illustration of the Location-aware Reflection Extraction (LRE) module. The position map $\mathbf{M}^c$ is explicitly introduced into the masked deformable attention (*MDA*) to drive the model to focus on the glass-related reflections.

Location-aware Reflection Extraction Module

Reflections may appear on both glass and non-glass surfaces (*e.g.*, smooth walls or floors), while those in non-glass regions can be highly distracting to glass surface detection. Hence, understanding the position information is crucial, and directly applying the position-unaware auto-encoder (Lin, He, and Lau 2021) for reflection capture does not work well, as validated in Table 2. To this end, we explore position-aware deformable attention for extracting localized reflection features.

Fig. 3 shows the architecture of the LRE module, which leverages the masked deformable attention mechanism (*i.e.*, MDA) to extract localized reflection features. Specifically, MDA takes the feature map $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ and the predicted intermediate glass mask $\mathbf{M}^c \in \mathbb{R}^{H \times W}$ from the mask head as inputs. It first divides the $\mathbf{X}$ into multiple $S \times S$ non-overlapping sub-windows, denoted as $\mathbf{X}^s \in \mathbb{R}^{S^2 \times \frac{HW}{S^2} \times C}$, and utilize three linear projection layers[1] to transform it to query, key and value tensors, $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$. Then, we conduct informative regions selection to pick the most informative regions and eliminate those sub-windows that are all non-glass within the local regions. We apply average pooling within each sub-window to $\mathbf{Q}$ and $\mathbf{K}$ to obtain region-level query and key, $\mathbf{Q}^r, \mathbf{K}^r \in \mathbb{R}^{S^2 \times C}$, and compute the region-to-region affinity matrix $\mathbf{A}^r = \mathbf{Q}^r (\mathbf{K}^r)^\top \in \mathbb{R}^{S^2 \times S^2}$. The row-wise ranking (Zhu et al. 2023) on the $\mathbf{A}^r$ is performed to select the top k most relevant sub-windows for each sub-window in $\mathbf{K}$ and $\mathbf{V}$, and obtain $\mathbf{K}^t, \mathbf{V}^t \in \mathbb{R}^{S^2 \times \frac{kHW}{S^2} \times C}$. The query features $\mathbf{Q}$, along with the selected key and value features $\mathbf{K}^t$ and $\mathbf{V}^t$, are then used to form the MDA[2]:

$$\mathit{MDA}(\mathbf{X}, \mathbf{P}) = \mathit{softmax}(\mathbf{Q}(\mathbf{K}^t)^\top + \mathbf{P})\mathbf{V}^t, \tag{1}$$

where the $\mathbf{P}$ is the position map and can be obtained by:

$$\mathbf{P}(x, y) = \begin{cases} 0 & \text{if } \mathbf{M}^c(x, y) = 1, \\ -\infty & \text{otherwise}, \end{cases} \tag{2}$$

We then encapsulate the MDA into the transformer block to form the masked deformable block:

$$\tilde{\mathbf{X}} = \mathit{MDA}(\mathit{LN}(\mathbf{X}), \mathbf{M}^c) + \mathbf{X}, \tag{3}$$

$$\tilde{\mathbf{F}} = \mathit{MLP}(\mathit{LN}(\tilde{\mathbf{X}})) + \tilde{\mathbf{X}}, \tag{4}$$

[1] No dim. reduction and the shape of Q,K,V, are same as the $\mathbf{X}_s$.

[2] Note that since the LRE will process all frames, all subscripts are omitted in this subsection for simplicity.

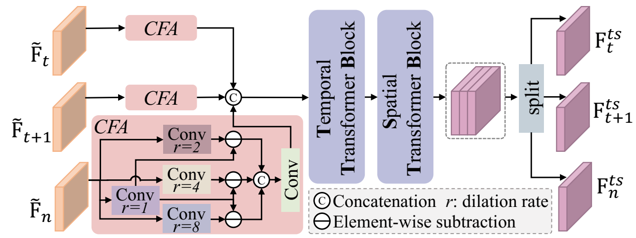


Figure 4: Overview of the proposed Context-enhanced Reflection Integration (CRI) module. The features of different frames undergo enhancement via a Contextual Feature Aggregation (*CFA*) operation, and are then combined and directed to the temporal and spatial transformer blocks for inter-frame information exchange and integration.

where *LN* and *MLP* denote the Layer Normalization and Multi-Layer Perception. $\mathbf{X}$ is the concatenation of the input frame $\mathbf{I}$, low-level features $\mathbf{F}^1$ and high-level features $\mathbf{F}^5$. In this way, the most informative neighboring glass regions are selected and the model can be prompted to focus on them.

Context-enhanced Reflection Integration Module

Despite the incorporation of localized reliable reflection features within the LRE module, reflections from discontinuous glass regions are often under-detected. This phenomenon is more noticeable across frames. To this end, we propose the Context-enhanced Reflection Integration (CRI) module, which has a *CFA* for contextual contrasts enrichment, a *TTB* for temporal reflection feature aggregation, and an *STB* for spatial reflection feature aggregation.

As shown in Fig. 4, the three input reflection feature maps are independently fed to the Contextual Feature Aggregation (*CFA*) to extract multi-scale contrast contextual semantics, which can facilitate the model to learn the contrast information between glass and non-glass regions. Here, we take the $\tilde{\mathbf{F}}_t$ as an example of how *CFA* works:

$$\tilde{\mathbf{F}}_t^c = \mathit{CFA}(\tilde{\mathbf{F}}_t) = [\mathit{CE}_2(\tilde{\mathbf{F}}_t), \mathit{CE}_4(\tilde{\mathbf{F}}_t), \mathit{CE}_8(\tilde{\mathbf{F}}_t)], \quad (5)$$

where the $[\cdot]$ is the concatenation operation, followed by a convolution layer with the kernel size of 3×3 for dimension reduction. $\mathit{CE}_r(\cdot)$ represents the contrast-enhancement mechanism (Ding et al. 2018) with a convolution dilation rate of r, with:

$$\mathit{CE}_r(\tilde{\mathbf{F}}_t) = f_l(\tilde{\mathbf{F}}_t) - f_r(\tilde{\mathbf{F}}_t), \quad (6)$$

where the $f(\cdot)$ represents the feature extraction unit that consists of a convolution layer with the kernel size of 3×3, a batch normalization, and a ReLU layer. $l = 1$ and r indicate dilation rate in $f(\cdot)$. In our method, we resort to the multi-scale parallel dilation mechanism (*i.e.*, $r \in \{2, 4, 8\}$) to merge the reflection features of glasses of different distances and sizes in the surrounding regions.

With the contextual-enhanced reflection features $\tilde{\mathbf{F}}_t^c$, $\tilde{\mathbf{F}}_{t+1}^c$, $\tilde{\mathbf{F}}_n^c$ extracted from the *CFA*, we first concatenate them to form a new feature $\mathbf{F} \in \mathbb{R}^{T\times C_f\times H_f\times W_f}$, where C_f, H_f, and W_f represent the channels, height, and width of $\mathbf{F}$, respectively, and T denotes the number of frames, set to 3 by default. $\mathbf{F}$ is reshaped to $\mathbf{F}_T \in \mathbb{R}^{(H_f\times W_f)\times T\times C_f}$ and fed to the Temporal Transformer Block (*TTB*) for inter-frame temporal information integration. The temporal-enhanced features are also reshaped to $\mathbf{F}_S \in \mathbb{R}^{T\times(H_f\times W_f)\times C_f}$ and fed to the Spatial Transformer Block (*STB*) for further inter-frame spatial information integration. The workflow is as follows:

$$\tilde{\mathbf{F}}_t^{ts}, \tilde{\mathbf{F}}_{t+1}^{ts}, \tilde{\mathbf{F}}_n^{ts} = \mathit{STB}(\mathit{TTB}(\tilde{\mathbf{F}}_t^c, \tilde{\mathbf{F}}_{t+1}^c, \tilde{\mathbf{F}}_n^c)). \quad (7)$$

TTB and *STB* share the same structure (*i.e.*, a vision transformer block (Dosovitskiy et al. 2020)), but are applied to different input dimensions. In this way, both reliable local features of the LRE and weakly responsive features from discontinuous glass regions are enhanced by temporal-spatial inter-frame feature integration.

Loss Function

We train our network with two loss items: the glass surface supervision term and the auxiliary reflection supervision term. The total loss $\mathcal{L}$ can be written as:

$$\mathcal{L} = \sum_{i\in\{t,\,t+1,\,n\}} \Big(\sum_{j\in\{c,\,f\}} \mathcal{L}_{mask}(\mathbf{M}_i^j, \mathbf{M}_i^*) + \mathcal{L}_{ref}(\mathbf{R}_i, \mathbf{R}_i^*)\Big), \quad (8)$$

where $\mathbf{M}_i^c$ and $\mathbf{M}_i^f$ are predicted intermediate and final glass masks. $\mathbf{R}_i$ is the predicted reflection map. $\mathbf{M}_i^*$ and $\mathbf{R}_i^*$ denote the ground truth glass masks and reflections. We adopt the pixel position-aware loss (Wei, Wang, and Huang 2020) as the mask loss $\mathcal{L}_{mask}$. For the reflection loss, we employ the reflection removal method (Dong et al. 2021) to generate pseudo-GT reflection maps, and $\mathcal{L}_{ref}$ is:

$$\mathcal{L}_{ref}(\mathbf{R}_i, \mathbf{R}_i^*) = \mathcal{L}_{mse}(\mathbf{R}_i, \mathbf{R}_i^* \odot \mathbf{P}_i^*) + \mathcal{L}_{pen}(\mathbf{R}_i, \mathbf{P}_i^*), \quad (9)$$

where $\odot$ denotes element-wise multiplication. $\mathcal{L}_{mse}$ is standard MSE loss. $\mathcal{L}_{pen}$ is a penalty item to constrain the reflection regions to exist within the glass regions only, as:

$$\mathcal{L}_{pen}(\mathbf{R}_i, \mathbf{P}_i^*) = ||\mathbf{R}_i \odot \mathbf{P}_i^* - \mathbf{R}_i||^2. \quad (10)$$

Video Glass Surface Detection Dataset

To facilitate the learning of the video glass surface detection problem, we build the first large-scale video glass surface detection dataset, named VGSD-D. It includes 19,166 image frames from 297 videos with diverse scenes, where all frames are carefully annotated with the corresponding masks. Some example video frames are shown in Fig. 5.

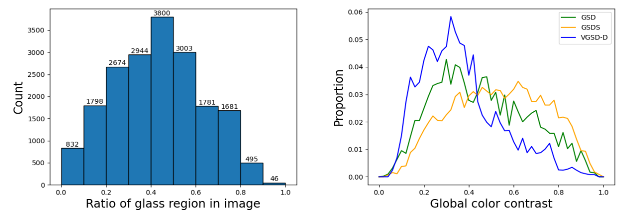
Dataset Construction

We use smartphones to collect the majority of videos with glass surfaces in diverse daily-life scenes (*e.g.*, office, classroom, mall), and we also collect another four video clips from the existing VSOD datasets (Perazzi et al. 2016; Xu et al. 2018). After collecting all videos, we manually trim the videos to make sure that each frame has at least one glass region. Then, we can obtain 297 video sequences with 19,166 image frames and 575 seconds of duration, where all

Figure 5: Visual display (left: frames; right: masks) of several examples of proposed Video Glass Surface Detection dataset.

Dataset	#Videos	#Labeled Frames	Time (s).	Max Reso.
GSD	-	4012	-	3456×4608
GSDS	-	4519	-	1024×1280
DAVIS	50	3455	144	1920×1080
VOS	200	7467	3870	800×800
DAVSOD	226	23,938	798	360×640
Visha	120	11,685	390	1280×720
VMD	269	15,066	502	1920×1080
Ours	297	19,166	575	1920×1080

Table 1: Statistical comparison between the dataset for relevant tasks and our proposed VGSD-D dataset.



(a) glass area distribution (b) color contrast distribution

Figure 6: Statistics of the constructed VGSD-D dataset.

frames are carefully annotated with corresponding ground truth glass surface masks by professional annotators. They are randomly divided into a training set (12,315 frames from 192 videos) and a testing set (6,851 frames from 105 videos). The frame rate is 30 fps for all video sequences.

Dataset Analysis

Table 1 summarizes our VGSD-D statistics compared to prior datasets from the relevant areas, including image-based glass detection (GSD (Lin, He, and Lau 2021) and GSDS (Lin, Yeung, and Lau 2022a)), salient video object detection (DAVIS (Perazzi et al. 2016), VOS (Li, Xia, and Chen 2017) and DAVSOD (Fan et al. 2019)), video shadow detection (Visha (Chen et al. 2021)) and video mirror detection (Lin, Tan, and Lau 2023). Fig. 6 provides a statistical analysis of the glass surface properties in our dataset.

Area distribution. Fig. 6(a) shows the ratio of glass area over the image area (glass area distribution). Our dataset contains mirrors covering a wide range of area ratios, and most glasses fall in the range of [0.1, 0.8]. Glass fall in the range of (0, 0.1] corresponds to images wherein the glass region is relatively small or situated distantly in the background. Detecting and classifying such glass surfaces could pose a considerable challenge for models due to potential distractions from the surrounding context. Conversely, ratios in the range [0.8, 1.0) represent images where the glass region dominates or entirely occupies the frame. Although detection in these instances may be less complex, comprehending the context of the image could also be a hurdle.

Color contrast distribution. We also analyze the global color contrast between the glass and non-glass regions by calculating the χ^2 distance between their RGB histograms, following the approach in (Li et al. 2014). Additionally, we compare this color contrast distribution with two existing image-based glass detection datasets, GSD (Lin, He, and Lau 2021) and GSDS (Lin, Yeung, and Lau 2022a), as shown in Fig. 6(b). It demonstrates that VGSD-D has a higher proportion of images with low color contrasts ($<$0.4) than the GSD and GSDS datasets. This results in increased complexity in detecting glass regions, underscoring the distinctiveness and challenges of the VGSD-D dataset.

Experiments

Implementation Details and Evaluation Metrics

We build our method using PyTorch toolbox and conduct all experiments on a Tesla V100 GPU with 32 GB memory. We adopt the Adam optimizer with a weight decay of 5×10^{-4} and a maximum learning rate of 5×10^{-5}. The cosine learning rate scheduler and warm-up are used to adjust the learning rate. The batch size and training epochs are 5 and 15. The input images were randomly flipped horizontally and were resized to 416×416 for network training. We employ ResNext-101 (Xie et al. 2017) pre-trained on ImageNet as the encoder. We set the number of masked deformable blocks in LRE to $m = 4$. The window size and k in MDA are empirically set to 7×7 and 4. The mask head contains two convolution layers with a batch normalization operation and a Sigmoid activation function.

We adopt four widely used dense prediction evaluation metrics: Intersection over Union (IoU), pixel accuracy, Balance Error Rate (BER), and Mean Absolute Error (MAE), to evaluate the performance of our video glass detection model.

Comparing to the State-of-the-art Methods

We systematically evaluate the efficacy of the proposed method by comparing it with 17 state-of-the-art methods from 7 relevant tasks, including salient object detection (GateNet (Zhao et al. 2020), MINet (Pang et al.

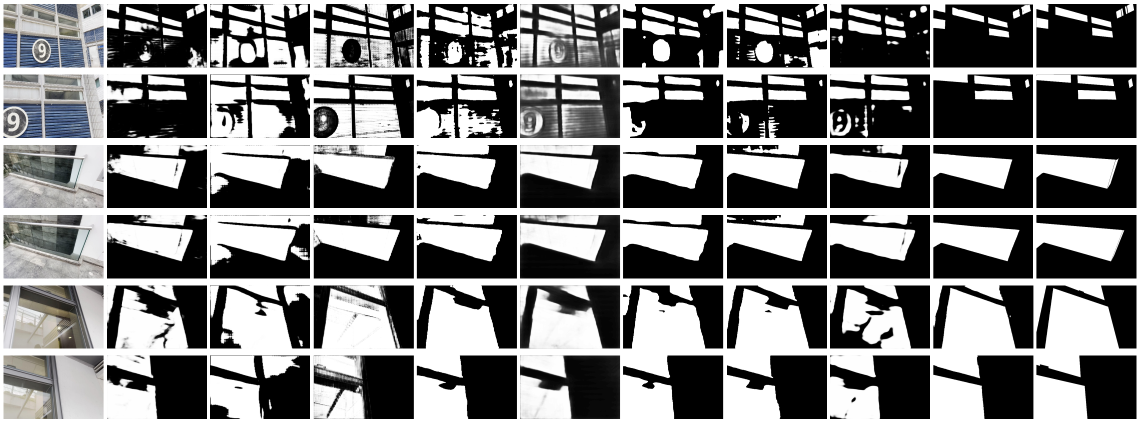


Figure 7: Qualitative comparison of eight state-of-the-art methods from seven relevant tasks and our approach.

Methods	Task	IoU↑	Accuracy↑	BER↓	MAE↓
GateNet	*SOD*	0.657	0.806	19.63	0.203
MINet		0.697	0.842	15.69	0.163
ZoomNet		0.741	0.865	13.30	0.138
UFO	*VSOD*	0.634	0.745	22.43	0.254
DeepLab	*SS*	0.705	0.845	16.67	0.155
Segformer		0.744	0.855	13.50	0.145
SAM		0.710	0.832	15.15	0.172
TVSD	*VSD*	0.728	0.860	13.52	0.140
SC-Cor		0.765	0.875	12.15	0.125
MirrorNet	*MD*	0.740	0.863	13.44	0.200
PMDNet		0.765	0.879	11.47	0.181
VCNet		0.751	0.873	12.17	0.168
VMD	*VMD*	0.763	0.878	12.44	0.123
GDNet	*GSD*	0.735	0.858	13.18	0.172
EBLNet		0.764	0.868	13.25	0.134
GlassNet		0.762	0.877	12.02	0.187
PGSNet		0.703	0.846	15.11	0.156
Ours		**0.802**	**0.899**	**9.54**	**0.099**

Table 2: Quantitative comparisons of our method with 17 relevant methods from 7 relevant tasks. Best results are shown in bold.

2020), ZoomNet (Pang et al. 2022)), video salient object detection (UFO (Su et al. 2023)), semantic segmentation (DeepLab (Chen et al. 2017), Segformer (Xie et al. 2021), SAM (Kirillov et al. 2023)), video shadow detection (TVSD (Chen et al. 2021), SC-Cor (Ding et al. 2022)), mirror detection (MirrorNet (Yang et al. 2019), PMDNet (Lin, Wang, and Lau 2020), VCNet (Tan et al. 2022)), video mirror detection (VMD (Lin, Tan, and Lau 2023)), glass detection (GDNet (Mei et al. 2020), EBLNet (He et al. 2021), GlassNet (Lin, He, and Lau 2021), PGSNet (Yu et al. 2022)).

Table 2 shows the quantitative comparison, and our method achieves the best performance on all metrics. Specifically, the proposed method outperforms the best single-image glass detection method GlassNet by 20.63% In comparison to the semantic segmentation methods, *e.g.*, Segformer and SAM, our method still outperforms them by a large margin. Particularly, SAM tends to misclassify the objects behind glass surfaces as the non-glass, due to the intricate reflection and refraction phenomena. As shown in Table 3, we also conduct efficiency comparisons of our method with existing glass detection methods, including the assessments of model parameters, FLOPs, and FPS. The results demonstrate that our model has fewer parameters and is faster than state-of-the-art image-based glass detection methods. In particular, our method surpasses PGSNet in IoU by 14.08%, with approximately $3\times$ fewer parameters and 21.63% FLOPs reduction.

Methods	Params.↓	FLOPs↓	FPS↑	IoU↑
GDNet	201.72M	271.69G	5.90	0.735
EBLNet	111.45M	303.86G	8.79	0.764
GlassNet	83.72M	108.98G	5.92	0.762
PGSNet	198.12M	113.02G	7.14	0.703
Ours	**64.06M**	**88.55G**	**15.04**	**0.802**

Table 3: Efficiency comparison between existing glass detection methods and our approach.

We also provide the visual comparisons with state-of-the-art methods in Fig. 7. Notably, most competing methods are susceptible to interference by objects in the non-glass region that are similar in shape or appearance to the glass surface. (*e.g.*, the number 9 in the first case, and the walls behind the glass in the second case). In addition, objects inside the glass can also interfere with the glass detection process, causing existing competing methods to misidentify them as non-glass areas (*e.g.*, window frames in the third case). In contrast, our method can reduce the distraction of these non-glass regions and detect all glass surfaces accurately with the incorporation of the location-aware and context-enhanced

	B	CRE	LRE	CRI *CFA*	CRI *TTB*	CRI *STB*	IoU↑	Accuracy↑	BER↓	MAE↓
①	✓						0.734	0.865	13.35	0.136
②	✓	✓					0.743	0.870	13.10	0.129
③	✓		✓				0.751	0.874	12.95	0.126
④	✓		✓	✓			0.774	0.883	11.39	0.117
⑤	✓		✓	✓	✓		0.794	0.895	10.32	0.105
⑥	✓		✓	✓		✓	0.786	0.892	10.38	0.107
⑦	✓		✓	✓	✓	✓	**0.802**	**0.899**	**9.54**	**0.099**

Table 4: Ablation analysis of the proposed network structure on the proposed VGSD-D dataset. B denotes the baseline network that uses only encoder and mask head for glass mask prediction. CRE indicates that we substitute the mask concatenation for the explicit position usage in the LRE.

Methods	IoU↑	Accuracy↑	BER↓	MAE↓
❶ Ours w/o $\mathcal{L}_{mask}(\mathbf{M}^c, \mathbf{M}^*)$	0.781	0.884	10.88	0.117
❷ Ours w/o $\mathcal{L}_{ref}(\mathbf{R}, \mathbf{R}^*)$	0.753	0.877	12.04	0.124
❸ Ours w/o $\mathcal{L}_{pen}(\mathbf{R}, \mathbf{P}^*)$	0.783	0.889	10.91	0.112
❹ Ours	**0.802**	**0.899**	**9.54**	**0.099**

Table 5: Ablation analysis of the used loss functions on the proposed VGSD-D dataset. All subscripts of the $\mathbf{M}$, $\mathbf{R}$, and $\mathbf{P}$ are omitted for clarity.

multi-view reflection cues.

Ablation Study

We perform internal analyses to verify the effectiveness of each component of our approach. ① We first construct the baseline model using the encoder, mask head, and $\mathcal{L}_{mask}(\mathbf{M}^c, \mathbf{M}^*)$. To validate the LRE module, in ②, we build a variant of the LRE by concatenating the predicted intermediate mask $\mathbf{M}^c$ with $\mathbf{I}$, $\mathbf{F}^1$ and $\mathbf{F}^5$ to form a new input to the standard deformable attention block. Then, in ③, we directly incorporate the LRE module into the baseline. Based on the variant in ③, we gradually insert *CFA*, *TTB*, and *STB* in the CRI module to validate their effectiveness in ④-⑥. Note that the total loss $\mathcal{L}$ are utilized for ②-⑦.

The quantitative results presented in Table 4, with following conclusions: (1) introducing reflection cues does improve model performance compared to the baseline (see ① *vs* ②); (2) explicit position map is more efficient than the naive implicit concatenation (see ② *vs* ③); (3) contextual semantics introduced by *CFA* (④) can boost the performance improvement even without cross-frames integration; and (4) While the *TTB* achieves slightly better performance than the *STB* (⑤ *vs* ⑥), the concurrent utilization of both elements promotes a significant performance improvement (⑤ *vs* ⑦ and ⑥ *vs* ⑦), which demonstrates the complementarity of the two blocks. We also display the visual comparisons in Fig. 8. Obviously, the baseline model can only locate part of the glass surfaces and is not sensitive to reflection cues. LRE can include more glass surface regions with the help of position-aware reflection features (see the left corner of

Figure 8: Visual comparison of different ablated models. B denotes the baseline in Table 4.

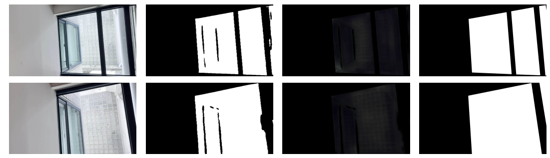


Figure 9: Failure cases illustration.

the glass surfaces). Finally, RCI further helps to improve regions with weak reflection within a single frame by fusing information between multiple frames.

In Table 5, we also investigate the impact of various loss components by independently disabling the intermediate mask supervision and the auxiliary reflection supervision in ❶ and ❷, and ablated the effect of $\mathcal{L}_{pen}$ in ❸ by removing it in the $\mathcal{L}_{ref}$. The results demonstrate that: (1) Intermediate mask supervision is essential because it ensures the accuracy of the position map used in the LRE module; (2) The reflection prior is the key of our model and is indispensable, without it, the BER metric plummets by 20.77%; and (3) The penalty term can further boost the glass detection accuracy by constraining the regions of generated reflection maps.

Conclusion

In this paper, we have explored the video glass surface detection problem. We address this problem from two aspects. First, we have proposed a VGSD-Net for video glass surface detection, which includes two novel modules: the Location-aware Reflection Extraction (LRE) module for extracting position-aware localized reflection features via masked deformable attention-based blocks, and the Context-enhanced Reflection Integration (CRI) module for incorporating multi-view dynamic reflections from the video sequences. Second, we have built the first large-scale video glass surface detection dataset. It contains 19,166 image frames from 297 videos (lasting 575 seconds in total). Finally, experimental comparisons also show that our method outperforms state-of-the-art methods from relevant tasks.

Nonetheless, our approach does have limitations. If temporal reflections in some regions of all video frames are too weak to be detected, our method may fail to detect all glass surfaces accurately. Fig. 9 shows that our method will incorrectly classify the frames of the window behind the main glass surfaces as non-glass regions due to weak reflection on these regions. Incorporating inherent structure cues (Liu et al. 2023c) in images is a potential solution.

Acknowledgments

This project is in part supported by a GRF grant from the Research Grants Council of Hong Kong (No.: 11211223).

References

Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; and Van Den Hengel, A. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 3674–3683.

Bhat, S. F.; Alhashim, I.; and Wonka, P. 2021. AdaBins: Depth Estimation Using Adaptive Bins. In *CVPR*, 4009–4018.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, Z.; Wan, L.; Zhu, L.; Shen, J.; Fu, H.; Liu, W.; and Qin, J. 2021. Triple-cooperative video shadow detection. In *CVPR*, 2715–2724.

Ding, H.; Jiang, X.; Shuai, B.; Liu, A. Q.; and Wang, G. 2018. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *CVPR*, 2393–2402.

Ding, X.; Yang, J.; Hu, X.; and Li, X. 2022. Learning shadow correspondence for video shadow detection. In *ECCV*, 705–722.

Dong, Z.; Xu, K.; Yang, Y.; Bao, H.; Xu, W.; and Lau, R. W. 2021. Location-aware single image reflection removal. In *ICCV*, 5017–5026.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Fan, D.-P.; Wang, W.; Cheng, M.-M.; and Shen, J. 2019. Shifting more attention to video salient object detection. In *CVPR*, 8554–8564.

Fritz, M.; Bradski, G.; Karayev, S.; Darrell, T.; and Black, M. 2009. An additive latent feature model for transparent object recognition. *NeurIPS*, 22.

Gu, Y.; Wang, L.; Wang, Z.; Liu, Y.; Cheng, M.-M.; and Lu, S.-P. 2020. Pyramid constrained self-attention network for fast video salient object detection. In *AAAI*, volume 34, 10869–10876.

Guan, H.; Lin, J.; and Lau, R. W. 2022. Learning semantic associations for mirror detection. In *CVPR*, 5941–5950.

Guo-Hua, C.; Jun-Yi, W.; and Ai-Jun, Z. 2019. Transparent object detection and location based on RGB-D camera. In *Journal of Physics: Conference Series*, volume 1183, 012011. IOP Publishing.

He, H.; Li, X.; Cheng, G.; Shi, J.; Tong, Y.; Meng, G.; Prinet, V.; and Weng, L. 2021. Enhanced boundary learning for glass-like object segmentation. In *ICCV*, 15859–15868.

He, R.; Lin, J.; and Lau, R. W. 2023. Efficient mirror detection via multi-level heterogeneous learning. In *AAAI*, volume 37, 790–798.

Huang, T.; Dong, B.; Lin, J.; Liu, X.; Lau, R. W.; and Zuo, W. 2023. Symmetry-aware transformer-based mirror detection. In *AAAI*, volume 37, 935–943.

Kalra, A.; Taamazyan, V.; Rao, S. K.; Venkataraman, K.; Raskar, R.; and Kadambi, A. 2020. Deep polarization cues for transparent object segmentation. In *CVPR*, 8602–8611.

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, H.; Chen, G.; Li, G.; and Yu, Y. 2019. Motion guided attention for video salient object detection. In *ICCV*, 7274–7283.

Li, J.; Xia, C.; and Chen, X. 2017. A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection. *IEEE TIP*, 27(1): 349–364.

Li, Y.; Hou, X.; Koch, C.; Rehg, J. M.; and Yuille, A. L. 2014. The secrets of salient object segmentation. In *CVPR*, 280–287.

Lin, J.; He, Z.; and Lau, R. W. 2021. Rich context aggregation with reflection prior for glass surface detection. In *CVPR*, 13415–13424.

Lin, J.; and Lau, R. W. 2023. Self-supervised Pre-training for Mirror Detection. In *ICCV*, 12227–12236.

Lin, J.; Tan, X.; and Lau, R. W. 2023. Learning To Detect Mirrors From Videos via Dual Correspondences. In *CVPR*, 9109–9118.

Lin, J.; Wang, G.; and Lau, R. W. 2020. Progressive mirror detection. In *CVPR*, 3697–3705.

Lin, J.; Yeung, Y.-H.; and Lau, R. 2022a. Exploiting Semantic Relations for Glass Surface Detection. *NeurIPS*, 35: 22490–22504.

Lin, J.; Yeung, Y. H.; and Lau, R. W. 2022b. Depth-aware glass surface detection with cross-modal context mining. *arXiv preprint arXiv:2206.11250*.

Liu, F.; Kong, Y.; Zhang, L.; Feng, G.; and Yin, B. 2023a. Local-global coordination with transformers for referring image segmentation. *Neurocomputing*, 522: 39–52.

Liu, F.; Liu, Y.; Kong, Y.; Xu, K.; Zhang, L.; Yin, B.; Hancke, G.; and Lau, R. 2023b. Referring image segmentation using text supervision. In *ICCV*, 22124–22134.

Liu, Y.; Guo, Q.; Fu, L.; Ke, Z.; Xu, K.; Feng, W.; Tsang, I. W.; and Lau, R. W. 2023c. Structure-Informed Shadow Removal Networks. *IEEE TIP*.

Liu, Y.; Xie, J.; Qiao, Y.; Tang, Y.; and Yang, X. 2021a. Prior-induced information alignment for image matting. *IEEE TMM*.

Liu, Y.; Xie, J.; Shi, X.; Qiao, Y.; Huang, Y.; Tang, Y.; and Yang, X. 2021b. Tripartite information mining and integration for image matting. In *ICCV*, 7555–7564.

Mei, H.; Dong, B.; Dong, W.; Peers, P.; Yang, X.; Zhang, Q.; and Wei, X. 2021. Depth-aware mirror segmentation. In *CVPR*, 3044–3053.

Mei, H.; Dong, B.; Dong, W.; Yang, J.; Baek, S.-H.; Heide, F.; Peers, P.; Wei, X.; and Yang, X. 2022. Glass segmentation using intensity and spectral polarization cues. In *CVPR*, 12622–12631.

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*, 3687–3696.

Pang, Y.; Zhao, X.; Xiang, T.-Z.; Zhang, L.; and Lu, H. 2022. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR*, 2160–2170.

Pang, Y.; Zhao, X.; Zhang, L.; and Lu, H. 2020. Multi-scale interactive network for salient object detection. In *CVPR*, 9413–9422.

Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Van Gool, L.; Gross, M.; and Sorkine-Hornung, A. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, 724–732.

Qiao, Y.; Dong, B.; Jin, A.; Fu, Y.; Baek, S.-H.; Heide, F.; Peers, P.; Wei, X.; and Yang, X. 2023. Multi-view Spectral Polarization Propagation for Video Glass Segmentation. In *ICCV*, 23218–23228.

Su, Y.; Deng, J.; Sun, R.; Lin, G.; Su, H.; and Wu, Q. 2023. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. *IEEE TMM*.

Tan, J.; Lin, W.; Chang, A. X.; and Savva, M. 2021. Mirror3D: Depth refinement for mirror surfaces. In *CVPR*, 15990–15999.

Tan, X.; Lin, J.; Xu, K.; Chen, P.; Ma, L.; and Lau, R. W. 2022. Mirror detection with the visual chirality cue. *IEEE TPAMI*, 45(3): 3492–3504.

Tibebu, H.; Roche, J.; De Silva, V.; and Kondoz, A. 2021. Lidar-based glass detection for improved occupancy grid mapping. *Sensors*, 21(7): 2263.

Wang, W.; Shen, J.; and Porikli, F. 2015. Saliency-aware geodesic video object segmentation. In *CVPR*, 3395–3402.

Wang, W.; Shen, J.; and Shao, L. 2017. Video salient object detection via fully convolutional networks. *IEEE TIP*, 27(1): 38–49.

Wei, J.; Wang, S.; and Huang, Q. 2020. F^3Net: fusion, feedback and focus for salient object detection. In *AAAI*, volume 34, 12321–12328.

Wei, Y.; Wen, F.; Zhu, W.; and Sun, J. 2012. Geodesic saliency using background priors. In *ECCV*, 29–42.

Xie, E.; Wang, W.; Wang, W.; Ding, M.; Shen, C.; and Luo, P. 2020. Segmenting transparent objects in the wild. In *ECCV*, 696–711.

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*, 34: 12077–12090.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *CVPR*, 1492–1500.

Xu, N.; Yang, L.; Fan, Y.; Yang, J.; Yue, D.; Liang, Y.; Price, B.; Cohen, S.; and Huang, T. 2018. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, 585–601.

Yang, S.-W.; and Wang, C.-C. 2008. Dealing with laser scanner failure: Mirrors and windows. In *ICRA*, 3009–3015. IEEE.

Yang, X.; Mei, H.; Xu, K.; Wei, X.; Yin, B.; and Lau, R. W. 2019. Where is my mirror? In *ICCV*, 8809–8818.

Ye, S.; Chen, D.; Han, S.; and Liao, J. 2021. Learning with noisy labels for robust point cloud segmentation. In *ICCV*, 6443–6452.

Ye, S.; Chen, D.; Han, S.; and Liao, J. 2022a. 3D question answering. *IEEE TVCG*.

Ye, S.; Chen, D.; Han, S.; and Liao, J. 2022b. Robust Point Cloud Segmentation With Noisy Annotations. *IEEE TPAMI*, 45(6): 7696–7710.

Yu, L.; Mei, H.; Dong, W.; Wei, Z.; Zhu, L.; Wang, Y.; and Yang, X. 2022. Progressive glass segmentation. *IEEE TIP*, 31: 2920–2933.

Zhang, M.; Liu, J.; Wang, Y.; Piao, Y.; Yao, S.; Ji, W.; Li, J.; Lu, H.; and Luo, Z. 2021. Dynamic context-sensitive filtering network for video salient object detection. In *ICCV*, 1553–1563.

Zhao, X.; Pang, Y.; Zhang, L.; Lu, H.; and Zhang, L. 2020. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, 35–51.

Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; and Lau, R. W. 2023. BiFormer: Vision Transformer with Bi-Level Routing Attention. In *CVPR*, 10323–10333.