MVGD-Net: A Novel Motion-aware Video Glass Surface Detection Network

Yiwei Lu*, **Hao Huang***, **Tao Yan**†

School of Artificial Intelligence and Computer Science, Jiangnan University, China
{6243112037, 6223110021}@stu.jiangnan.edu.cn, yantao.ustc@gmail.com

Abstract

Glass surface ubiquitous in both daily life and professional environments presents a potential threat to vision-based systems, such as robot and drone navigation. To solve this challenge, most recent studies have shown significant interest in Video Glass Surface Detection (VGSD). We observe that objects in the reflection (or transmission) layer appear farther from the glass surfaces. Consequently, in video motion scenarios, the notable reflected (or transmitted) objects on the glass surface move slower than objects in non-glass regions within the same spatial plane, and this motion inconsistency can effectively reveal the presence of glass surfaces. Based on this observation, we propose a novel network, named MVGD-Net, for detecting glass surfaces in videos by leveraging motion inconsistency cues. Our MVGD-Net features three novel modules: the Cross-scale Multimodal Fusion Module (CMFM) that integrates extracted spatial features and estimated optical flow maps, the History Guided Attention Module (HGAM) and Temporal Cross Attention Module (TCAM), both of which further enhances temporal features. A Temporal-Spatial Decoder (TSD) is also introduced to fuse the spatial and temporal features for generating the glass region mask. Furthermore, for learning our network, we also propose a large-scale dataset, which comprises 312 diverse glass scenarios with a total of 19,268 frames. Extensive experiments demonstrate that our MVGD-Net outperforms relevant state-of-the-art methods.

Code and Datasets —
https://github.com/YT3DVision/MVGDNet

Instruction

Glass surfaces, such as glass windows, walls and doors, are ubiquitous in our everyday lives. They are always transparent and colorless, posing significant challenges for computer vision tasks, such as robot and drone navigation (Li et al. 2024), depth estimation (Liang et al. 2023) and 3D reconstruction (Li, Yeh, and Chandraker 2020). Previous methods have explored various prior cues or assumptions for single-image glass surface detection (GSD), such as contrasted context features (Mei et al. 2020), boundary cues (He et al. 2021), reflection phenomena (Lin, He, and Lau 2021), ghosting effect (Yan et al. 2025), semantic relation (Lin, Yeung, and Lau 2022), and visual blurriness (Qi et al. 2024). Multimodal image-based GSD methods have also explored RGB-Depth (Lin et al. 2025), Polarization (Mei et al. 2022), RGB-Thermal (Huo et al. 2023) and RGB-NIR (Yan et al. 2024) for more effective and robust GSD.

Figure 1: Examples of the motion inconsistency in indoor and outdoor scenes. The $1st$ and $2nd$ columns show the frames at time t and time $t+10$, respectively. The $3rd$ column shows the inconsistent motion cues on glass surfaces, indicated by arrows. The $4th$ column shows the optical flow maps.

However, real-world applications like robotic navigation and autonomous driving are video-centric rather than image-centric, and above methods are tailored for single images, failing to leverage temporal information. Thus, most recently, Liu *et al.* (Liu et al. 2024) proposed the first VGSD method, named VGSD-Net, which leverages temporal information in videos to improve detection. Though VGSD-Net exploits reflections from each frame to aid GSD, it may fail to obtain accurate reflections in challenging scenes (as shown in Fig. 2) due to the absence of ground-truth reflection supervision.

Neuroscience studies demonstrate that humans rely on dynamic perceptual cues to identify glass regions in daily life. The references (Tamura et al. 2016; Tamura, Higashi, and Nakauchi 2017, 2018) reveal that rotational, parallax, forward, and backward motions can induce motion inconsistencies, which are useful cues for GSD. In real-world scenarios, we have observed that objects in the reflection (or transmission) layer appear farther from the glass surfaces. Consequently, in video motion scenarios, the notable reflected (or transmitted) objects on the glass surface move slower than objects in non-glass regions within the same spatial plane. As shown in the $1st$ scene of Fig. 1, reflection

*These authors contributed equally.
†Tao Yan is the corresponding author.

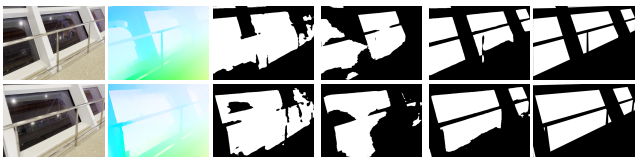


Figure 2: Existing methods may under/over-detect glass surfaces in challenge scenes. Our method utilizes motion inconsistency cues to guide GSD and outperforms competitors.

moves more slowly than non-glass regions when the camera moves in front of glass surfaces. These inconsistent motion cues imply the existence of reflections, which could be further exploited for GSD. Moreover, as illustrated in the $2nd$ scene of Fig. 1, even in indoor scenes with weak reflections, motion inconsistency can still arise due to transmitted objects being located at greater depths. Unlike Warren *et al.* (Warren et al. 2024)'s mirror detection method exploiting motion inconsistency, our method further accounts for special cases of GSD, such as open doors and windows, by applying the primary mask to filter inconsistent motion cues from obvious non-glass regions. Specifically, we adopt RAFT (Teed and Deng 2020) to calculate optical flow maps, which can reveal the potential location of glass surfaces, as shown in Fig. 2.

In this paper, we propose the **M**otion-aware **V**ideo **G**lass Surface **D**etection **Net**work (MVGD-Net), comprising three key modules: (1) the Cross-scale Multimodal Fusion Module (CMFM), which integrates features from two modalities across multiple scales; (2) the History-Guided Attention Module (HGAM), which leverages historical information to refine current frame predictions; and (3) the Temporal-Spatial Decoder (TSD), which balances the fusion of temporal and spatial features in the decoding process. To train our network, we also construct a large-scale dataset, MVGD-D, comprising $19,268$ frames across 312 videos, with manually annotated glass surface masks for each frame. Our main contributions are summarized as follows:

- We propose a novel network, named MVGD-Net, which exploits motion inconsistency from video sequence for VGSD task. MVGD-Net features three novel modules: the CMFM for cross-scale multimodal fusion, the HGAM for exploiting temporal context, and the TSD for effectively balancing temporal and spatial feature integration.
- We propose the MVGD-D dataset, a large-scale dataset containing 312 videos with a total of $19,268$ frames, each with corresponding manual annotations, for GSD.
- Extensive experiments demonstrate the outperformance of our proposed method compared with the relevant state-of-the-art methods.

Related Work

Glass Surface Detection (GSD). Early works primarily focused on leveraging single-frame visual cues, such as contrasted contextual features (Mei et al. 2020), reflection priors (Lin, He, and Lau 2021), boundary detection (He et al. 2021; Fan et al. 2023), semantic correlations (Lin, Yeung, and Lau 2022), blurry effects (Qi et al. 2024), and ghosting cues (Yan et al. 2025). To capture more comprehensive scene information, recent methods have explored multi-modal imaging for GSD. For example, RGB-polarization (Mei et al. 2022), RGB-thermal (Huo et al. 2023), RGB-NIR (Yan et al. 2024), and RGB-D (Lin et al. 2025) approaches leverage additional inputs to provide richer and more distinctive glass cues compared to RGB-only methods.

Despite these advancements, all the above methods are designed for singles images and fail to utilize the temporal information present in video sequences.

Video Salient Object Detection (VSOD). Video Salient Object Detection aims to identify the most visually significant objects across video frames by integrating spatial and temporal information. To improve detection accuracy, early works have utilized motion cues (Wang, Shen, and Shao 2017) and optical flow maps (Li et al. 2019) to localize salient objects. More recent approaches (Oh et al. 2019; Cheng, Tai, and Tang 2021; Cheng et al. 2024) further enhance temporal consistency by introducing memory modules. However, these methods often detect the objects behind the glass rather than the glass surface itself, due to the glass region may not always represent the most visually significant objects within a scene.

Our Method

Pipeline

Our key observation is that glass surfaces always show motion inconsistency compared to non-glass regions located in the same position. This phenomenon arises from the varied depth perception created by the reflection (or transmission) layer. This observation motivates us to fully exploit motion inconsistency contained within optical flow maps to aid VGSD. The structure of our proposed MVGD-Net is shown in Fig. 3.

Specifically, three adjacent frames (I_{N-2}, I_{N-1}, I_N) are first input into RAFT (Teed and Deng 2020) to estimate optical flow maps (f_{N-1} and f_N), while simultaneously passing through a weight-shared Swin backbone for feature extraction. A coarse glass surface mask P_{N-1} is generated by direct decoding of the intermediate features $\{G_i^{N-1}\}_{i=1}^4$, which is then used to exclude inconsistent motion cues from non-glass areas. Following feature extraction via another weight-shared Swin transformer, the optical flow features ($\{O_i^{N-1}\}_{i=1}^4$, $\{O_i^N\}_{i=1}^4$) and RGB features ($\{G_i^{N-2}\}_{i=1}^4$, $\{G_i^{N-1}\}_{i=1}^4$, $\{G_i^N\}_{i=1}^4$) are fed into the CMFM module to extract spatial features ($\{S_i^{N-1}\}_{i=1}^4$, $\{S_i^N\}_{i=1}^4$). At the same time, RGB features are passed through the TCAM and HGAM modules to capture temporal features ($\{T_i^{N-2}\}_{i=1}^4$, $\{T_i^{N-1}\}_{i=1}^4$, $\{T_i^N\}_{i=1}^4$). HGAM and TCAM aggregate temporal information across frames, stabilizing the model against local unreliable flow cues. Finally, the TSD module fuses the spatial features and the temporal features to produce the final prediction of the glass surface masks (M_{N-2}, M_{N-1}, M_N).

Cross-scale Multimodal Fusion Module (CMFM)

Camera shake and other factors can cause significant variations in motion between adjacent frames, so we estimate optical flow between adjacent frames to improve accuracy, where

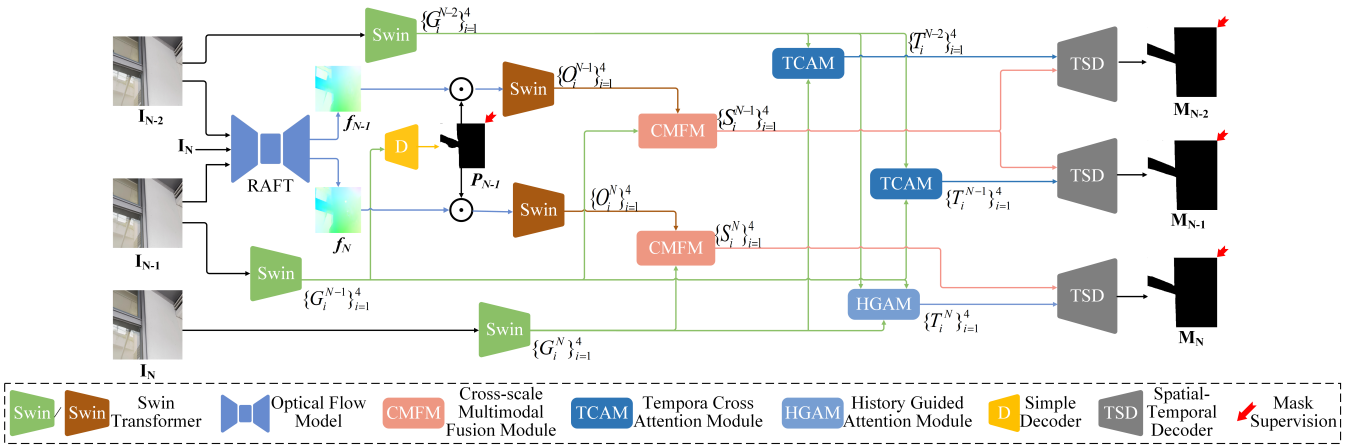


Figure 3: The structure of our proposed **M**otion-aware **V**ideo **G**lass Surface **D**etection **Net**work (MVGD-Net).

$f_{N-1} = \text{RAFT}(I_{N-2}, I_{N-1})$ and $f_N = \text{RAFT}(I_{N-1}, I_N)$. The optical flow maps f_{N-1} and f_N, refined by the glass mask P_{N-1}, can indicate the potential locations of the glass surfaces. After being encoded by a Swin-Transformer (Liu et al. 2022), the four-scale optical flow features $\{O_i^t\}_{i=1}^4$ and RGB features $\{G_i^t\}_{i=1}^4$ are fed into the CMFM to generate spatial features $\{S_i^t\}_{i=1}^4$, where $t \in \{N-1, N\}$.

The CMFM shown in Fig. 4 leverages the cues provided by optical flow to guide the learning of the GSD. The input features $\{G_i\}_{i=1}^4$ and $\{O_i\}_{i=1}^4$ are first all projected to the channel size C_1 to reduce the computational cost, as follows:

$$X_i^G = \text{Conv}(\text{CBAM}(G_i)), \tag{1}$$

where CBAM$(\cdot)$ (Woo et al. 2018) is a lightweight attention module for feature refinement, and Conv$(\cdot)$ denotes a 1×1 convolution used for channel dimension reduction. C_1 is set to 128. Similarly, X_i^O is generated from O_i in the same way.

CMFM consists of seven cross-scale cross-attention blocks, each with independent parameter matrices. The fusion of all features is achieved through a U-shaped loop that proceeds from left to right (top branch), and then from right to left (bottom branch). At each attention block, the new query features iteratively update the memory representations, enabling the progressive fusion of all eight feature maps.

In the left-to-right process of CMFM (top branch of Fig. 4), the feature maps are gradually downsampled, which can be regarded as a process of compressing spatial features to extract more effective representations. In the i-th attention block, the corresponding query, key, and value feature vectors are obtained via the projection matrices W_i^Q, W_i^K, and W_i^V, respectively. The feature vector X_i^K is then transposed using a reshape operation. While $i = 1$, $X_i^K = W_1^K(X_1^G)$, $X_i^V = W_1^V(X_1^G)$, and $X_{i+1}^Q = W_1^Q(X_2^O)$. The attention map $Att_{i+1,i}^{top}$ is then calculated as follow:

$$Att_{i+1,i}^{top} = \text{SoftMax}(X_{i+1}^Q \otimes X_i^K), \tag{2}$$

where $i \in [1, 3]$, and SoftMax$(\cdot)$ denotes the normalization operation that maps values to the range $[0, 1]$. The attention

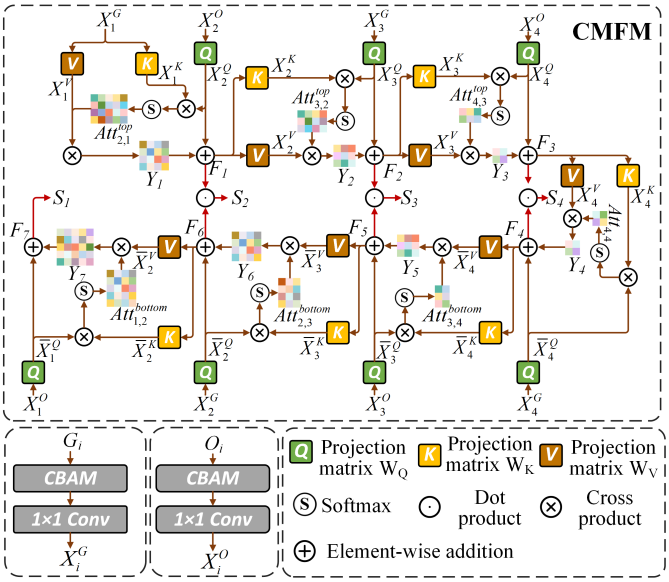


Figure 4: The structure of our CMFM.

map $Att_{i+1,i}^{top}$ is then applied to the memory (X_i^V) to obtain the fused features $Y_i \in \mathbb{R}^{H_{i+1}W_{i+1} \times C_1}$, as follows:

$$Y_i = Att_{i+1,i}^{top} \otimes X_i^V, \tag{3}$$

Subsequently, the query X_{i+1}^Q is concatenated with the fused features:

$$F_i = \text{R}(\text{MLP}(\text{LN}(Y_i))) + X_{i+1}^Q, \tag{4}$$

where MLP$(\cdot)$ represents the function of multi-layer perceptron, and LN$(\cdot)$ represents layer normalization. R$(\cdot)$ is the function of reshape operation. Particularly, while $i = 4$, X_4^O and X_4^G are serving as the connection point between the top and bottom branches.

In the right-to-left process of CMFM (bottom branch of CMFM), the feature maps are gradually upsampled, which can be regarded as a progressive enhancement of the important features. The attention map $Att_{j-1,j}^{bottom}$, as well as the

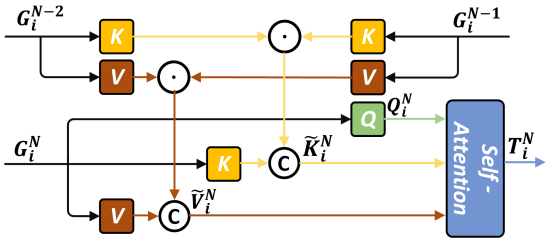


Figure 5: The structure of our HGAM.

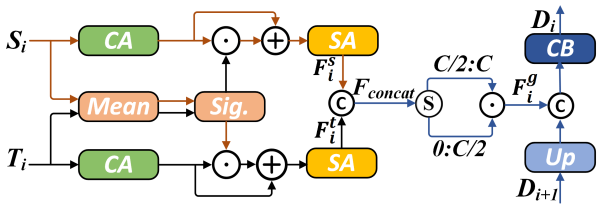


Figure 6: The structure of our TSD.

computation of Y_i and F_i, follows a similar procedure to that described above. Here, $j \in \{4, 3, 2\}$ and $i = 9 - j$.

Finally, the features F_i at the same scale are paired and fused via element-wise multiplication to obtain the final output features $\{S_i\}_{i=1}^4$. This process can be formulated as:

$$S_i = \begin{cases} F_7 & i = 1, \\ F_{i-1} \odot F_{8-i} & i = 2, 3, 4. \end{cases} \quad (5)$$

History Guided Attention Module (HGAM)

Adjacent frames in a video always contain rich temporal information, as the positions of glass surfaces remain roughly consistent between neighboring frames. Effectively leveraging this temporal similarity can greatly enhance the robustness of our method. Therefore, HGAM is proposed to enhance the current frame with previous frames.

Since the frame I_N is acquired after the frames I_{N-1} and I_{N-2}, its prediction M_N can be improved with the help of the predictions for the previous frames, termed M_{N-2} and M_{N-1}. As shown in Fig. 5, when the feature of current frame G_i^N is used as the query ($Q_i^N \in \mathbb{R}^{H_i W_i \times C_i}$), the K and V are calculated from G_i^{N-2} and G_i^{N-1}, yielding rich features $\tilde{K}_i^N \in \mathbb{R}^{C_i \times H_i W_i}$ and $\tilde{V}_i^N \in \mathbb{R}^{H_i W_i \times C_i}$, as follows:

$$\tilde{K}_i^N = [\mathrm{W_K}(G_i^{N-2}) \odot \mathrm{W_K}(G_i^{N-1}), \mathrm{W_K}(G_i^N)], \quad (6)$$

$$\tilde{V}_i^N = [\mathrm{W_V}(G_i^{N-2}) \odot \mathrm{W_V}(G_i^{N-1}), \mathrm{W_V}(G_i^N)], \quad (7)$$

where $\odot$ denotes element-wise multiplication, and $[\cdot]$ represents channel-wise concatenation. These features, along with the query Q_i^N, are then fed into a self-attention to generate the temporal output feature T_i^N, as follows:

$$T_i^N = \mathrm{SelfAttn}(Q_i^N, \tilde{K}_i^N, \tilde{V}_i^N). \quad (8)$$

Temporal Cross Attention Module (TCAM)

TCAM employs a standard cross-attention mechanism to capture inter-frame dependencies, while minimizing redundant computation. Specifically, $T_i^{N-1} = \mathrm{TCAM}(G_i^{N-1}, G_i^{N-2})$ and $T_i^{N-2} = \mathrm{TCAM}(G_i^{N-2}, G_i^N)$. The cross-attention between frame I_{N-1} and I_{N-2} captures short-term temporal dependencies and motion trends, while the cross-attention between frame I_{N-2} and I_N facilitates long-range temporal consistency and provides complementary information for recovering occluded or degraded regions. Taking T_i^{N-1} as an example, the feature G_i^{N-1} serves as the Query and the feature G_i^{N-2} serves as the Key and Value to form the attention map. Then, the cross-attention map serves as a residual term to produce the output features, as follows:

$$T_i^{N-1} = G_i^{N-1} + \mathrm{CrossAttn}(G_i^{N-1}, G_i^{N-2}, G_i^{N-2}). \quad (9)$$

and the process for the T_i^{N-2} is similar.

Temporal-Spatial Decoder (TSD)

The temporal feature set obtained by HGAM and TCAM has a channel configuration of $\{C_i\}_{i=1}^4 = \{2^{i-1}C_1\}_{i=1}^4$, while the spatial features produced by CMFMs share a uniform channel dimension of C_1 across all scales. This channel inconsistency between the temporal and spatial domains hinders direct feature fusion. Thus, TSD is proposed to balance the channel dimensions of both domains in the process of feature fusion and decoding.

As shown in Fig. 6, the input temporal features $T_i \in \mathbb{R}^{H_i \times W_i \times C_i}$ and spatial features $S_i \in \mathbb{R}^{H_i \times W_i \times C_1}$ are fed into a Channel Attention module ($\mathrm{CA}(\cdot)$) for enhancement. Then, feature weights are generated via a channel-wise mean and subsequent sigmoid activation. These weights then enhance the corresponding feature map through element-wise multiplication. This process can be formulated as:

$$F_i^t = \mathrm{SA}(\mathrm{CA}(T_i) \odot \mathrm{Sigmoid}(\mathrm{M}(S_i)) + \mathrm{CA}(T_i)), \quad (10)$$

$$F_i^s = \mathrm{SA}(\mathrm{CA}(S_i) \odot \mathrm{Sigmoid}(\mathrm{M}(T_i)) + \mathrm{CA}(S_i)). \quad (11)$$

$\mathrm{SA}(\cdot)$ and $\mathrm{CA}(\cdot)$ represent Spatial Attention and Channel Attention, respectively. $\mathrm{M}(\cdot)$ denotes channel-wise mean function. The features F_i^t and F_i^s are then concatenated along the channel dimension. However, these features are unbalanced as the temporal information still dominates the features. Inspired by the simple gate of the NAFNet (Chen et al. 2022), we evenly split the concatenated features into two halves along the channel dimension. Then, each corresponding pair of channels from the two halves is combined through element-wise multiplication to generate the gated output features F_i^g. This process can be formulated as:

$$F_{concat} = \mathrm{Concat}(F_i^t, F_i^s), \quad (12)$$

$$F_i^g = F_{concat}^{[:C/2]} \odot F_{concat}^{[C/2:C]}, \quad (13)$$

where $\odot$ denotes element-wise multiplication.

The output from the previous decoding layer is upsampled and fused with the current gated output F^g, and then passed through a Convolution block with Batch Normalization, termed $\mathrm{CB}(\cdot)$, to obtain the decoded feature for this layer. This process can be formulated as follows:

$$D_i = \begin{cases} F_4^g & i = 4 \\ \mathrm{CB}(\mathrm{Concat}(F_i^g, \mathrm{UP}(D_{i+1}))) & i = 1, 2, 3. \end{cases} \quad (14)$$

Finally, we apply a 1×1 convolution to the feature D_1 to generate the segmentation mask of the glass region.

Loss Function

We employ BCE loss (De Boer et al. 2005) and IoU loss (Qin et al. 2019) to supervise the predicted results at the pixel level and the region level, respectively. The overall loss consists of the loss for the primary glass surface mask P_{N-1} and the losses for the predicted masks of three adjacent frames.

The loss for the primary mask P_{N-1} is defined as:

$$\mathcal{L}_P = \mathcal{L}_{BCE}(P_{N-1}, \hat{M}_{N-1}) + \mathcal{L}_{IoU}(P_{N-1}, \hat{M}_{N-1}), \quad (15)$$

where $\hat{M}_{N-1}$ denotes the ground truth of the glass surface mask in the frame of N-1.

We also employ the same loss for the prediction across the three adjacent frames. This can be formulated as:

$$\mathcal{L}_M = \sum_{i=N-2}^{N} (\mathcal{L}_{\text{BCE}}(M_i, \hat{M}_i) + \mathcal{L}_{\text{IoU}}(M_i, \hat{M}_i)), \quad (16)$$

where M denotes the prediction of the glass surface mask.

The total loss of the network consists of the above two losses and can be defined as:

$$\mathcal{L} = \alpha \mathcal{L}_P + \mathcal{L}_M, \quad (17)$$

where α is empirically set to $1/8$ for balancing the weights between $\mathcal{L}_P$ and $\mathcal{L}_M$.

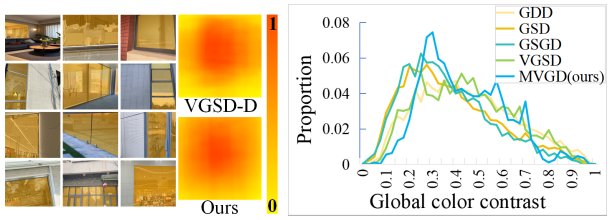
Task	Dataset	Publication	Videos	Frames
VGSD	VGSD-D	AAAI'24	297	19166
	PVG-117	ICCV'23	117	21485
VMD	MMD	CVPR'24	37	9727
	VMD-D	CVPR'23	269	15066
VGSD	MVGD-D(Ours)	-	312	19268

Table 1: Comparison of Datasets for VGSD/VMD task.

Dataset

Though Liu *et al.* (Liu et al. 2024) proposed the dataset VGSD-D for VGSD, this dataset covers only a limited variety of scenes, as illustrated in Fig. 7. Moreover, VGSD-D

Figure 7: Comparison between VGSD-D(1st-2nd columns) and our MVGD-D(3rd-8th columns). For each scene, from top to bottom, we show two selected frames and the mask for the second frame.



(a) Glass Region Distribution (b) Color Contrast Distribution

Figure 8: Statistics of our MVGD-D.

suffers from biased distributions of glass surface locations and overly pronounced color contrast between glass and non-glass regions. To alleviate these limitations, we construct a new large-scale dataset, named MVGD-D, which comprises 312 real-world video clips (19, 268 images) with glass surfaces, manually annotated masks and flow maps produced by RAFT (Teed and Deng 2020). MVGD-D covers diverse static scenes with dynamic camera movements.

Dataset Construction

Our MVGD-D consists of 236 real-world videos captured by handheld devices and 76 videos sampled from streaming platforms (which are under the Creative Commons licenses). The handheld videos are recorded using DSLR cameras and smartphones, while the streaming videos are collected from platforms such as YouTube and Bilibili. Each video has a duration of approximately 3-10 seconds. Videos were captured at 30 FPS but sampled at 15 FPS to balance redundancy and motion observability. Subsequently, experienced volunteers were engaged to carefully annotate video frames. The size of our dataset is comparable to that of existing representative video glass datasets. A comparison of the VGSD datasets (VGSD-D (Liu et al. 2024) and PVG117 (Qiao et al. 2023)) and the VMD datasets (MMD (Warren et al. 2024) and VMD-D (Lin, Tan, and Lau 2023)) is shown in Tab. 1.

Dataset Analysis

- **Glass Location.** As shown in Fig. 8 (a), MVGD-D exhibits diverse glass surface locations, covering nine distinct spatial patterns. We overlap all masks to obtain the binary distribution of all glass/mirror surfaces within VGSD-D and MVGD-D, respectively. The distribution exhibited in Fig. 8 (a) shows that though both datasets exhibit glass primarily in the center region, the more uniform color distribution in MVGD-D reflects a more balanced underlying distribution, which suggests that MVGD-D is more effective in avoiding the center bias problem.
- **Color Contrast.** A lower color contrast between glass and non-glass regions indicates a higher similarity between the glass surface and its surrounding scene, thereby increasing the difficulty for VGSD. We compute the χ^2 distance to evaluate the color contrast across several glass surface datasets(GDD (Mei et al. 2020), GSD (Lin, He, and Lau 2021), GSGD (Yan et al. 2025), VGSD (Liu et al. 2024)). Fig. 8 (b) shows that the contrast values in our dataset are mainly distributed within the range of

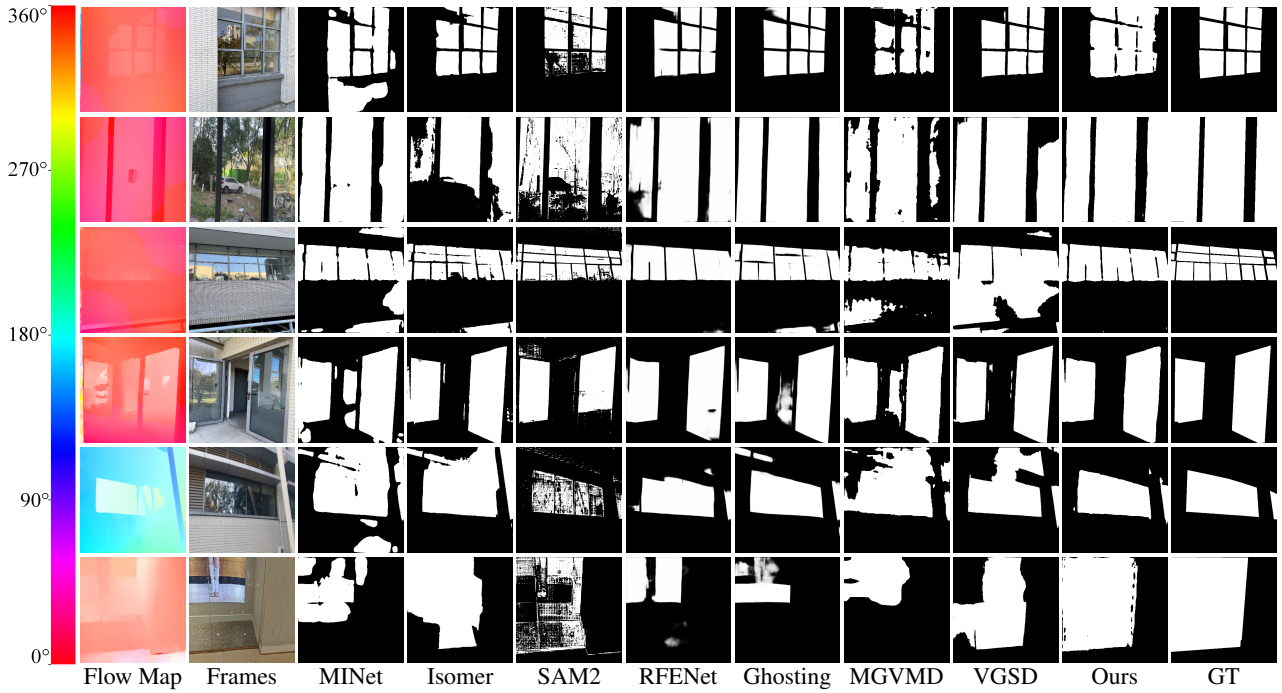


Figure 9: Qualitative comparison of our method and competing methods on MVGD-D. The leftmost subfig shows the optical flow color bar.

$[0.2, 0.6]$, suggesting that MVGD-D is more varied and challenging for VGSD.

Experiments

Implementations

All methods were retrained and evaluated on the two VGSD datasets using PyTorch on an NVIDIA RTX 4090 (24GB). We resized all frames to 384×384 pixels, keeping other settings consistent with their original implementations. For all optical flow-dependent methods, unified flow maps are generated by RAFT (Teed and Deng 2020). We do not apply data augmentation, as it may disrupt temporal consistency. The evaluation metrics we chose include Intersection over Union (IoU↑), F-measure (F_β ↑), Mean Absolute Error (MAE↓), Balanced Error Rate (BER↓) and Accuracy (ACC↑).

Comparisons

We evaluate our method against 11 state-of-the-art approaches, including 1 SOD method (MINet (Pang et al. 2020)), 3 VSOD methods (FSNet (Ji et al. 2021), ISomer (Yuan et al. 2023) and UFO (Su et al. 2023)), 1 Semantic segmentation(SS) method (SAM2 (Ravi et al. 2024)), 3 GSD methods (GSD (Lin, He, and Lau 2021), RFENet (Fan et al. 2023) and GhostingNet (Yan et al. 2025)), one VGSD method (VGSDNet (Liu et al. 2024)), and 2 VMD methods (VMD-Net (Lin, Tan, and Lau 2023) and MGVMD (Warren et al. 2024)). As shown in Table 2, our method achieves superior performance in both datasets. Specifically, compared to the second-best performing method VGSD-Net (Liu et al. 2024) on VGSD-D, our method performs better by IoU: 7.20% ↑, F_β: 4.46% ↑, MAE: 35.35% ↓, BER: 36.45% ↓, and ACC: 4.12% ↑. Furthermore, despite sharing the same backbone as GhostingNet, our method surpasses it, showing notable improvements on MVGD-D, as IoU: 3.26% ↑, F_β: 0.44% ↑, MAE: 13.46% ↓, BER: 7.45% ↓, and ACC: 1.63% ↑. Notably, FSNet obtains a higher ACC but a lower IoU in two datasets, as it relies heavily on optical flow and tends to overdetect glass regions in most scenes. It can also be seen from the table that single-image-based GSD methods perform worse than video-based GSD methods in most cases, mainly because they cannot exploit temporal information for GSD. Similarly, SOD methods perform worse than GSD methods, as glass surfaces typically lack obvious appearances and require effective cues.

The qualitative visual results are shown in Fig. 9. In the $1st$ scene, optical flow map exhibits obvious motions of reflections, but networks such as Isomer, RFENet, MGVMD, and VGSD fail in the top-left region with weak reflections. In the $2nd$ and $6th$ scenes, the reflections on the glass are relatively weak in the indoor environment, the optical flow maps display motion inconsistencies from the transmitted layer, allowing our network to achieve better detection. It is worth noting that SAM2 and Isomer are affected by the semantic information of the transmitted layer, resulting in limited performance. In the $3rd$ scene, the presence of railings at the bottom area creates a glass-like region. Compared with Isomer and MGVMD, our method refines the optical flow map to avoid introducing erroneous information. In these scenes, the proposed method performs better than all competing methods, as our method is able to effectively exploit motion inconsistency for GSD task.

Task	Method	Year	VGSD-D (Liu et al. 2024)					MVGD-D (Ours)					Time(ms)
			IoU↑	F_β↑	MAE↓	BER↓	ACC↑	IoU↑	F_β↑	MAE↓	BER↓	ACC↑	
SOD	MINet(Pang et al. 2020)	CVPR'20	71.84	81.94	0.162	0.157	0.869	71.29	81.29	0.152	0.132	0.885	7.33
VSOD	FSNet(Ji et al. 2021)	ICCV'21	69.88	78.22	0.188	0.179	0.914	66.90	74.88	0.190	0.190	**0.938**	11.62
	ISomer(Yuan et al. 2023)	ICCV'23	77.58	85.18	0.127	0.129	0.905	77.25	85.81	0.126	0.114	0.892	16.14
	UFO(Su et al. 2023)	TMM'23	64.69	74.49	0.235	0.228	0.873	67.05	77.16	0.211	0.192	0.885	22.28
SS	SAM2(Ravi et al. 2024)	arXiv'24	78.60	89.28	0.131	0.125	0.844	78.18	88.25	0.121	0.113	0.841	84.88
GSD	GSDNet(Lin, He, and Lau 2021)	CVPR'21	78.19	86.46	0.116	0.111	0.902	76.34	85.50	0.125	0.111	0.894	58.40
	RFENet(Fan et al. 2023)	IJCAI'23	79.21	88.60	0.109	0.105	0.910	76.45	87.29	0.124	0.109	0.883	39.14
	GhostingNet(Yan et al. 2025)	TPAMI'24	80.40	88.83	0.100	0.093	0.905	80.01	88.75	0.104	0.094	0.915	32.94
VGSD	VGSDNet(Liu et al. 2024)	AAAI'24	80.72	88.57	0.099	0.096	0.898	77.27	85.59	0.126	0.110	0.904	72.44
VMD	VMD-Net(Lin, Tan, and Lau 2023)	CVPR'23	74.12	84.35	0.136	0.133	0.879	70.74	81.92	0.155	0.139	0.865	54.40
	MG-VMD(Warren et al. 2024)	CVPR'24	76.56	84.68	0.125	0.123	0.912	73.69	83.74	0.134	0.122	0.887	190.04
VGSD	Ours	-	**86.57**	**92.53**	**0.064**	**0.061**	**0.935**	**82.62**	**89.14**	**0.090**	**0.087**	0.930	190.9

Table 2: Quantitative comparison on VGSD-D and MVGD-D. **Red** and Cyan denote the best and second-best values, respectively.

Module	IoU↑	F_β ↑	MAE↓	BER↓	ACC↑
A. BS + BD	74.31	80.87	0.140	0.135	0.905
B. BS + RAFT + BF + BD	75.59	82.12	0.136	0.131	0.908
C. BS + CMFM + BT + TSD	79.80	86.33	0.109	0.107	0.915
D. BS + BF + TAM + TSD	78.74	85.24	0.117	0.112	0.915
E. BS + CMFM + TAM + BD	80.08	86.58	0.104	0.101	0.922
F. w/o P	80.36	86.93	0.107	0.098	0.922
G. BS + CMFM + TAM + TSD	**82.62**	**89.14**	**0.090**	**0.087**	**0.930**

Table 3: Ablation study of our network. 'BS' denotes the Swin Transformer backbone. 'BD' represents a simple decoder used to ablate TSD. Temporal Attention Module (TAM) is the collective term for HGAM and TCAM 'BT' represents a basic temporal module replacing TAM, and 'BF' represents a basic fusion module replacing CMFM. 'P' denotes the initial glass mask prediction used to refine optical flow map.

Figure 10: The visual comparison of different ablated models.

Ablation Study

We report ablation results on MVGD-D in Tab. 3. The improvement of *B* over *A* on IoU demonstrates the effectiveness of motion cues in guiding GSD. The comparison between *D* and *G* indicates that CMFM brings a significant improvement to the network. Moreover, it can be seen from *F* that the mask P_{N-1} plays a crucial role in improving the performance of our network. Visual comparisons are shown in Fig. 10. The comparison between B and A further confirms that motion cues can guide GSD, though further optimization is still needed. The comparisons of Models C–E with Model G show that each proposed module contributes positively to the overall performance.

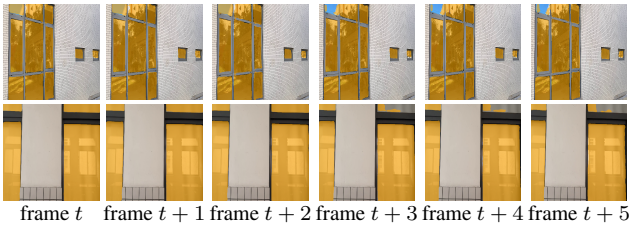


Figure 11: Typical failure cases.

Conclusion

In this paper, we have proposed a novel network, named MVGD-Net, for VGSD based on motion inconsistency cues. We have also proposed three effective modules to exploit spatial and temporal features from videos to improve detection performance. Additionally, we have constructed a large-scale dataset for VGSD. Extensive experiments demonstrate that our method outperforms existing related SOTA methods.

Our method does have limitations, as it takes only three consecutive frames as input, which limits its ability to capture long-term temporal dependencies. As shown in Fig. 11, in the $1st$ scene, the top-left glass region is missed in the $4th$ frame despite being consistently detected in previous frames. In the $2nd$ scene, the top-right glass region is under-detected in the $4th$ frame. These failures highlight our model's limitation in maintaining consistent detection under appearance changes. Moreover, similar to existing image-based GSD methods, our method may mis-detect some open door/window, specifically glass-like regions enclosed by their frames, as glass regions.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61902151).

References

Chen, L.; Chu, X.; Zhang, X.; and Sun, J. 2022. Simple baselines for image restoration. In *Proc. ECCV*.

Cheng, H. K.; Oh, S. W.; Price, B.; Lee, J.-Y.; and Schwing, A. 2024. Putting the object back into video object segmentation. In *Proc.CVPR*.

Cheng, H. K.; Tai, Y.-W.; and Tang, C.-K. 2021. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *Proc. NeurIPS*.

De Boer, P.-T.; Kroese, D. P.; Mannor, S.; and Rubinstein, R. Y. 2005. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134: 19–67.

Fan, K.; Wang, C.; Wang, Y.; Wang, C.; Yi, R.; and Ma, L. 2023. RFENet: Towards Reciprocal Feature Evolution for Glass Segmentation. In *Proc. IJCAI*, 717–725.

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 *Proc. ICCV*.

Huo, D.; Wang, J.; Qian, Y.; and Yang, Y.-H. 2023. Glass Segmentation With RGB-Thermal Image Pairs. *IEEE TIP*, 32: 1911–1926.

Ji, G.-P.; Fu, K.; Wu, Z.; Fan, D.-P.; Shen, J.; and Shao, L. 2021. Full-duplex strategy for video object segmentation. In *Proc. ICCV*.

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

Li, Z.; Li, K.; Wang, S.; Lan, S.; Yu, Z.; Ji, Y.; Li, Z.; Zhu, Z.; Kautz, J.; Wu, Z.; et al. 2024. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. In *Proc. CVPR*.

Li, Z.; Yeh, Y.-Y.; and Chandraker, M. 2020. Through the looking glass: Neural 3d reconstruction of transparent shapes. In *Proc. CVPR*.

Liang, Y.; Deng, B.; Liu, W.; Qin, J.; and He, S. 2023. Monocular Depth Estimation for Glass Walls With Context: A New Dataset and Method. *IEEE TPAMI*, 45(12): 15081–15097.

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

Lin, J.; Tan, X.; and Lau, R. W. 2023. Learning to detect mirrors from videos via dual correspondences. In *Proc. CVPR*.

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

Lin, J.; Yeung, Y.-H.; Ye, S.; and Lau, R. W. 2025. Leveraging RGB-D Data with Cross-Modal Context Mining for Glass Surface Detection. In *Proc. AAAI*.

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

Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proc. CVPR*.

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 *Proc. CVPR*.

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 *Proc. CVPR*.

Oh, S. W.; Lee, J.-Y.; Xu, N.; and Kim, S. J. 2019. Video object segmentation using space-time memory networks. In *Proc. ICCV*.

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

Qi, F.; Tan, X.; Zhang, Z.; Chen, M.; Xie, Y.; and Ma, L. 2024. Glass Makes Blurs: Learning the Visual Blurriness for Glass Surface Detection. *IEEE TII*, 20(4): 6631–6641.

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 *Proc. ICCV*.

Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; and Jagersand, M. 2019. Basnet: Boundary-aware salient object detection. In *Proc. CVPR*.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.

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*, 26: 313–325.

Tamura, H.; Higashi, H.; and Nakauchi, S. 2017. Multiple cues for visual perception of mirror and glass materials. *Journal of Vision*, 17(10): 765–765.

Tamura, H.; Higashi, H.; and Nakauchi, S. 2018. Dynamic visual cues for differentiating mirror and glass. *Scientific reports*, 8(1): 8403.

Tamura, H.; Tsukuda, M.; Higashi, H.; and Nakauchi, S. 2016. Perceptual segregation between mirror and glass material under natural and unnatural illumination. *Journal of Vision*, 16(12): 634–634.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Proc. ECCV*.

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

Warren, A.; Xu, K.; Lin, J.; Tam, G. K.; and Lau, R. W. 2024. Effective video mirror detection with inconsistent motion cues. In *Proc. CVPR*.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. CBAM: Convolutional Block Attention Module. In *Proc. ECCV*.

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*, 47(1): 323–337.

Yan, T.; Xu, S.; Huang, H.; Li, H.; Tan, L.; Chang, X.; and Lau, R. W. 2024. NRGlassNet: Glass surface detection from visible and near-infrared image pairs. *KBS*, 294: 111722.

Yuan, Y.; Wang, Y.; Wang, L.; Zhao, X.; Lu, H.; Wang, Y.; Su, W.; and Zhang, L. 2023. Isomer: Isomerous transformer for zero-shot video object segmentation. In *Proc. ICCV*.