Multi-view Spectral Polarization Propagation for Video Glass Segmentation

Yu Qiao, Bo Dong*, Ao Jin, Yu Fu, Seung-Hwan Baek, Felix Heide
Pieter Peers, Xiaopeng Wei*, Xin Yang*

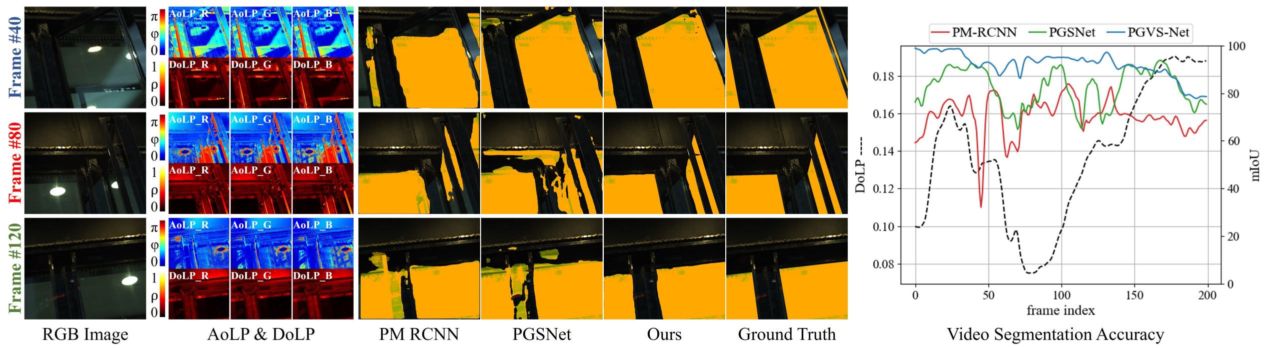


Figure 1. A comparison of PGVS-Net on 3 selected frames from an RGB-P video sequence from our novel PGV-117 dataset against the method of Kalra *et al.* [14] and Mei *et al.* [29]. From left to right: RGB, AoLP, and DoLP input images; glass segmentation results for selected frames compared to the ground truth segmentation; evolution of error over the video sequence. Both prior methods are temporally unstable due to significant temporal changes in DoLP. In contrast, PGVS-Net combines multi-view polarization cues for segmentation propagation, yielding a temporally stable glass segmentation.

Abstract

In this paper, we present the first polarization-guided video glass segmentation propagation solution (PGVS-Net) that can robustly and coherently propagate glass segmentation in RGB-P video sequences. By leveraging spatio-temporal polarization and color information, our method combines multi-view polarization cues and thus can alleviate the view dependence of single-input intensity variations on glass objects. We demonstrate that our model can outperform glass segmentation on RGB-only video sequences as well as produce more robust segmentation than per-frame RGB-P single-image segmentation methods. To train and validate PGVS-Net, we introduce a novel RGB-P Glass Video dataset (PGV-117) containing 117 *video sequences of scenes captured with different types of camera paths, lighting conditions, dynamics, and glass types.*

1. Introduction

Transparent objects, such as glass, are common in man-made environments. Despite their prevalence, such objects pose numerous challenges for industrial and academic vision processing algorithms due to their view-dependent appearance, statistical similarity to the background environment, and lack of texture. These challenges are particularly acute for segmentation tasks that underpin autonomous driving, robotics, and aerial drone navigation. Traditional learning-based solutions [8, 30, 50, 51] that rely on RGB textures for feature extraction often fail to adequately model the illumination and view-dependent glass features. Consequently, recent advances have explored richer modalities such as light fields [44] and polarization [14, 18, 21, 29, 46, 49] for more robust segmentation of glass objects. Polarization has proven to be a strong cue for glass segmentation in still images [14, 29]. The availability of commercial off-the-shelf RGB-P cameras (*e.g.*, Flir and Lucid), makes this last category an attractive modality for glass segmentation.

Polarization of reflected light depends on various factors such as surface normals, material types, and illumination and view directions [2]. Due to the polarimetric near-specular behavior, glass materials are particularly challenging; small changes in object orientation and/or view/light direction can induce a rapid change in polarization state. For example, consider the glass areas '1' in Figure 2 that show a glass door in various stages of being closed. In this example, the RGB textures remain almost unchanged, but

*Corresponding authors.

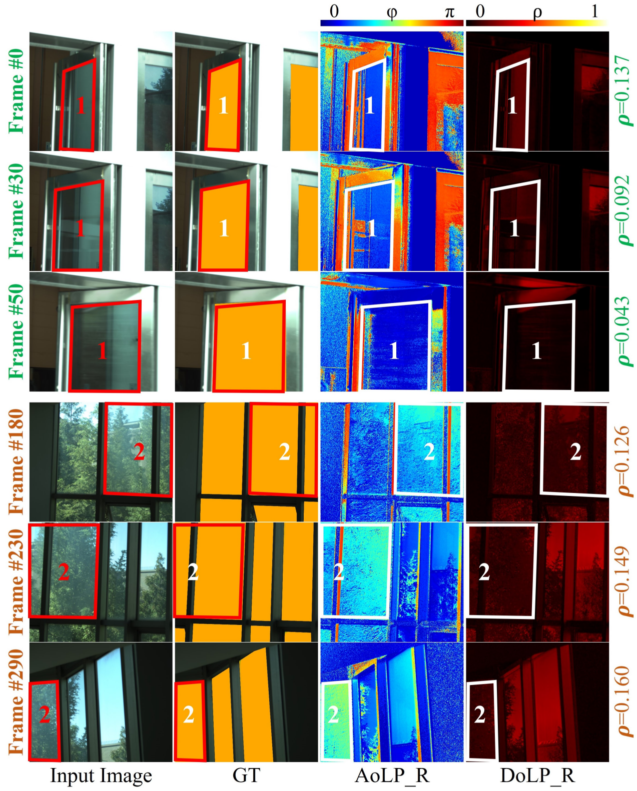


Figure 2. Representative exemplars from the PGV-117 dataset containing a total of 117 video sequences captured with a variety of camera movement patterns, lighting conditions, scene dynamics, and materials. Glass areas ‘1’ depict a door in various stages of being closed. Glass areas ‘2’ are captured by moving the camera.

changes in the surface normal induce a significant decrease in the Angle of Linear Polarization (AoLP) from close to π (red-blue mixture) to less than half π (full blue), and a noticeable reduction in the Degree of Linear Polarization (DoLP) from 0.137 to 0.043. Similarly, moving the camera will induce a similar change in the polarization (*e.g.*, the glass areas ‘2’ in Figure 2). These observations suggest that polarization cues are beneficial for glass segmentation and can provide additional guidance. Furthermore, the DOLP in glass area ’1’ is unstable across the entire span of the video sequence, with noticeable DoLP reductions at specific angles of incidence ($\rho = 0.043$). Such types of instabilities can confuse methods [14, 29] that only rely on a single-frame RGB-P as input (Figure 1, right). Moreover, extracting glass features from RGB textures or polarization cues from a single view may also be difficult (*e.g.*, under low light conditions as in Figure 1 left). As a consequence, polarization cues that are robust for static single image segmentation methods, can still yield temporally disjoint and inconsistent segmentation results when applied to video sequences, especially when object orientation and/or viewpoint changes.

In this paper, we introduce a Polarization Video Glass Segmentation network (PGVS-Net) to robustly and coherently propagate an initial glass segmentation mask over an RGB-P video sequence. PGVS-Net perceives multi-view polarimetric features from long-range memory and short-range integration. A variation of the Spatio-Temporal Memory network (STM) [33] is introduced to leverage spectral polarization cues for constructing long-range view dependency. A well-designed Polarization-Guide Integration module (PGI) is responsible for the frame identity and historical polarization matching. Furthermore, a novel Polarization Temporal Forward module (PTF) integrates previous polarization cues to promote short-range spectral consistency during the decoder process. We also leverage cross-modal and cross-temporal attention (CMTA) to integrate multimodal features and temporal representations from different frames. To train and test PGVS-Net, we capture and annotate a large-scale video glass segmentation dataset, PGV-117, which includes 21,485 annotated RGB-P frames and masks. We demonstrate, using this newly created dataset, that our PGVS-Net outperforms competing single-frame RGB-P glass segmentation methods as well as RGB-based video segmentation methods both quantitatively and qualitatively.

In summary, our contributions are:

- A novel Polarization Video Glass Segmentation network (PGVS-Net) to robustly provide glass segmentation propagation in RGB-P video sequences;
- A series of polarimetric consistency and integration modules to enhance spatio-temporal correlation of spectral polarization cues;
- A large-scale video glass segmentation dataset, PGV-117, containing 21,485 frames with spectral polarization cues and densely-annotated masks.

2. Background and Related Work

Polarization. Light is an electro-magnetic wave, whose polarization state describes the orientation of the transverse electric field, which can be unpolarized (*i.e.*, random), linearly polarized (*i.e.* biased towards a single direction), or circularly polarized. We focus our discussion on linear polarization as supported by recent commercial polarization-array CMOS sensors, which allows us to record four distinct linear-polarization states per pixel: I_{0°, I_{45°, I_{90°, and I_{135°, where I_x describes the intensity of linearly-polarized light at an angle x. These four linear polarization measurements allow us to compute the linear-polarization components of the Stokes vector $S = [S_0, S_1, S_2]$:

$$\begin{aligned} S_0 &= I_{0^\circ} + I_{90^\circ} = I_{45^\circ} + I_{135^\circ}, \\ S_1 &= I_{0^\circ} - I_{90^\circ}, \\ S_2 &= I_{45^\circ} - I_{135^\circ}, \end{aligned} \tag{1}$$

where S_0 is the total light intensity, S_1 and S_2 are the ratio of the $0°$ and $45°$, respectively, linear polarization intensity over its perpendicular counterpart. Given the linear Stokes vector, the degree of linear polarization (DoLP) and the angle of linear polarization (AoLP) are defined as:

$$\mathrm{DoLP} = \frac{\sqrt{S_1^2 + S_2^2}}{S_0}, \quad \mathrm{AoLP} = \frac{1}{2}\arctan\left(\frac{S_2}{S_1}\right). \tag{2}$$

Glass Object Segmentation. RGB texture information has proven to be an effective cue for image detection [41, 42, 43] and segmentation [24, 36, 37]. However, accurately segmenting glass objects in images remains a challenging problem because of the dynamic texture patterns that vary with viewing conditions due to reflection and transmission that share the statistics of the surrounding scene. Prior glass segmentation methods operating exclusively on RGB images have leveraged contextual information [30, 52, 53] or boundary cues [9, 50] to address these challenges. However, the effectiveness of RGB-only appearance cues depends on the lighting conditions, the amount of scene clutter, and the presence of nefarious actors (e.g., print-out-spoofs) [14]. One avenue to improve glass segmentation robustness is to include additional modalities that further differentiate glass objects from others, such as scene depth [28], thermal information [12], and polarization [14, 29]. Our approach falls in this latter category, and we rely on RGB and polarization (RGB-P) to robustly segment glass objects. However, the strength of the polarization cues can vary from frame to frame in a video sequence due to changes in view or lighting conditions yielding inconsistent object segmentation between frames. In this work, we leverage multi-view observations to recover a more complete, less view-dependent, description of the polarization properties of objects in the scene, resulting in a temporally more robust glass segmentation.

Video Segmentation and Detection. Video sequences can offer additional temporal cues to aid segmentation and detection tasks, but simultaneously also introduce the additional challenge of temporal consistency. Semantic video segmentation methods enhance temporal consistency via a variety of temporal propagating approaches such as optical flow[13, 32, 55], spatially varying convolutions [23], and recurrent network architectures [40]. In the absence of semantic cues, video object segmentation approaches estimate consistent class-agnostic object masks and instance IDs by exploiting spatial and temporal information to propagate an initial mask [4, 7, 11, 20, 26, 33, 39]. Panoptic segmentation [16] is a popular and successful segmentation approach that unifies semantic and instance segmentation, and which has also been extended to video segmentation [15, 22, 35, 47]. In addition to segmentation, the related problem of object detection has recently also been extended to video sequences in different application domains such as salient object detection [17, 19], camouflaged object detection [5], shadow detection [3, 25], and video action detection [54]. Similar to prior work, we also exploit spatial and temporal cues. In addition, we also leverage the richer embedded cues in RGB-P videos and gather robust polarization cues from multi-view observations to aid in propagating glass segmentation masks in video sequences. To the best of our knowledge, none of the prior segmentation propagation methods can effectively leverage polarization cues that are uncorrelated with RGB features.

3. Polarimetric Video Glass Segmentation

Our Polarization Video Glass Segmentation network (PGVS-Net) sequentially processes video frames beginning with the second frame, utilizing the ground truth annotation provided in the initial frame. As the video processing unfolds, preceding frames with object masks—established in the first frame or inferred in subsequent frames—are employed as memory frames. Furthermore, the current frame, excluding the object mask, assumes the role of the query frame. Note that each frame consists of trichromatic intensity I, AoLP φ, and DoLP ρ information.

Inspired by the successful Spatial-Temporal Memory (STM) framework [4, 33], we encode each frame into a paired set of key and value maps. Specifically, for a given modality x, we utilize a ResNet50 architecture [10] to encode the modality into three distinct feature maps: f_{b1}^x, f_{b2}^x, and f_{b3}^x, corresponding to the outputs of the first, second, and third blocks, respectively. The key feature map of each frame is derived using our specially devised Polarization-Guided Integration module (PGI; detailed in Section 3.1), based on f_{b1}^x, f_{b2}^x, and f_{b3}^x. Importantly, both memory and query frames employ the same methodology to generate key maps. For memory frames, the value map is generated through the following steps:

1. Trichromatic I, φ, ρ, and the glass mask are concatenated along the channel dimension and fed as input to a ResNet18 model [10] to produce a feature map;
2. The resulting feature map is concatenated with f_{b3}^x and subjected to two residual blocks [10] and a Convolutional Block Attention Module (CBAM) [48] attention block;
3. Finally, the memory value maps are produced by aggregating the multimodality maps from step 2.

In essence, relying on the query key and the pairs of mem-

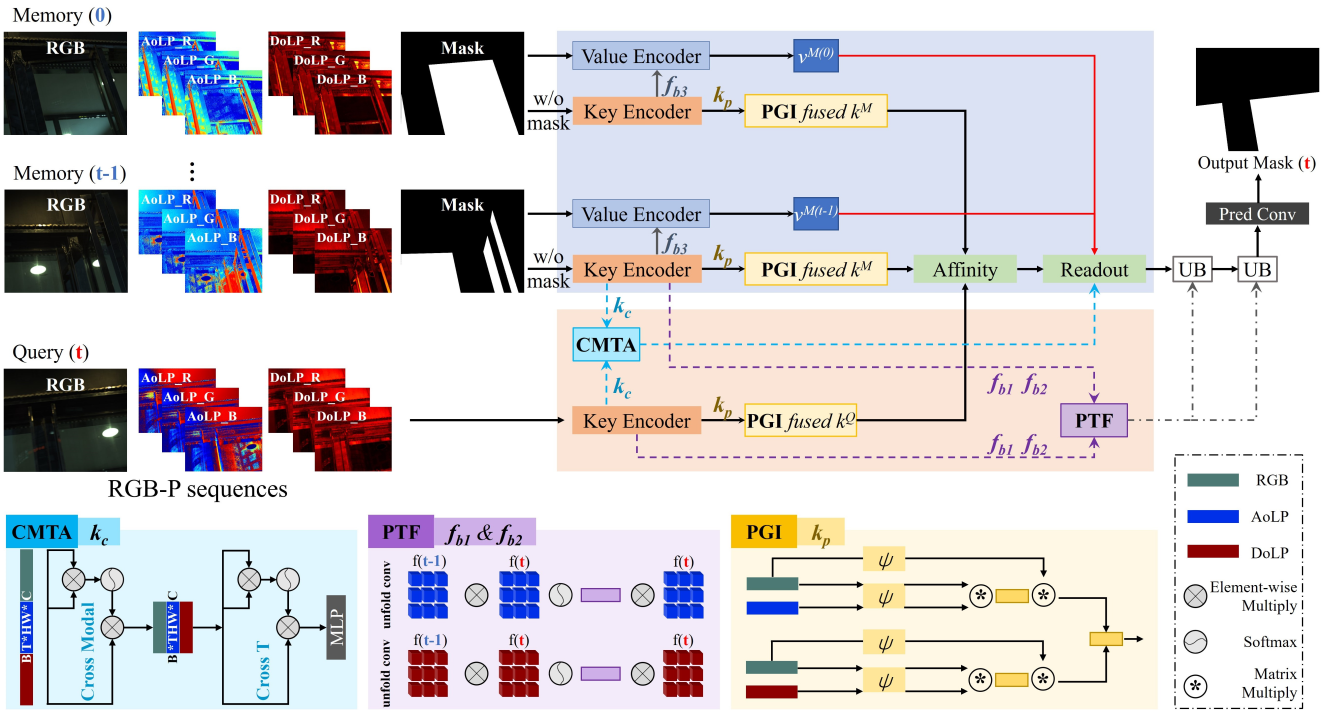


Figure 3. Overview of PGVS-Net. The RGB-P Memory and PGI fused keys are designed to store long-range multi-view polarization cues. The short-range spectral propagation are implemented by the CMTA and PTF modules.

ory key and value maps, PGVS-Net endeavors to estimate the value map (discussed in Section 3.2) for a given query frame through multi-view polarimetric propagation (described in Section 3.3) and the extraction of temporal correlations (explained in Section 3.4). The resultant value map of the query frame is then translated into the glass mask of the query frame via a decoder (introduced in Section 3.5). Our PGVS-Net is summarized in Figure 3.

3.1. Polarization-Guided Integration

Polarization information provides strong cues on the material properties. We therefore leverage these polarization cues as guidance to combine multimodal key feature maps for each frame. To achieve this, we introduce a novel Polarization-Guided Integration (PGI) module. Formally, the PGI is defined as:

$$\begin{aligned} k_x &= \Psi(f^x_{b3}), \\ \mathcal{Y}_1 &= \gamma_1 * \{[\Psi(k_I) \circledast \Psi(k_\varphi)] \circledast \Psi(k_I)\}, \\ \mathcal{Y}_2 &= \gamma_2 * \{[\Psi(k_I) \circledast \Psi(k_\rho)] \circledast \Psi(k_I)\}, \\ k &= \Psi([k_I, \mathcal{Y}_1, \mathcal{Y}_2]) \end{aligned} \tag{3}$$

where x is either I, φ, or ρ; γ_1 and γ_2 are learnable parameters; $\Psi(\cdot)$ represents a convolution layer, followed by a regional pooling operation (output size equals 11) and an additional convolution layer; $\circledast$ indicates a matrix multiplication. $k_x \in \mathbb{R}^{BT\times H\times W\times C}$, where B, T, H, W, C are the batch size, temporal length, and the height, width, and channel size of a feature map, respectively.

3.2. Query Value Prediction

In the context of a query frame, we establish a connection between RGB-P information from multiple views by investigating the interplay between the query key and memory keys. This interaction results in corresponding value features grounded in memory affinity. Our model employs an STM variant for RGB-P memory, dynamically amalgamating multimodal inputs and establishing correlations across diverse views over an extensive temporal span.

Concretely, when provided with a query key k^Q, we anticipate the query value v^Q through the utilization of a softmax-normalized affinity matrix, which is computed from the pairing of (k^M, k^Q) [4]. This computation allows us to extract the most pertinent memory values. The computation of memory affinity adheres to the methodology outlined by Cheng *et al.* [4] (for a comprehensive explanation, please refer to the supplementary material). Memory values encapsulate historical RGB-P information of a scene, ensuring that the affinity mechanism facilitates the far-reaching propagation of multi-view spectral polarization within v^Q.

3.3. Cross Modal-Temporal Attention

To this point, the query value prediction process only utilizes long term memory by leveraging the collective information contained in the memory frames. However, captured trichromatic information also exhibits rapid changes across different modalities, indicating potential target area boundaries over short time-intervals. These changes across diverse modalities are not well captured by a long-term memory framework. To address this shortcoming, we introduce a Cross Modal-Temporal Attention module (CMTA) to model the short-term interactions between the query and its one preceding memory frame.

Central to our CMTA module is a multi-head self-attention mechanism inspired by the transformer architecture [45]. This mechanism is first applied along the modality dimension and then extended to the temporal dimension. Formally, the CMTA can be defined as follows:

$$\begin{aligned}\chi_m &= [\mathcal{R}_m(k_I), \mathcal{R}_m(k_\varphi), \mathcal{R}_m(k_\rho)],\\ F_m &= \mathcal{L}[\Upsilon(\mathcal{L}(\chi_m)) + \chi_m],\\ \chi_t &= [\mathcal{R}_t(\mathcal{C}_I(F_m)), \mathcal{R}_t(\mathcal{C}_\varphi(F_m)), \mathcal{R}_t(\mathcal{C}_\rho(F_m))],\\ F_t &= \mathcal{L}[\Upsilon(\mathcal{L}(\chi_t)) + \chi_t],\end{aligned} \tag{4}$$

where $[\cdot]$ indicates the concatenation on the first dimension; Υ indicates a multi-head self-attention block [45]; $\mathcal{L}$ refers to Layer Normalization [1]. $\mathcal{R}_m$ is a modality reshape operator, which transfers the shape of k_c from $\mathbb{R}^{B\times T\times H\times W\times C}$ into $\mathbb{R}^{BT\times HW\times C}$. $\mathcal{R}_t$ is a reshape operator in the temporal domain, which reshape a feature map from $\mathbb{R}^{B\times T\times H\times W\times C} \rightarrow \mathbb{R}^{B\times THW\times C}$. $\mathcal{C}_x$ is the chunked vector for modality x. Finally, the derived F_t is concatenated with the query value acquired through the query value prediction process, and subsequently forwarded as input to the decoder.

3.4. Polarimetric Temporal Forward

The examples in Figure 2 demonstrate that the polarization cues provide strong temporal continuity cues. The core idea of leveraging polarimetric continuity is that the polarization for the query frame can benefit from the preceding memory frame's polarization cues. To actualize this, the Polarimetric Temporal Forward (PTF) module engages f^x_{b1} and f^x_{b2}, sourced from the query and the preceding memory frame, as inputs. This module carries out an unfolding operation ($\mathcal{U}$) on these inputs. In order to distinguish between the features from the query and memory frames, an additional subscript is appended to the feature samples. For instance, f^x_{M-b1} and f^x_{Q-b2} represent the feature map of block 1 from a memory frame and the feature map of block 2 from a query frame for the modality x, respectively. Formally, given f^x_{M-b1} and f^x_{Q-b1}, the PTF operation is defined as:

$$\begin{aligned}U^M_\varphi, U^M_\rho &= \mathcal{U}(f^\varphi_{M-b1}), \mathcal{U}(f^\rho_{M-b1}),\\ F^Q_\varphi, F^Q_\rho &= \mathcal{R}(f^\varphi_{Q-b1}), \mathcal{R}(f^\rho_{Q-b1}),\\ S_\varphi &= \sigma(\mathcal{S}(F^Q_\varphi \otimes U^M_\varphi)) \otimes U^M_\varphi + f^\varphi_{b1},\\ S_\rho &= \sigma(\mathcal{S}(F^Q_\rho \otimes U^M_\rho)) \otimes U^M_\rho + f^\rho_{b1},\\ F^{PTF}_{b1} &= S_\varphi + S_\rho + f^I_{b1},\end{aligned} \tag{5}$$

where $\mathcal{R}$ and $\mathcal{S}$ denote the reshape operation and sum operation applied to the first dimension, respectively; $\otimes$ indicates the element-wise multiplication; σ is the softmax function. $U_\varphi, U_\rho \in \mathbb{R}^{C\times P\times H\times W}$, where P denotes the window size during the unfold convolution, and $f^I_{b2}, f^\varphi_{b2}, f^\rho_{b2} \in \mathbb{R}^{C\times 1\times H\times W}$.

PTF further applies the described operations to f^φ_{b2}, f^ρ_{b2}, and f^I_{b2}, yielding the output F^{PTF}_{b2}. In our implementation, we establish the window sizes for unfold convolutions as 7 and 9 for f_{b2} and f_{b1}, respectively. Through this design, PTF enables dynamic incorporation via unfold convolutions, fostering forward propagation among adjacent polarimetric features and thereby achieving inter-frame view dependency.

3.5. Decoder

We employ a straightforward yet effective decoder to translate the acquired query value into the corresponding glass masks of the query frame. Our decoder encompasses two UNet-like [38] upsampling blocks, tasked with reinstating the resolution of the query value map to $1/4$ of the query frame's dimensions. Subsequently, a prediction convolution transforms the channel configuration of the resultant feature map to a single channel, which is then upsampled to match the resolution of the query frame. In addition to the query value, this decoder taps into the feature maps offered by the PTF module to reinforce temporal coherence. Specifically, $F^{PTF}b1$ and $F^{PTF}b2$ serve as inputs to the first and second upsampling blocks, respectively.

3.6. Implementation Details

We implement PGVS-Net in PyTorch. For training, input images are randomly cropped to 384×384 and augmented by horizontal random flipping, translations, and affine transforms. During training, the sequence takes a group of three frames as the input of the network. The masks of ground truth glass areas in the first frame are provided as input, and we employ a binary cross-entropy (BCE) loss to supervise the output mask of the last two frames. For optimization, we use the Adam optimizer with a weight decay of $1e-7$. The learning rate is fixed at $1e-5$. Training takes three days to converge on a dual TITAN V100 graphics card setup after 150,000 iterations with a batch size of 4. During inference, we update the memory bank every three frames.

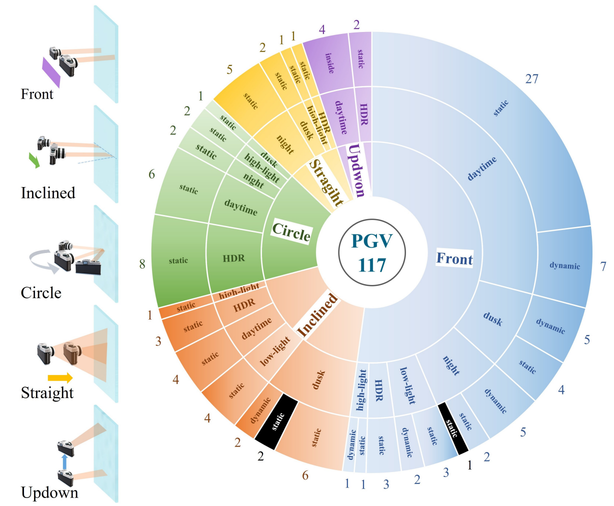


Figure 4. Summary of camera motions (left) and the PGV-117 dataset characteristics (right). To avoid biasing PGV-117 into certain types of camera motions, we implement different camera motion strategies. The blue areas represent the glass target. The 117 video sequences are captured with different types of camera paths, lighting conditions, dynamics, and materials (dark indicates ground glass).

Datasets	Task	Frames	AD	P	Seq(tr+v/te)
DAVIS2017 [34]	VOS	10,459	#1		90/60
DAVSOD [6]	VSOD	23,938	#1		36/90
MoCA-Mask [5]	VCOD	22,939	#5		71/16
RGBP-Glass [29]	IGS	4,511	#1	✓	-
PGV-117	VGS	21,485	#1	✓	85/32

Table 1. **Statistics of PGV-117 compared to representative video datasets and glass datasets.** In each column, we list: the primary *Task* for which the dataset was designed (Video Object Segmentation (VOS), Video Saliency Object Detection (VSOD), Video Camouflaged Object Detection (VCOD), and Image Glass Segmentation (IGS)); *AD*: the annotation interval in #frames; *P*: whether the dataset includes polarization; and *Seq(tr+v)/te*: the number of training and validation exemplars versus test exemplars.

4. PGV-117 dataset

We introduce a novel RGB-P-video-based dataset, PGV-117, to support and stimulate research in video-based glass segmentation. The dataset is collected using a trichromatic (RGB) polarizer-array camera (LUCID PHX050S), equipped with four directional linear-polarizers, i.e., 0°, 45°, 90° and 135°. The frame rate is fixed at 30fps, and we vary the exposure time from 3,000 μs to 39,999 μs depending on the scene and lighting conditions. Each frame in the PGV-117 dataset includes: an RGB image, four directional polarization images, inferred spectral AoLP and DoLP maps, and the captured RAW camera data. Table 1 summarizes PGV-117 statistics in comparison to prior datasets. Care was taken during creation of PGV-117 to ensure a sufficient variation in lighting conditions, camera paths, and scene dynamics.

Lighting conditions. PGV-117 is collected in various indoor environments, such as shopping malls, office buildings, and classrooms, as well as in a variety of outdoor environments. For outdoor environments, we capture sequences spanning during the full time-range from daytime to night. As such, PGV-117 offers four types of lighting variations: normal, low-light, saturated light, and high-dynamic-range. Additionally, we also consider diverse glass types in PGV-117, including clear, frosted, and patterned glasses, tinted glass on vehicles, and outdoor and indoor glass walls.

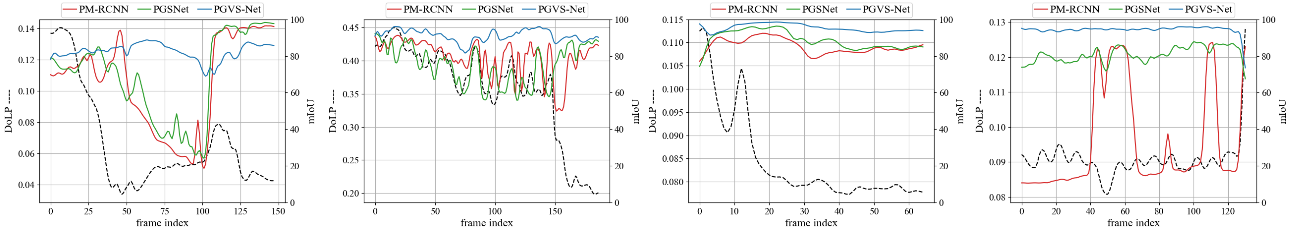
Camera motion patterns. Polarization cues are angle dependent. To avoid biasing PGV-117 to certain view angles and camera paths, we randomly pick one of five predetermined camera motion patterns: front, inclined, circular, straight-forward, and up-down; see Figure 4 for a visual summary. During acquisition, we also recapture some glass areas back and forth to enrich the robustness and diversity.

Dynamics. PGV-117 captures both static and dynamic scenes (*e.g.*, doors/windows opening and closing, and pedestrians occluding glass).

Frame removal. Frames characterized by motion blur and overexposure are excluded due to the inherent ambiguity in their annotations. Typically, we remove 1 to 2 frames per static-camera sequence, and about 20% frames for a moving-camera sequence.

5. Experiments

In this section, we first compare the performance of PGVS-Net under continuous polarization to demonstrate the robustness of combining multi-view and temporal consistency (subsection 5.1). Next, we validate the efficacy of our model by quantitatively and qualitatively comparing PGVS-Net against polarization-based and SOTA video methods retrained on the PGV-117 dataset. To quantify performance, we adopt four prevalent evaluation metrics: intersection over union (IoU), weighted F-measure (F_β) [27], mean absolute error (MAE), and balance error rate (BER) [31]. Finally, we perform an ablation study to show the impact of the different components that comprise PGVS-Net (subsection 5.3). Please refer to the supplementary materials for additional images and video results.



dynamic_door_daytime5 outside_building_daytime_front24 dynamic_kettle_night2 HDR_building_straight1

Figure 5. Error plots comparing PGVS-Net to PM-RCNN and PGSNet for four different sequences. For reference, the DoLP is also included to indicate which frames pose a greater challenge for RGB-P methods.

5.1. Multi-view Polarimetric Consistency

Due to their near-specular reflection and refraction behavior, glass materials are more sensitive to changes in view/light conditions. While RGB-P cues have proven to be essential for robust single-image glass segmentation, ignoring the temporal instability can result in inconsistent glass segmentation over time. PGVS-Net is specially designed to leverage and aggregate temporal changes in polarization cues. Figure 5 illustrates the correlation between variations in polarization intensity (DoLP) and the corresponding performance levels (measured by mIoU) in comparison to two image-based polarimetric techniques [14, 29] (applied independently to each frame) across diverse video sequences. The distinct abrupt shifts in the curve within Figure 5 due to the removal frames with motion blur (section 4), which breaks the temporal continuity and may degrade the multi-view dependency of PGVS-Net.

From the first result in Figure 5, we can observe that when DoLP changes significantly, prior single-frame image-based methods fail to produce a consistent, coherent segmentation. In contrast, our model not only stores and leverages historical polarization information, but also enhances the current frame with the prior polarization information through the CMTA and PTF modules, resulting in good temporal consistency. The second result in Figure 5 shows that our model can still maintain a stable performance when the overall DoLP changes within the sequence cover a relatively large range (between 0.2 and 0.45). The third and fourth result in Figure 5 show that PGVS-Net can achieve better performance in low light and high brightness environments with extremely low overall DoLP over the sequence (less than 0.12).

5.2. Comparison to the State-of-the-art

We compare PGVS-Net to prior single-frame image-based glass segmentation methods and state-of-the-art video segmentation methods. We retrain all methods on the PGV-117 dataset and evaluate them on the 32 test sequences. We compare against: RGB-input GSD [30] where the input and supervision are replaced with our dataset; RGB-P glass segmentation methods EAF [49], PM-RCNN [14] and PGS [29]; STICT [25] (since this is an unsupervised method, we use the polarimetric dataset from [29] as labeled images for training); ViSha [3] is a recent video shadow detection method; STM-based video segmentation methods STCN [4] and RDE [20] (with ground truth masks during training and using the mask from GSD [30] during evaluation). In all our examples, during inference, the initial masks for PGVS-Net are provided by the polarization-aware PGSNet [29]. The quantitative and qualitative results are shown in Table 2 and Figure 6.

Methods	IoU↑	F_β↑	MAE↓	BER↓
GSD_{IS} [30]	80.29	0.828	**0.115**	12.34
EAF_{IS} [49]	41.69	0.556	0.450	43.91
PM-RCNN_{IS} [14]	76.13	0.797	0.140	15.83
PGS_{IS} [29]	79.69	**0.873**	0.119	**12.31**
STICT_{VU} [25]	50.68	0.577	0.299	32.03
ViSha_{VS} [3]	77.42	0.805	0.136	13.77
STCN_{VS} [4]	**80.85**	0.826	0.120	12.77
RDE_{VS} [20]	74.87	0.763	0.169	15.95
PGVS-Net (Ours)	**84.60**	**0.867**	**0.099**	**10.02**

Table 2. Quantitative comparison of PGVS-Net with prior SOTA methods (method subscripts: *IS*: Image-based supervised, *VS*: Video-based supervised, and *VU*.: Video-based without labels). The top and second best results are highlighted in red and blue, respectively.

PGVS-Net outperforms state-of-the-art methods by a significant margin and improves IoU, MAE, and BER by 3.75, 0.016, and 2.29 over the next best method, demonstrating the effectiveness of multi-view polarization propagation for video glass segmentation. PGVS-Net performs slightly less than PGS-Net on the F_β metric, which again illustrates the robustness of polarization cues for glass segmentation for both still-image and video sequences. Figure 6 further qualitatively illustrates the robustness of PGVS-Net. We refer to the video sequences in the supplementary material to qualitatively evaluate the temporally

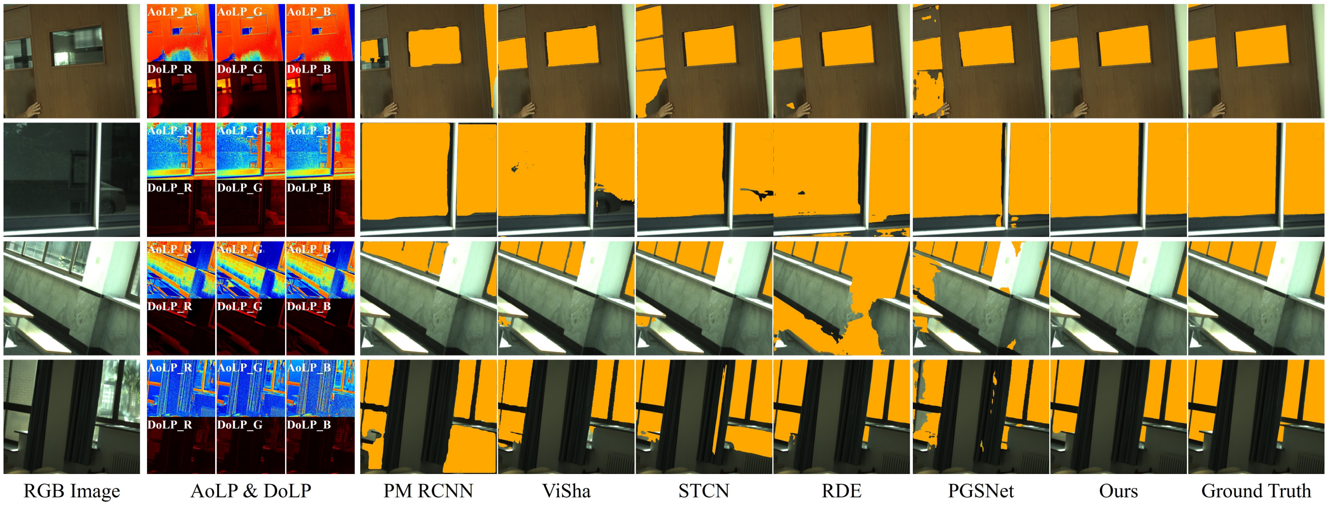


Figure 6. Qualitative comparison of our model against state-of-the-art image-based glass segmentation methods and video detection/segmentation methods. All models are retrained on PGV-117.

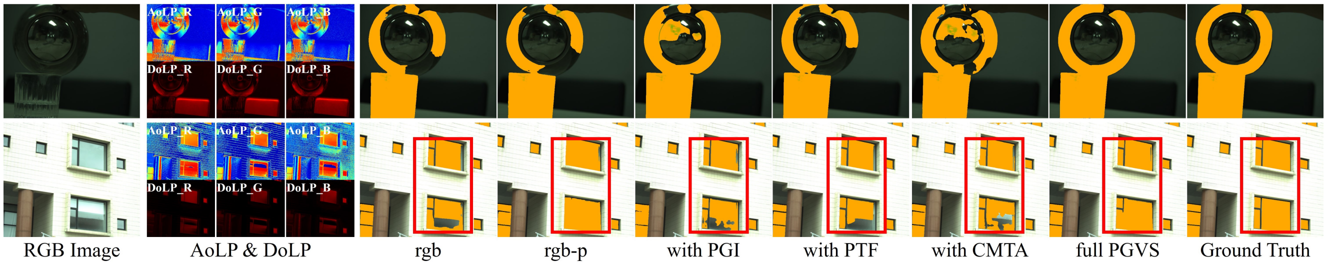


Figure 7. Qualitative comparison of the different ablation combinations of PGVS-Net.

Networks	IoU↑	F_β↑	MAE↓	BER↓
RGB	78.96	0.807	0.142	13.88
RGB-P	81.01	0.827	0.131	13.16
RGB-P+PGI	83.68	0.860	0.101	10.43
RGB-P+PTF	81.78	0.838	0.122	12.38
RGB-P+CMTA	83.81	0.862	0.100	10.88
RGB-P+PGI+CMTA+PTF(M5)	80.76	0.826	0.134	12.82
PGVS-Net	**84.60**	**0.867**	**0.099**	**10.02**

Table 3. Summary of ablation experiments: Polarization cues (RGB-P) outperform RGB-only cues. The combination of all modules (PGVS-Net) produces the overall best results. 'M5' indicates that the long-range memory is updated every five frames during inference.

coherent glass segmentation propagation.

5.3. Ablation Study

We validate the impact of each component in PGVS-Net by disabling one or more components and comparing the performance on the PGV-117 test set (Table 3). The network named 'RGB' uses only RGB textures to populate and query the memory to predict glass areas without any other modules. The 'RGB-P' network uses RGB-P information to store, read and calculate memory features. All other networks start from the 'RGB-P' baseline with the listed modules enabled. Finally, we indicate with 'M5' that the memory is updated every five frames during the inference stage as opposed to every three frames for PGVS-Net.

The improved performance of the 'RGB-P' network over 'RGB' shows that polarization provides additional cues for glass segmentation. PGI combines tripartite key features to retrieve affinity from historical views. Its attribute performance gain shows that integrated RGB-P streams can better mine multi-view similarity than single-stream matching. The PTF module further improves the segmentation accuracy (*e.g.*, IoU increases from 81.01 to 81.78). Adding the CMTA module further improves the performance on the four metrics, demonstrating that the cross-modal and cross-temporal integration provides rich short-range references for query frames. Finally, the full PGVS-Net with all modules enabled achieves the best results when combined with a memory update every three frames.

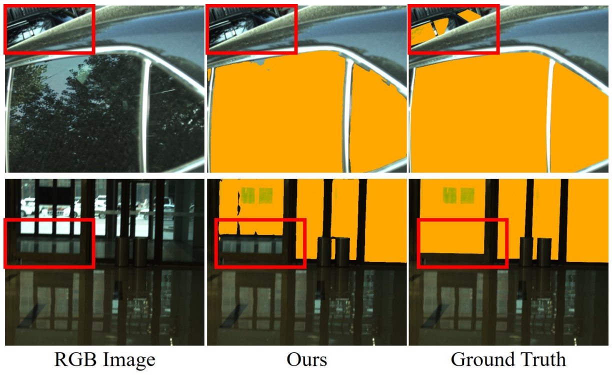


Figure 8. Two failure cases with inaccurate glass segmentation highlighted in the red boxes.

5.4. Failure Cases and Limitations

Our method propagates a single initial mask, and thus its overall quality is affected by the quality of the initial mask. For glass areas with no initial or historical references, little information can be propagated from multi-view observations. While PGVS-Net can augment glass segmentation in such scenarios by harnessing spatial characteristics across both the RGB and polarization domain, it is possible that the segmentation can result in gaps in the glass segmentation results (two failure cases are shown in Figure 8). An interesting avenue for future research would be to determine the most reliable RGB-P frame for initial mask estimation before propagating it to the rest of the sequence. Furthermore, although PGVS-Net significantly outperforms other video segmentation methods in accuracy, its tripartite-modal integration and temporal consistency strategy of polarization cues require additional inference time. Our unoptimized implementation currently operates at 7fps.

6. Conclusion

In this paper, we introduced a Polarization Video Glass Segmentation network (PGVS-Net), that exploits historical multi-view spectral polarization cues to segment glass areas, and which outperforms state-of-the-art polarization-related image and video segmentation networks. PGVS-Net leverages RGB-P memory and PGI-fused keys to correlate input frames with view-dependent polarization information from prior spectral cues. We show that the continuous changes in illumination are better captured by polarimetric features, and that long-range affinity helps to better balance single-input instabilities than single image methods. In addition, CMTA and PTF provide short-range polarization cues for low-level polarimetric features, providing rich wavelength details in the decoder phase. We also create a novel video glass segmentation dataset (PGV-117) to train and evaluate PGVS-Net.

Acknowledgements

This work was supported in part by National Key Research and Development Program of China (2022ZD0210500), the National Natural Science Foundation of China under Grants 61972067/U21A20491, and the Distinguished Young Scholars Funding of Dalian (No. 2022RJ01). Pieter Peers was supported in part by NSF grant IIS-1909028. Felix Heide was supported by an NSF CAREER Award (2047359), a Packard Foundation Fellowship, a Sloan Research Fellowship, a Sony Young Faculty Award, a Project X Innovation Award, and an Amazon Science Research Award. Seung-Hwan Baek acknowledges the support from NRF Korea (RS-2023-00211658, 2022R1A6A1A03052954) and Samsung Electronics.

References

[1] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

[2] Seung-Hwan Baek, Tizian Zeltner, Hyunjin Ku, Inseung Hwang, Xin Tong, Wenzel Jakob, and Min H Kim. Image-based acquisition and modeling of polarimetric reflectance. *ACM TOG*, 2020.

[3] Zhihao Chen, Liang Wan, Lei Zhu, Jia Shen, Huazhu Fu, Wennan Liu, and Jing Qin. Triple-cooperative video shadow detection. In *CVPR*, 2021.

[4] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *NeurIPS*, 2021.

[5] Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In *CVPR*, 2022.

[6] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, 2019.

[7] Wenbin Ge, Xiankai Lu, and Jianbing Shen. Video object segmentation using global and instance embedding learning. In *CVPR*, 2021.

[8] Hao He, Xiangtai Li, Guangliang Cheng, Jianping Shi, Yunhai Tong, Gaofeng Meng, Veronique Prinet, and LuBin Weng. Enhanced boundary learning for glass-like object segmentation. In *ICCV*, 2021.

[9] Hao He, Xiangtai Li, Guangliang Cheng, Jianping Shi, Yunhai Tong, Gaofeng Meng, Véronique Prinet, and LuBin Weng. Enhanced boundary learning for glass-like object segmentation. In *ICCV*, 2021.

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

[11] Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, and Rong Jin. Learning position and target consistency for memory-based video object segmentation. In *CVPR*, 2021.

[12] Dong Huo, Jian Wang, Yiming Qian, and Yee-Hong Yang. Glass segmentation with rgb-thermal image pairs. *arXiv preprint arXiv:2204.05453*, 2022.

[13] Samvit Jain, Xin Wang, and Joseph E Gonzalez. Accel: A corrective fusion network for efficient semantic segmentation on video. In *CVPR*, 2019.

[14] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *CVPR*, 2020.

[15] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *CVPR*, 2020.

[16] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019.

[17] Trung-Nghia Le and Akihiro Sugimoto. Deeply supervised 3d recurrent fcn for salient object detection in videos. In *BMVC*, 2017.

[18] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *CVPR*, 2020.

[19] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *CVPR*, 2018.

[20] Mingxing Li, Li Hu, Zhiwei Xiong, Bang Zhang, Pan Pan, and Dong Liu. Recurrent dynamic embedding for video object segmentation. In *CVPR*, 2022.

[21] Rui Li, Simeng Qiu, Guangming Zang, and Wolfgang Heidrich. Reflection separation via multi-bounce polarization state tracing. In *ECCV*, 2020.

[22] Xiangtai Li, Wenwei Zhang, Jiangmiao Pang, Kai Chen, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. Video k-net: A simple, strong, and unified baseline for video segmentation. In *CVPR*, 2022.

[23] Yule Li, Jianping Shi, and Dahua Lin. Low-latency video semantic segmentation. In *CVPR*, 2018.

[24] Yuhao Liu, Jiake Xie, Xiao Shi, Yu Qiao, Yujie Huang, Yong Tang, and Xin Yang. Tripartite information mining and integration for image matting. In *CVPR*, 2021.

[25] Xiao Lu, Yihong Cao, Sheng Liu, Chengjiang Long, Zipei Chen, Xuanyu Zhou, Yimin Yang, and Chunxia Xiao. Video shadow detection via spatio-temporal interpolation consistency training. In *CVPR*, 2022.

[26] Xiankai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. Video object segmentation with episodic graph memory networks. In *ECCV*, 2020.

[27] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, 2014.

[28] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. In *CVPR*, 2021.

[29] Haiyang Mei, Bo Dong, Wen Dong, Jiaxi Yang, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, and Xin Yang. Glass segmentation using intensity and spectral polarization cues. In *CVPR*, 2022.

[30] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson W.H. Lau. Don't hit me! glass detection in real-world scenes. In *CVPR*, 2020.

[31] Vu Nguyen, Tomas F Yago Vicente, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. Shadow detection with conditional generative adversarial networks. In *ICCV*, 2017.

[32] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *CVPR*, 2018.

[33] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019.

[34] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[35] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *CVPR*, 2021.

[36] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *CVPR*, 2020.

[37] Yu Qiao, Jincheng Zhu, Chengjiang Long, Zeyao Zhang, Yuxin Wang, Zhenjun Du, and Xin Yang. Cpral: Collaborative panoptic-regional active learning for semantic segmentation. In *AAAI*, 2022.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015.

[39] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *ECCV*, 2020.

[40] Evan Shelhamer, Kate Rakelly, Judy Hoffman, and Trevor Darrell. Clockwork convnets for video semantic segmentation. In *ECCV*, 2016.

[41] Xin Tian, Ke Xu, Xin Yang, Lin Du, Baocai Yin, and Rynson WH Lau. Bi-directional object-context prioritization learning for saliency ranking. In *CVPR*, 2022.

[42] Xin Tian, Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Weakly-supervised salient instance detection. In *BMVC*, 2020.

[43] Xin Tian, Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to detect instance-level salient objects using complementary image labels. *IJCV*, 2021.

[44] Dorian Tsai, Donald G. Dansereau, Thierry Peynot, and Peter Corke. Distinguishing refracted features using light field cameras with application to structure from motion. *IEEE Robotics and Automation Letters*, 2019.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017.

[46] Patrick Wieschollek, Orazio Gallo, Jinwei Gu, and Jan Kautz. Separating reflection and transmission images in the wild. In *ECCV*, 2018.

[47] Sanghyun Woo, Dahun Kim, Joon-Young Lee, and In So Kweon. Learning to associate every segment for video panoptic segmentation. In *CVPR*, 2021.

[48] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.

[49] Kaite Xiang, Kailun Yang, and Kaiwei Wang. Polarization-driven semantic segmentation via efficient attention-bridged fusion. *Optics Express*, 2021.

[50] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting rent objects in the wild. In *ECCV*, 2020.

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

[52] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson WH Lau. Where is my mirror? In *ICCV*, pages 8809–8818, 2019.

[53] Letian Yu, Haiyang Mei, Wen Dong, Ziqi Wei, Li Zhu, Yuxin Wang, and Xin Yang. Progressive glass segmentation. *IEEE TIP*, 2022.

[54] Jiaojiao Zhao, Yanyi Zhang, Xinyu Li, Hao Chen, Bing Shuai, Mingze Xu, Chunhui Liu, Kaustav Kundu, Yuanjun Xiong, Davide Modolo, et al. Tuber: Tubelet transformer for video action detection. In *CVPR*, 2022.

[55] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *CVPR*, 2017.