Multi-view Spectral Polarization Propagation for Video Glass Segmentation (Supplementary Material)

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

This supplementary document provides more details of the proposed PGV-117 dataset (§ 1), the formal definitions of the four quantitative metrics (§ 2), and the detailed calculation of key affinity (§ 3).

1. PGV-117 Dataset

The ground truth masks of the proposed PGV-117 dataset are annotated by annotation professionals, resulting in 144, 686 glass masks. Each ground truth mask is manually checked to ensure the quality of the annotations.

The proposed dataset consists of 117 sequences and 21, 485 frames. The training set offers 85 sequences, 15, 838 frames, and the testing set provides 32 sequences, 5, 647 frames. Figure 1 and Figure 2 show the number of frames for each sequence in the training and testing set, respectively.

2. Formal Definition of Evaluation Metrics

We adopt the four metrics used by Mei *et al.* [5] for evaluating all competing approaches, which are intersection over union (IoU), weighted F-measure (F_β) [4], mean absolute error (MAE), and balance error rate (BER) [6]. Here, we provide the formal definitions of these four metrics.

Intersection over union (IoU)

$$IoU = \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} (G(i,j) * P_b(i,j))}{\sum_{i=1}^{H} \sum_{j=1}^{W} (G(i,j) + P_b(i,j) - G(i,j) * P_b(i,j))},$$
(1)

where G is the ground truth mask in which the values of the glass region are 1 while those of the non-glass region are 0; P_b is the predicted mask binarized with a threshold of 0.5; and H and W are the height and width of the ground truth mask, respectively.

Weighted F-measure (\mathbf{F}_{β}) takes a prediction map's precision and recall into account, which is a common metric used in salient object detection tasks. Based on recent studies [2, 3], the weighted F-measure [4] is more reliable than the traditional \mathbf{F}_{β} [5], and it is used in our evaluation.

Mean Absolute Error (MAE) calculates the element-wise distance between a prediction map P and the corresponding ground truth mask G:

$$MAE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} |P(i,j) - G(i,j)|$$
(2)

where P(i, j) indicates the predicted probability score at location (i, j).

dynamic_door_daytime1 dynamic_door_daytime3 dynamic_door_daytime4 dynamic door daytime6 dynamic_door_dusk1 dynamic_door_dusk2 dynamic_door_inclined_dusk4 dynamic_glass_night2 dynamic_high-light_window1 dynamic_kettle_night1 dynamic_low-light_door1 dynamic_low-light_window1 dynamic_window_daytime1 dynamic_window_dusk1 dynamic_window_dusk3 dynamic_window_night1 HDR_building_circle2 HDR_building_circle3 HDR building circle4 HDR_building_circle5 HDR_building_circle7 HDR_building_circle8 HDR_building_front2 HDR_building_front3 HDR_building_inclined1 HDR_building_inclined2 HDR_building_updown1 HDR_building_updown2 high-light_building_circle1 high-light_building_front1 high-light_building_inclined1 inside_building_daytime_circle1 inside_building_daytime_circle3 inside building daytime circle4 inside_building_daytime_inclined1 inside_building_dusk_inclined1 inside_building_updown1 inside_shopping_dusk_circle1 inside_shopping_dusk_front1 inside_shopping_dusk_inclined_ground1 inside_shopping_dusk_inclined_ground2 inside_shopping_dusk_inclined1 inside_shopping_dusk_inclined2 inside_shopping_dusk_inclined3 inside shopping dusk straight1 inside_shopping_night_circle1 inside_shopping_night_circle2 inside_shopping_night_front_ground1 inside_shopping_night_front1 inside_shopping_night_front2 inside_shopping_night_straight1 inside_shopping_night_straight3 low-light_building_inclined2 low-light_inside_building_daytime_front1 low-light_inside_building_night_front2 low-light_shopping_inclined1 low-light_shopping_inclined2 outside_building_daytime_front1 outside_building_daytime_front2 outside building daytime front3 outside_building_daytime_front5 outside_building_daytime_front6 outside_building_daytime_front7 outside_building_daytime_front8 outside_building_daytime_front9 outside_building_daytime_front10 outside_building_daytime_front12 outside_building_daytime_front13 outside_building_daytime_front15 outside_building_daytime_front16 outside building daytime front17 outside_building_daytime_front18 outside_building_daytime_front19 outside_building_daytime_front21 outside building daytime front22 outside_building_daytime_front23 outside_building_daytime_front26 outside_building_daytime_front27 inside_building_daytime_inclined1 outside_building_daytime_updown1 outside_building_daytime_updown3 outside building dusk front1 outside_car_daytime_circle2 outside_kettle_night_straight1



Figure 1. Summary for training set distribution of PGV-117, which includes 85 video sequences in total.



Figure 2. Summary for testing set distribution of PGV-117, which includes 32 video sequences in total. In order not to lose generality, all lighting conditions, camera motion patterns, and dynamics are also included in the testing set.

Balance error rate (BER) is a common metric used in shadow detection tasks. Formally, it is defined as:

$$BER = \left(1 - \frac{1}{2}\left(\frac{TP}{N_p} + \frac{TN}{N_n}\right)\right) \times 100$$
(3)

where TP, TN, N_p , and N_n represent the numbers of true positive pixels, true negative pixels, glass pixels, and non-glass pixels, respectively.

3. Computing Affinity

For the query frame t, we relate multi-view RGB-P information by exploring the relationship between the PGI key of t with the keys in the memory (0 to t - 1). After generating the query key k^Q and memory keys k^M , we refer to [7, 1] to calculate the affinity between k^Q and k^M :

$$a = \xi[(k^M)^2],$$

$$b = 2 * [(k^M)^T \circledast k^Q,$$

$$A = (-a+b)/\sqrt{CK},$$

$$A = \frac{exp(A_{ij})}{\sum_n (exp(A_{nj}))}.$$

(4)

where $\xi[\cdot]$ represents the summation and unsqueeze operation, \circledast means matrix multiplication. CK is the number of channels for the key features, and *i* denotes the affinity value at the *i*-th position. The affinity considers both RGB similarity and multi-view spectral consistency between the query and memory frames. The value encoder output v^M corresponding to k^M contains features from the tripartite memory values, and the multiplication of affinity and v^M correlates the query frame and historical information to obtain v^Q to participate in the readout of the memory bank.

References

- 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. 3
- [2] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In ICCV, 2017. 1
- [3] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, 2018. 1
- [4] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In CVPR, 2014. 1
- [5] 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. 1
- [6] Vu Nguyen, Tomas F Yago Vicente, Maozheng Zhao, Minh Hoai, and Dimitris Samaras. Shadow detection with conditional generative adversarial networks. In *ICCV*, 2017. 1
- [7] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In ICCV, 2019. 3