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

Anonymous ICCV submission

Paper ID 6969

This supplementary document provides more details of the proposed PGV-117 dataset (\S 1), the formal definitions of the four quantitative metrics (\S 2), and the detailed calculation of key affinity (\S 3). Five processed video sequences are provided along with this document. The teaser and the visual results shown in section 5.2 of this submission are from these five videos. We also offer additional video results through Google Drive.

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.

 $\begin{array}{c} \textbf{Mean Absolute Error (MAE)} \\ \textbf{ground truth mask } G: \\ \textbf{Mean Absolute Error (MAE)} \\ \textbf{ground truth mask } G: \\ \textbf{Mean Absolute Error (MAE)} \\ \textbf{Mean Absolute Error (MAE)} \\ \textbf{ground truth mask } G: \\ \textbf{Mean Absolute Error (MAE)} \\ \textbf{Mean Absolute Error (MAE)} \\ \textbf{ground truth mask } G: \\ \textbf{Mean Absolute Error (MAE)} \\ \textbf{Mean Absolute Error (MAE)} \\ \textbf{ground truth mask } G: \\ \textbf{Mean Absolute Error (MAE)} \\ \textbf{Mean Absolute Error (MAE)} \\ \textbf{Mean Absolute Error (MAE)} \\ \textbf{G: } \\ \textbf{G: } \\ \textbf{Mean Absolute Error (MAE)} \\ \textbf{G: } \\ \textbf$

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

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

ICCV 2023 Submission #6969. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

ICCV #6969

108	dynamic_door_daytime1	92	162
100	dynamic_door_daytime3	208	162
105	dynamic_door_daytime4	132	105
110	dynamic_door_daytime6	1/0	164
444	dynamic_door_dusk1	155	165
	dynamic_door_inclined_dusk4	239	105
112	dynamic glass night?	201	166
440	dynamic high-light window1	200	107
113	dynamic kettle night1	103	107
114	dynamic low-light door1	222	168
	dynamic low-light window1	249	
115	dynamic_window_daytime1	118	169
116	dynamic_window_dusk1	230	170
110	dynamic_window_dusk3	134	
117	dynamic_window_night1	244	171
110	HDR_building_circle2	100	170
110	HDR_building_circle3	147	172
119	HDR_building_circle4	193	173
100	HDR_building_circle5	220	174
120	HDR_building_circle/	134	174
121	HDR_building_critices	116	175
100	HDR_building_front3	200	170
122	HDR building inclined1	125	170
123	HDR_building_inclined2	125 166	177
	HDR building updown1	152	170
124	HDR building updown2	200	178
125	high-light building circle1	142	179
120	high-light_building_front1	194	
126	high-light_building_inclined1	150	180
107	inside_building_daytime_circle1	222	101
121	inside_building_daytime_circle3	123	101
128	inside_building_daytime_circle4	72	182
400	inside_building_daytime_inclined1	201	100
129	inside_building_dusk_inclined1	87	183
130	inside_building_updown1	200	184
	inside_shopping_dusk_circle1	131	
131	inside_shopping_dusk_front1	200	185
132	inside_shopping_dusk_inclined_ground1	200	186
102	inside_shopping_dusk_inclined_ground2	125	100
133	inside_shopping_dusk_inclined?	123	187
124	inside shopping dusk inclined3		100
104	inside shopping dusk straight 1	151	100
135	inside shopping night circle1	175	189
100	inside shopping night circle2	122	100
130	inside_shopping_night_front_ground1	167	190
137	inside_shopping_night_front1	150	191
100	inside_shopping_night_front2	120	100
138	inside_shopping_night_straight1	150	192
139	inside_shopping_night_straight3	171	193
	low-light_building_inclined2	200	
140	low-light_inside_building_daytime_front1	200	194
1/1	low-light_inside_building_night_front2	140	105
141	low-light_shopping_inclined	180	195
142	low-light_shopping_inclined2	80	100 196
140	outside_building_daytime_front1	225	107
145	outside_building_daytime_front3	304	197
144	outside_building_daytime_front5	157	198
	outside_building_daytime_front6	289	100
145	outside building daytime front7	258	199
146	outside building daytime front8	225	200
	outside_building_daytime_front9	235	
147	outside_building_daytime_front10	147	201
148	outside_building_daytime_front12	242	202
	outside_building_daytime_front13	310	202
149	outside_building_daytime_front15	161	203
150	outside_building_daytime_front16	280	204
150	outside_building_daytime_front17	149	204
151	outside_building_daytime_tront18	240	205
150	outside_building_daytime_front19	240	000
152	outside_building_daytime_front21	229	200
153	outside_building_daytime_front22	219 250	207
454	outside_building_daytime_front26	200	
154	outside building davtime front27	150	208
155	inside building davtime inclined	218	209
	outside building davtime undown1	150	200
156	outside building davtime updown3	191	210
157	outside building dusk front1	131	011
107	outside car daytime circle2	244	211
158	outside_kettle_night_straight1	172	212
150	outside_shooping_dusk_inclined1	202	010
109	E 1.0 6 4	nining and distribution of DCW 117 which is 1, 1, 05, 11	213
160	Figure 1. Summary for t	raining set distribution of PGv-117, which includes 85 video sequences in total.	214
404			0.17
161			215

ICCV 2023 Submission #6969. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.



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 :

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

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

$$A = (-a+b)/\sqrt{CK},\tag{4}$$

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

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 [1] 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 fore-ground 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