EVOLVE: Event-Guided Deformable Feature Transfer and Dual-Memory Refinement for Low-Light Video Object Segmentation

Supplementary Material

Method	$\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
Cutie	0.635	0.652	0.667	0.579	0.642
OneVOS	0.642	0.657	0.663	0.591	0.657
EVOLVE	**0.753**	**0.742**	**0.780**	**0.703**	**0.788**

Table 1. Quantitative comparison on the LLE-YoutubeVOS 2019 dataset. The best results are boldfaced.

S-1. Comparison with state-of-the-arts on LLE-YoutubeVOS 2019

We validate EVOLVE on larger dataset, YouTube-VOS 2019, by constructing synthetic low-light images with event following the setting in [2]. As shown in Table 1, our method consistently outperforms recent state-of-the-art VOS [1, 3], demonstrating generalization to large-scale dataset.

K	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	FPS
0	0.709	0.731	0.720	45.3
1	0.741	0.760	0.750	42.7
5	<u>0.779</u>	**0.783**	**0.781**	37.6
7	**0.780**	<u>0.781</u>	<u>0.780</u>	34.2
Ours (3)	0.775	0.771	0.773	39.3

Table 2. Results according to the number of DMOT blocks K.

N	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
0	0.709	0.731	0.720
1	0.751	0.740	0.745
5	<u>0.766</u>	<u>0.761</u>	<u>0.763</u>
Ours (3)	**0.775**	**0.771**	**0.773**

Table 3. Results according to the number of reference frames N.

Q	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
8	0.760	0.759	0.759
24	<u>0.771</u>	<u>0.768</u>	<u>0.770</u>
32	0.764	0.762	0.763
Ours (16)	**0.775**	**0.771**	**0.773**

Table 4. Results according to the number of memory tokens Q.

S-2. Hyperparameter

Tables 2, 3, and4 provide the results according to the number of DMOT blocks K, reference frames N, memory tokens Q, respectively, on the LLE-VOS dataset. In Table 2, we determine that setting the number of DMOT blocks to three achieves the best trade-off between performance and FPS. As shown in Tables 3 and 4, we select three reference frames and 16 memory tokens, which yield the highest performance. Table 5 presents the results according to the confidence score threshold for binarization of auxiliary masks in MRM. Experimentally, we achieve the optimal performance when the threshold is set to 0.65.

Confidence score thresholds	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
0.4	0.763	0.768	0.765
0.5	0.767	0.762	0.764
0.6	<u>0.772</u>	<u>0.770</u>	<u>0.771</u>
0.7	0.768	0.759	0.763
Ours (0.65)	**0.775**	**0.771**	**0.773**

Table 5. Results according to confidence score thresholds for mask binarization of auxiliary masks.

Method	Noise rate 0%	Noise rate 10%	Noise rate 20%
Cutie	0.672	0.623 (4.9 ↓)	0.598 (7.4 ↓)
OneVOS	0.683	0.633 (5.0 ↓)	0.606 (7.7 ↓)
EVOLVE	0.773	0.748 (2.5 ↓)	0.729 (4.4 ↓)

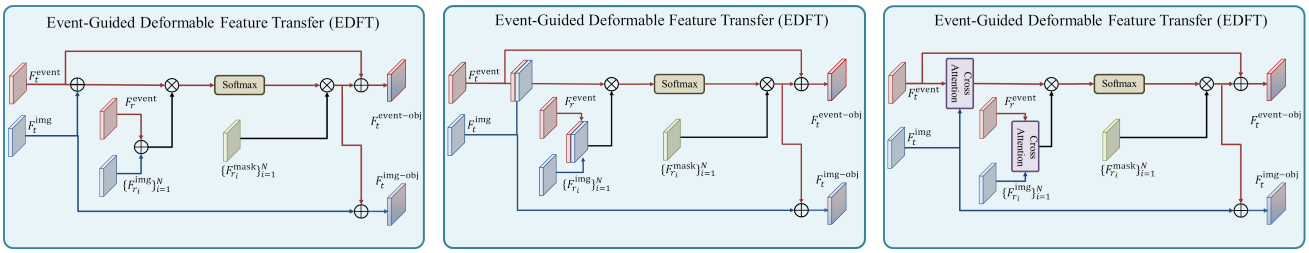
Table 6. Results according to the rate of event noise.

S-3. Robustness on Event Noise

Table 6 show the results on LLE-VOS under different rate of event noise. The event noise is simulated by randomly flipping event polarities according to the specified noise ratio. Compared to recently VOS methods [1, 3], the proposed method shows less performance drop under event noise, indicating strong robustness to noisy event data.

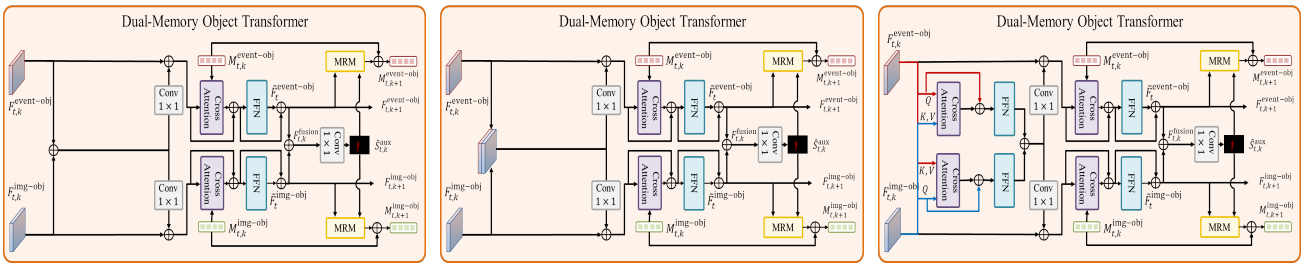
S-4. Loss function

Table 7 lists results according to different loss settings on LLE-VOS. Removing the cross-entropy loss or Dice loss degrades the segmentation performance. Moreover, we observe that including Dice loss for auxiliary masks obtained from DMOT imporves the performance. This indicates that the predicted auxiliary masks provide informative cues for effective memory updates.



S_1 in EDFT S_2 in EDFT S_3 in EDFT

Figure 1. Diagrams of different feature fusion approaches in EDFT. S_1 in EDFT means image and event features are simple added. S_2 in EDFT means image and event image features are concatenated. Finally, S_3 in EDFT means employ cross-attention using image features as queries and event features as keys and values.



(a) S_4 in DMOT (b) S_5 in DMOT (c) S_6 in DMOT

Figure 2. Diagrams of different feature fusion approaches in DMOT. S_4 and S_5 in DMOT mean $F_t^{\text{img-obj}}$ and $F_t^{\text{event-obj}}$ simply added and concatenated, respectively. S_6 in DMOT means two separate cross-attentions between $F_t^{\text{img-obj}}$ and $F_t^{\text{event-obj}}$ are used.

Loss function	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
w/o cross-entropy loss	0.732	0.742	0.759
w/o Dice loss	0.711	0.689	0.700
w/o Dice loss for auxiliary masks	0.751	0.748	0.759
Ours	**0.775**	**0.771**	**0.773**

Table 7. Ablation study on different settings of loss function.

Methods	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
S_1 in EDFT	0.732	0.729	0.731
S_2 in EDFT	0.736	0.734	0.735
S_3 in EDFT	0.759	0.748	0.754
Ours	**0.775**	**0.771**	**0.773**

Table 8. Ablation study on different feature fusions in EDFT.

Methods	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
w/o self-attention	0.728	0.741	0.735
S_4 in DMOT	0.741	0.750	0.746
S_5 in DMOT	0.744	0.754	0.749
S_6 in DMOT	0.773	0.770	0.771
Ours	**0.775**	**0.771**	**0.773**

Table 9. Ablation study on different feature fusions in DMOT.

S-5. Different feature fusion strategies in EDFT and DMOT

Tables 8 and 9 present the ablation study results on different feature fusion methods in event-guided deformable feature transfer (EDFT) and dual-memory object transformer (DMOT), respectively. Detailed descriptions of these methods are provided in Figures 1 and 2.

As shown in Table 8, in the case of EDFT, S_1 in EDFT, simply adding image and event features, and S_2 in EDFT, concatenating them, result in low performances with $\mathcal{J}\&\mathcal{F}$ scores of 0.732 and 0.736, respectively. S_3 in EDFT, which employ cross-attention using image features as queries and event features as keys and values, achieves a $\mathcal{J}\&\mathcal{F}$ score of 0.754. In contrast, the proposed EVOLVE designs event-based deformable convolution to effectively align features before feature transfer, addressing spatial misalignment in low-light environments and achieving the highest $\mathcal{J}\&\mathcal{F}$ score of 0.773.

As shown in Table 9, removing self-attention in DMOT reduces the $\mathcal{J}\&\mathcal{F}$ score to 0.735, demonstrating the importance of global feature refinement. S_4 and S_5 in DMOT, where $F_t^{\text{img-obj}}$ and $F_t^{\text{event-obj}}$ are simply added and concatenated, achieve the lower $\mathcal{J}\&\mathcal{F}$ scores of 0.746 and 0.749, respectively. S_6 in DMOT, applying cross-attention to $F_t^{\text{img-obj}}$ and $F_t^{\text{event-obj}}$ each other, yields similar performance to our method. However, two separate cross-attention operations for each iteration require more computational costs than self-attention in EVOLVE. In contrast, the proposed DMOT provides the best $\mathcal{J}\&\mathcal{F}$ score of 0.773.

S-6. Qualitative Results

Figures 3, 4 and 5 illustrate qualitative comparisons of the proposed EVOLVE with Cutie [1] and OneVOS [3] on

LLE-DAVIS, LLE-VOS and LLE-YoutubeVOS 2019, respectively. The proposed method effectively segments objects that are closer to the ground-truth compared to Cutie and OneVOS. Figures 6 and 7 provides qualitative comparisons of EVOLVE according to the input modalities. We see that faithful segmentation results are obtained when both image and event data is used for VOS.

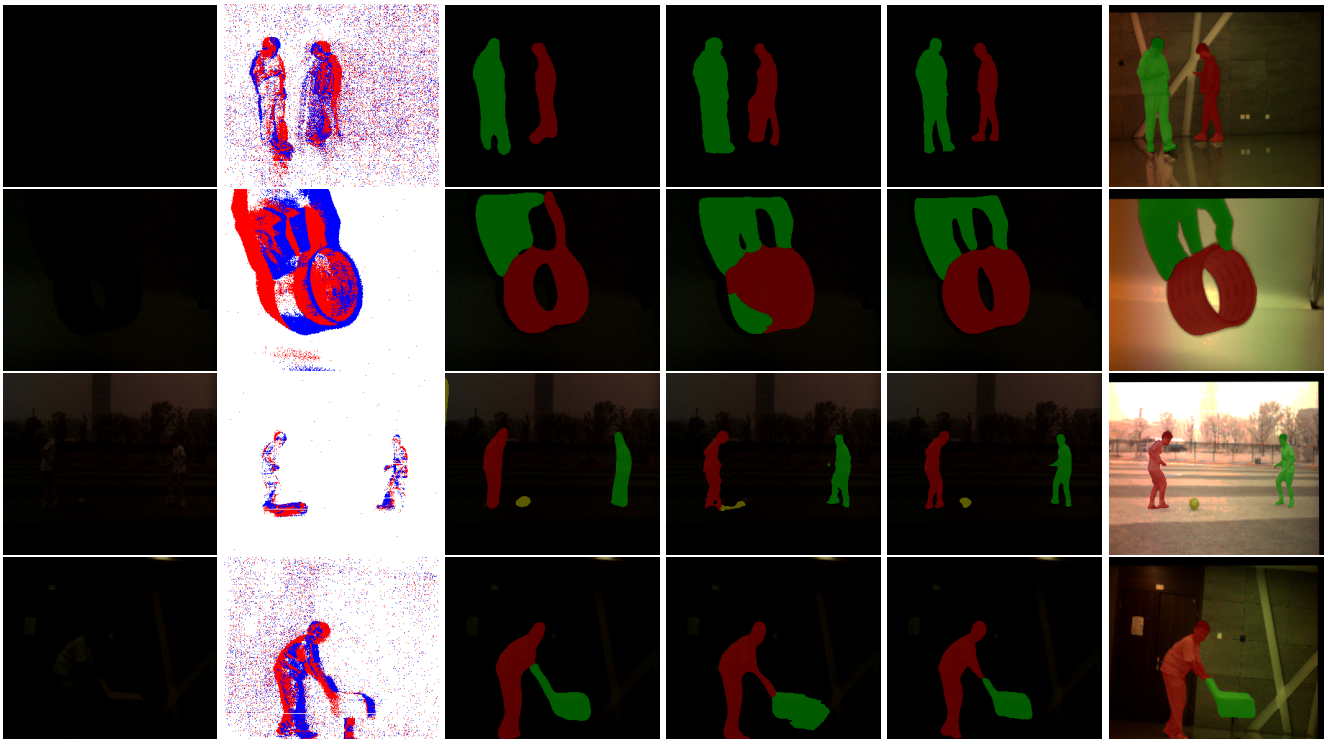
Source code. The source code for the EVOLVE implementation is included in the supplementary materials. We publicly release it after the paper is accepted.

References

[1] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In *CVPR*, pages 3151–3161, 2024. 1, 2

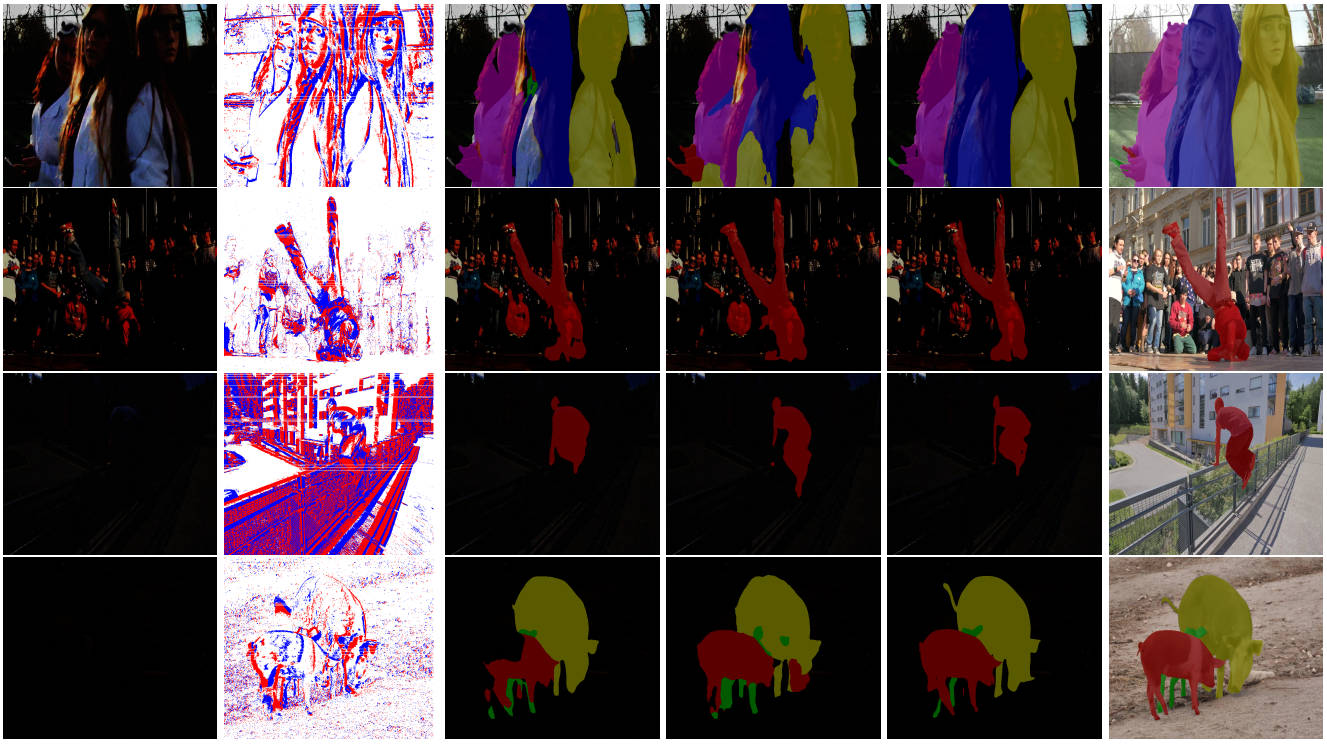
[2] Hebei Li, Jin Wang, Jiahui Yuan, Yue Li, Wenming Weng, Yansong Peng, Yueyi Zhang, Zhiwei Xiong, and Xiaoyan Sun. Event-assisted low-light video object segmentation. In *CVPR*, pages 3250–3259, 2024. 1

[3] Wanyun Li, Pinxue Guo, Xinyu Zhou, Lingyi Hong, Yangji He, Xiangyu Zheng, Wei Zhang, and Wenqiang Zhang. Onevos: Unifying video object segmentation with all-in-one transformer framework. In *ECCV*, pages 20–40, Cham, 2024. Springer Nature Switzerland. 1, 2



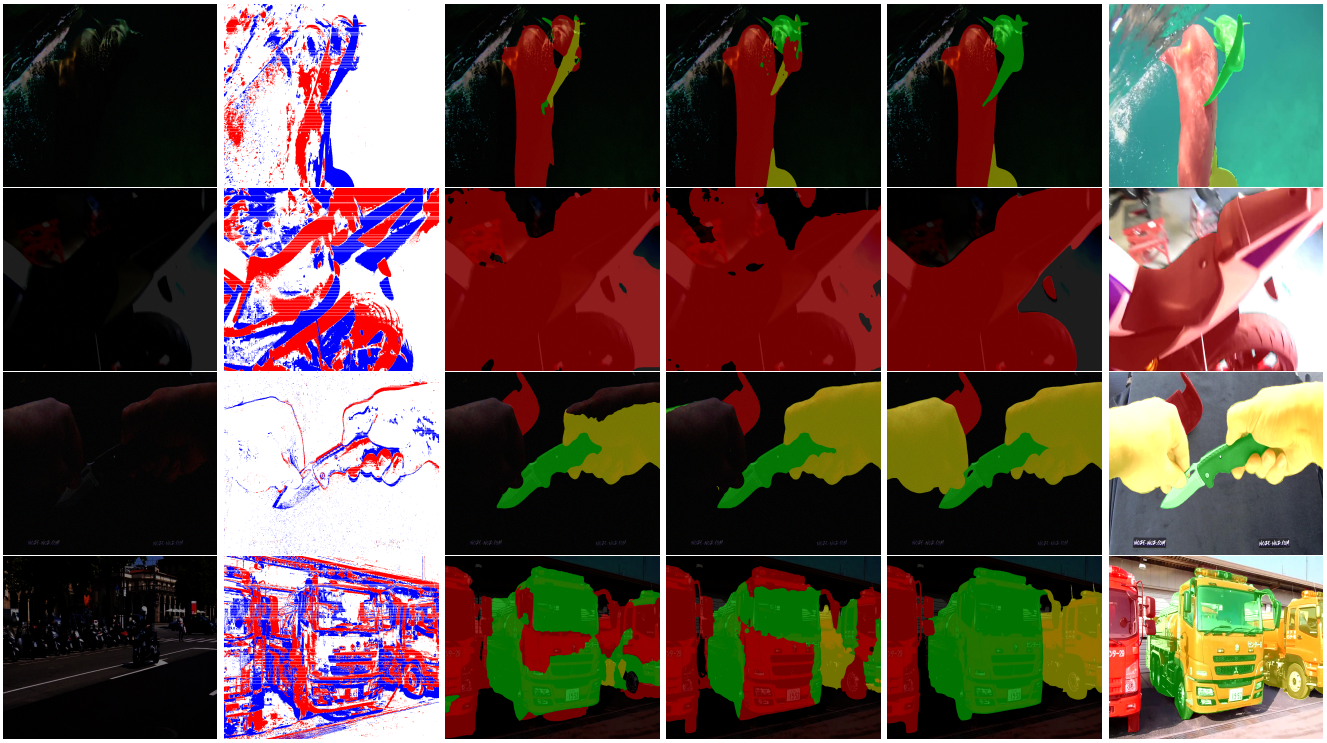
(a) Input image (b) Event (c) Cutie (d) OneVOS (e) Ours (f) Ground-truth

Figure 3. Qualitative comparisons on LLE-VOS.



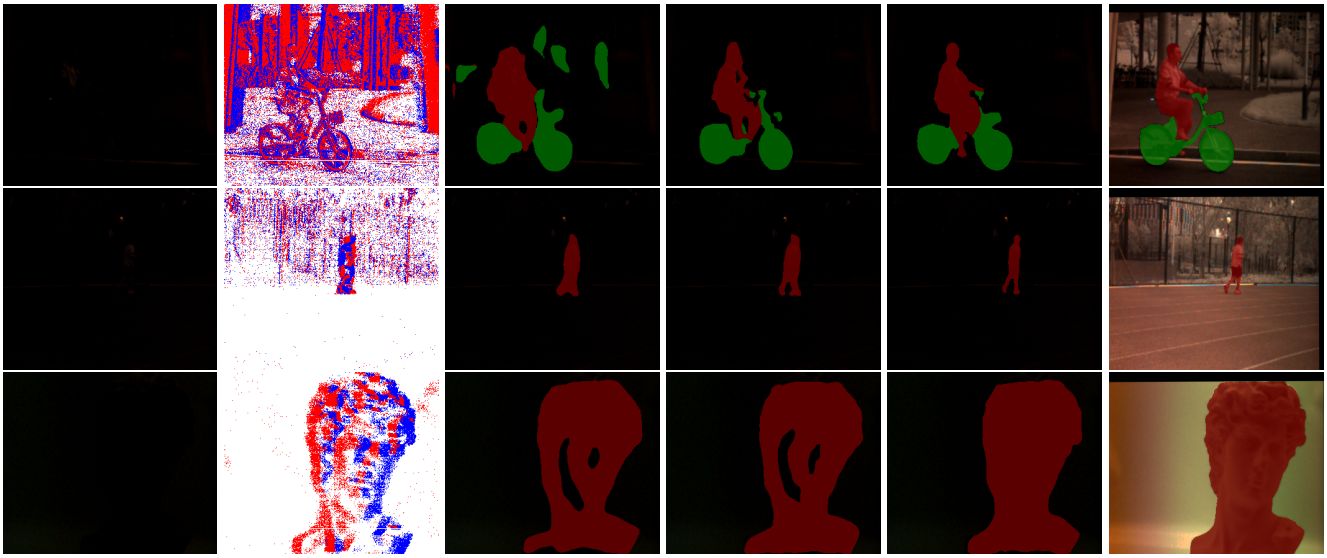
(a) Input image (b) Event (c) Cutie (d) OneVOS (e) Ours (f) Ground-truth

Figure 4. Qualitative comparisons on LLE-DAVIS.



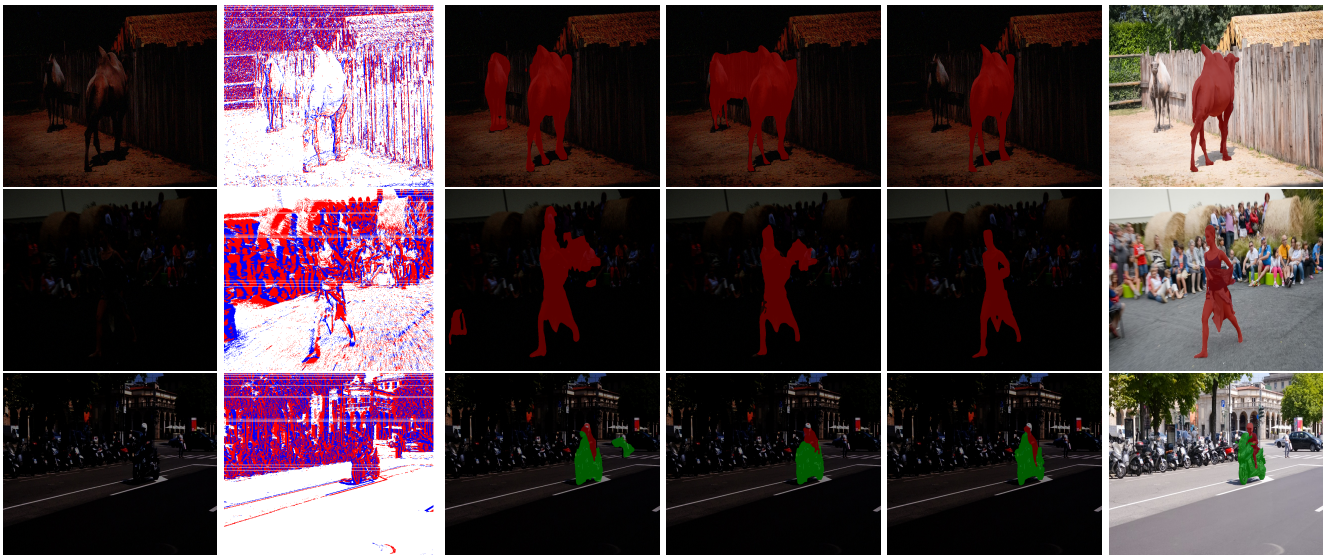
(a) Input image (b) Event (c) Cutie (d) OneVOS (e) EVOLVE (f) Ground-truth

Figure 5. Qualitative comparisons on LLE-YoutubeVOS 2019.



(a) Input image (b) Event (c) Image only (d) Event only (e) Ours (f) Ground-truth

Figure 6. Qualitative results of EVOLVE according to different input modalities on LLE-VOS.



(a) Input image (b) Event (c) Image only (d) Event only (e) Image + Event (f) Ground-truth

Figure 7. Qualitative results of EVOLVE according to different input modalities on LLE-DAVIS.