Supplementary Material: Event-assisted Low-Light Video Object Segmentation

Hebei Li [1] Jin Wang [1] Jiahui Yuan [1] Yue Li [1] Wenming Weng [1]
Yansong Peng [1] Yueyi Zhang [1,*] Zhiwei Xiong [1] Xiaoyan Sun [1, 2]

[1] University of Science and Technology of China

[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

{lihebei, jin01wang, yuanjiahui, yueli65, wmweng, pengyansong}@mail.ustc.edu.cn,
{zhyuey, zwxiong, sunxiaoyan}@ustc.edu.cn

In the supplementary material, we first describe the detailed collection pipeline of our real-world dataset LLE-VOS in Sec. 1 (as a supplement to Sec. 3 in the main paper), including the temporal synchronization, geometric calibration, and the segmentation annotation. Then, we conduct more experimental comparisons with other methods on LLE-DAVIS and LLE-VOS datasets in Sec. 2. Finally, we present the diversity of our proposed dataset and provide more qualitative results to further demonstrate the superiority of our method in Sec. 4.

1. LLE-VOS Collection Pipeline

1.1. Hybrid Camera System

The hybrid camera system is equipped with two Davis346 event cameras [3] and a beam splitter guaranteeing that both cameras record identical scenes. We provide details about the temporal synchronization and geometric calibration between the two cameras below.

For **temporal synchronization**, we first connect the synchronization signal trigger/receive ports of two cameras using a synchronous cable. Subsequently, one camera is designated as the master camera, and the other as the slave camera. When the master camera starts capturing data, it sends a trigger signal to the slave camera. Upon receiving this signal, the slave camera begins synchronized data capture. To ensure the synchronous capture of frames and events, we employ software-based triggering for the recording process. It is important to note that the communication time between the two cameras is negligible. Ultimately, this approach ensures synchronized data acquisition between the two cameras. For **geometric calibration**, we utilize a 5x8 chessboard calibration pattern, waving it in front of the camera setup to ensure visibility by both cameras. We select a corresponding pair of images, one from the normal-light camera view, denoted as I_{norm} and the other from the low-light camera view, denoted as I_{low}. Using a corner detection algorithm, we detect the chessboard corners in both images. Based on these detected corners, we employ an affine transformation matrix to model the transformation between the two cameras. Mathematically, we formulate it as:

$$p_i^{low} = H \cdot p_i^{norm}, \tag{1}$$

where H denotes a 3×3 transformation matrix, $p_i^{low} = [x_i^{low}, y_i^{low}, 1]^T$ and $p_i^{norm} = [x_i^{norm}, y_i^{norm}, 1]^T$ are the homogeneous coordinates in the low/normal-light pairs. We compute the transformation matrix

$$H = \begin{bmatrix} 1.0088e{+}0 & 7.0321e{-}3 & -1.2667e{+}1 \\ -1.3069e{-}3 & 1.0103e{+}0 & 1.2477e{+}1 \\ 3.4419e{-}6 & 2.0457e{-}5 & 1.0000e{+}0 \end{bmatrix}. \tag{2}$$

The computed reprojection RMS error is 0.2814 pixel. Ultimately, we apply the calculated matrix H to transform the normal-light frame, achieving spatial alignment between two camera views.

Figure 1. The platform we use to generate the annotation.

*Corresponding Author

Method	Input	Indoor Scenes			Outdoor Scenes			Overall		
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
STCN [NIPS2021] [2]	I	0.486	0.321	0.403	0.400	0.309	0.354	0.445	0.316	0.380
XMem [ECCV2022] [1]	I	0.664	0.528	0.596	0.507	0.456	0.481	0.590	0.494	0.542
AOT [NIPS2021] [6]	I	0.699	0.618	0.659	0.592	0.571	0.581	0.649	0.596	0.623
DeAOT [NIPS2022] [5]	I	0.716	0.643	0.680	0.580	0.580	0.580	0.653	0.614	0.633
STCN [NIPS2021] [2]	E+I	0.522	0.360	0.441	0.460	0.354	0.407	0.493	0.358	0.425
XMem [ECCV2022] [1]	E+I	0.732	0.616	0.674	0.525	0.465	0.495	0.635	0.545	0.590
AOT [NIPS2021] [6]	E+I	0.745	0.674	0.709	0.590	0.574	0.582	0.673	0.627	0.650
DeAOT [NIPS2022] [5]	E+I	0.746	0.678	0.712	0.596	**0.604**	**0.600**	0.675	0.643	0.659
Ours	E+I	**0.789**	**0.710**	**0.749**	**0.604**	0.588	0.596	**0.702**	**0.653**	**0.678**

Table 1. Quantitative comparisons of various VOS methods on the real-world LLE-VOS dataset. 'I' and 'E' represent the input of images and events, respectively. The best results are marked in bold.

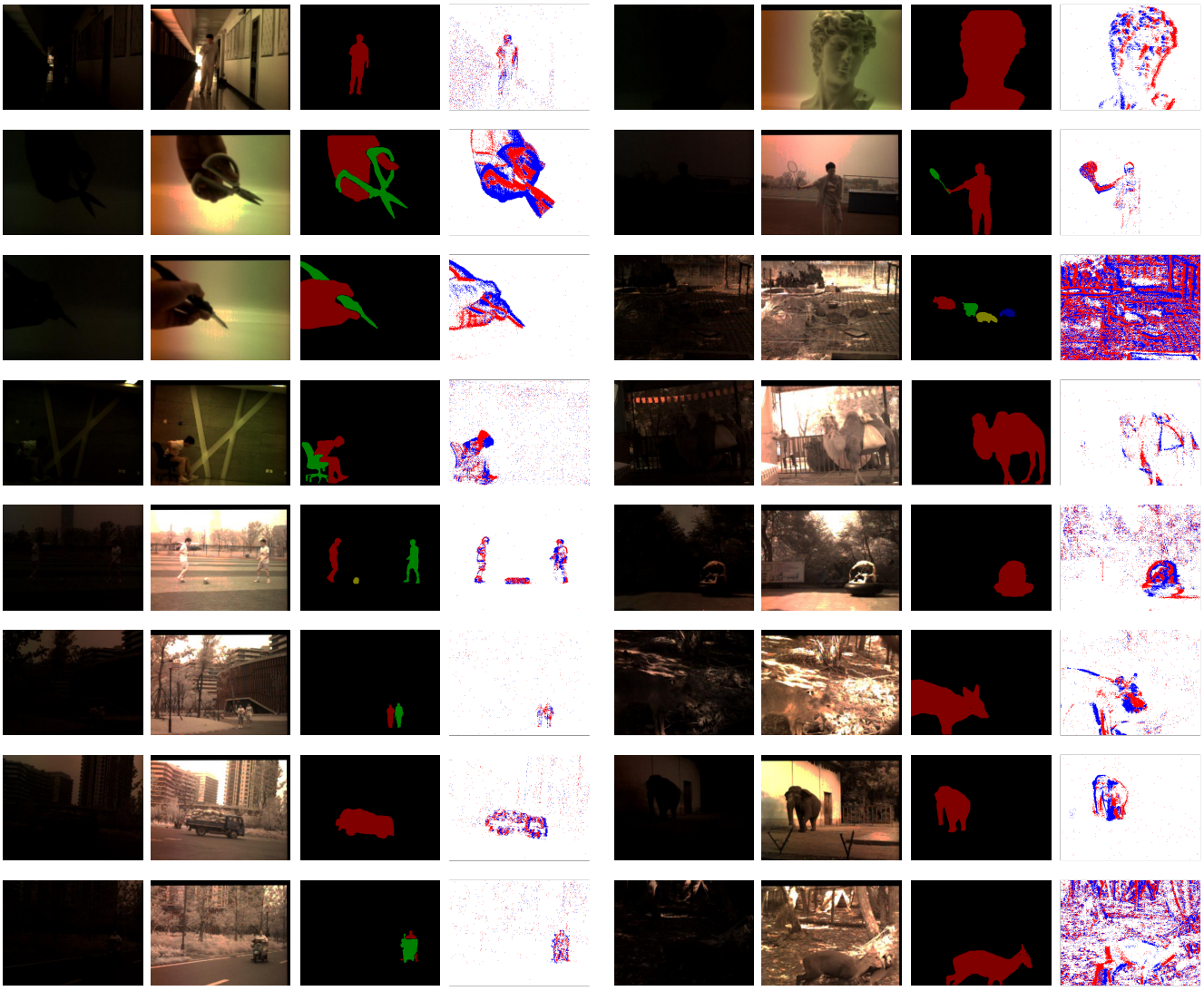

Figure 2. Examples of our real-world LLE-VOS dataset

.

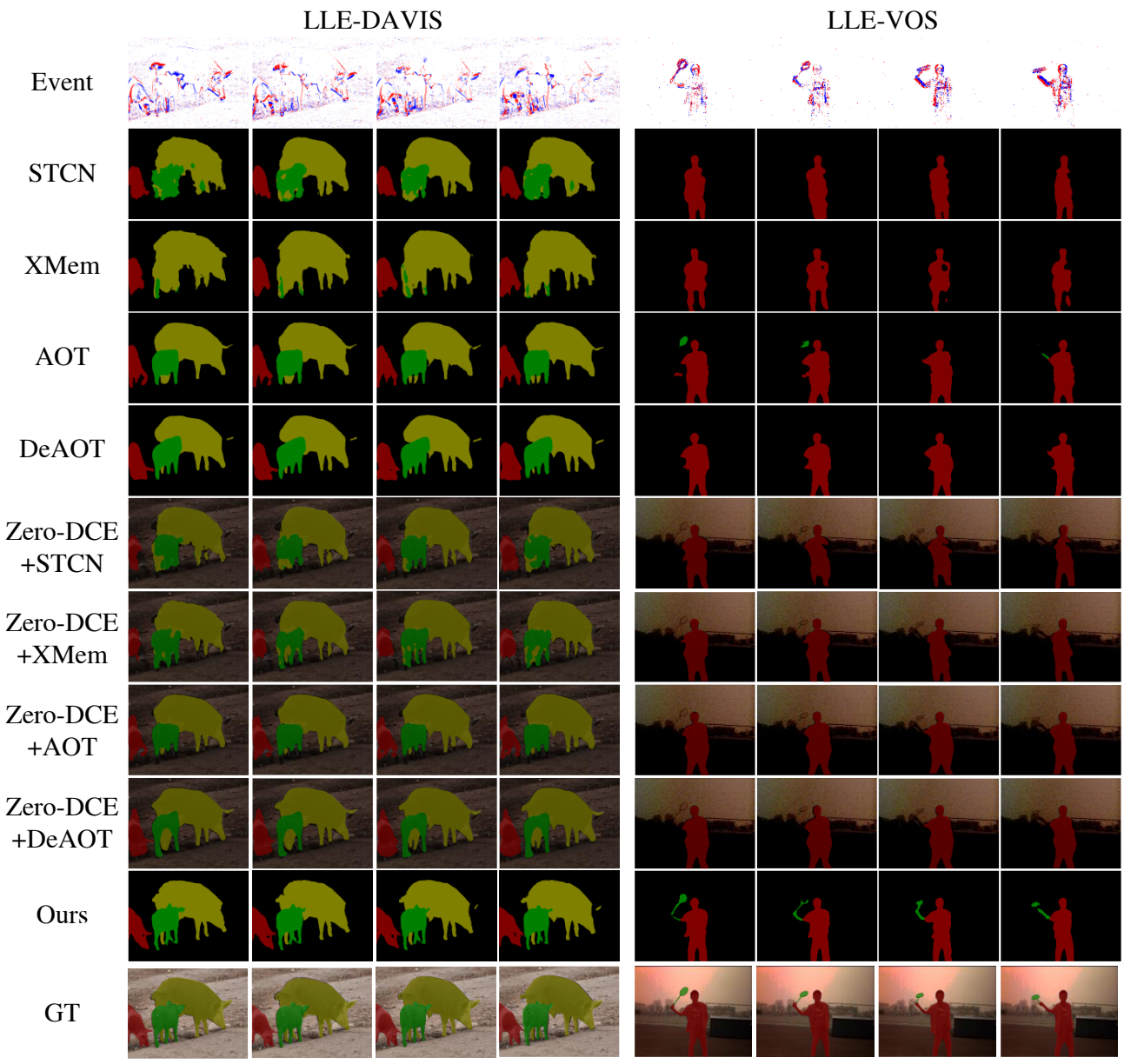

Figure 3. Qualitative comparisons with other methods on the synthetic LLE-DAVIS and the real-world LLE-VOS datasets.

1.2. Segmentation Annotations

After collecting the event and image pairs, we employ 20 volunteers to assist us with the annotation. The annotation platform is Label Studio [4], shown in Fig. 1. We eventually export the results in the JSON format and process them into masks.

2. Comparison with methods with E+I inputs

In this section, additional experimental results on the LLE-VOS and LLE-DAVIS datasets are presented. For a fair comparison, we utilize the event and image modalities as input for recent state-of-the-art methods. First, we use an event encoder and image encoder to extract the features. Then we concatenate these two features as a query to compute similarity in the memory bank. Finally, the matching features are sent to the decoder to generate masks.

Tab. 1 provides a quantitative comparison of our VOS method against existing state-of-the-art methods on the real-world LLE-VOS dataset. The results are split into indoor and outdoor scenes, along with an overall score. From these results, it is clear that outdoor scenes present more of a challenge, generally yielding lower $\mathcal{J}$ and $\mathcal{F}$ scores across all methods. Moreover, the addition of events shows a marked

Method	Input	LLE-DAVIS		
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
STCN [NIPS2021] [2]	I	0.424	0.453	0.438
XMem [ECCV2022] [1]	I	0.465	0.477	0.471
AOT [NIPS2021] [6]	I	0.540	0.578	0.559
DeAOT [NIPS2022] [5]	I	0.541	0.571	0.556
STCN [NIPS2021] [2]	E+I	0.450	0.498	0.474
XMem [ECCV2022] [1]	E+I	0.507	0.534	0.521
AOT [NIPS2021] [6]	E+I	0.555	0.614	0.584
DeAOT [NIPS2022] [5]	E+I	0.566	0.608	0.587
Ours	E+I	**0.602**	**0.654**	**0.628**

Table 2. Quantitative comparisons of various VOS methods on the synthetic LLE-DAVIS dataset. 'I' and 'E' represent the input of images and events, respectively. The best results are marked in bold.

improvement in performance over image-only inputs. This suggests that event data provides valuable information that significantly helps VOS, especially in Indoor Scenes. Our method demonstrates this improvement, achieving the highest $\mathcal{J}\&\mathcal{F}$ score of 0.678 overall, which highlights the effectiveness of our approach on the LLE-VOS dataset.

Tab. 2 presents the performance comparison between our method and other state-of-the-art approaches on the synthetic LLE-DAVIS dataset, utilizing various combinations of image and event data. It is observed that integrating images with event data results in improvement over methods that rely solely on images. This enhancement confirms the significant role of events, as they provide crucial supplementary information that helps image data in video object segmentation under low-light conditions. Compared to these methods, our approach achieves the best. Specifically, against the DeAOT method, our approach shows an increase of 0.041 in the $\mathcal{J}\&\mathcal{F}$ score. These results clearly demonstrate the effectiveness of our method in enhancing VOS performance under challenging lighting conditions.

3. Real-world Dataset Showcase

We show more scenes in our LLE-VOS dataset, shown in Fig. 2. It can be seen that our collected real-world dataset has a diverse array of objects and scenes.

4. More Qualitative Results

Our dataset includes diverse scenes including indoor scenes and outdoor scenes. We show the additional VOS results on LLE-DAVIS and LLE-VOS datasets in Fig. 3. Our method generates the relative complete mask and distinguishes the overlap objects. Besides, our method could segment the invisible and fast objects in the LLE-VOS dataset.

References

[1] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European Conference on Computer Vision (ECCV)*, pages 640–658. Springer, 2022. 2, 4

[2] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021. 2, 4

[3] Gemma Taverni, Diederik Paul Moeys, Chenghan Li, Celso Cavaco, Vasyl Motsnyi, David San Segundo Bello, and Tobi Delbruck. Front and back illuminated dynamic and active pixel vision sensors comparison. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 65(5):677–681, 2018. 1

[4] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2022. Open source software available from https://github.com/heartexlabs/label-studio. 3

[5] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. *Advances in Neural Information Processing Systems*, 35:36324–36336, 2022. 2, 4

[6] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34:2491–2502, 2021. 2, 4