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

Abstract

In the realm of video object segmentation (VOS), the challenge of operating under low-light conditions persists, resulting in notably degraded image quality and compromised accuracy when comparing query and memory frames for similarity computation. Event cameras, characterized by their high dynamic range and ability to capture motion information of objects, offer promise in enhancing object visibility and aiding VOS methods under such low-light conditions. This paper introduces a pioneering framework tailored for low-light VOS, leveraging event camera data to elevate segmentation accuracy. Our approach hinges on two pivotal components: the Adaptive Cross-Modal Fusion (ACMF) module, aimed at extracting pertinent features while fusing image and event modalities to mitigate noise interference, and the Event-Guided Memory Matching (EGMM) module, designed to rectify the issue of inaccurate matching prevalent in low-light settings. Additionally, we present the creation of a synthetic LLE-DAVIS dataset and the curation of a real-world LLE-VOS dataset, encompassing frames and events. Experimental evaluations corroborate the efficacy of our method across both datasets, affirming its effectiveness in low-light scenarios. The datasets are available at https://github.com/HebeiFast/EventLowLightVOS.

1. Introduction

Video Object Segmentation (VOS) refers to a computer vision domain focusing on algorithms for segmenting objects within a sequence of video frames. Its applications span across various fields such as autonomous driving, surveillance, and interactive video editing [29, 37]. VOS is commonly classified into two categories, contingent on the availability of annotation for the initial frame: semi-supervised and unsupervised VOS. Our research primarily delves into the realm of semi-supervised VOS.

Under low-light conditions, insufficient illumination gives rise to a diminished level of detail and compromised color accuracy. Conventional VOS methods [5, 6, 12, 34], predicated upon the availability of high-quality visual inputs, tend to exhibit a marked decline in performance when confronted with such challenging lighting circumstances.

Unlike traditional imaging methods, event cameras signify a fundamental change and offer significant advantages in demanding lighting scenarios [11, 13, 39]. These cameras excel in high dynamic range and provide detailed edge and movement data for objects [2]. These unique features are crucial in segmenting video objects under low-light conditions. Despite their impressive capabilities, event cameras lack the ability of capture texture and color information, which limits their effectiveness in segmentation tasks. Hence, our paper aims to explore the integration of event-based and frame-based modalities to improve VOS, particularly in challenging low-light environments.

In pursuit of event-assisted low-light VOS, three pivotal challenges must be addressed: (i) The absence of a dedicated dataset tailored to low-light conditions for video object segmentation, encompassing both frames and events. Existing datasets are captured in standard lighting, leading to a significant gap in accurately representing low-light situations. (ii) Effectively exploiting complementary information from frame and event modalities under low-light conditions remains a complex task. Traditional approaches to integrate this data fall short due to inherent noise in low-light settings, necessitating the development of a more robust integration strategy for these modalities. (iii) Optimizing the utilization of event assistance for matching poses a substantial challenge. Current methods [5, 32, 33] focus

*Corresponding Author

on enhancing the matching mechanism, assuming normal lighting conditions and relying solely on image data. Consequently, exploring effective means of leveraging event assistance in low-light scenarios becomes essential.

To tackle the above challenges, we build low-light datasets and introduce a novel end-to-end framework designed for low-light VOS, filling a critical research gap in VOS. For low-light VOS data, we construct a synthetic Low-Light Event DAVIS (LLE-DAVIS) dataset and collect a Low-Light Event Video Object Segmentation (LLE-VOS) dataset. These datasets serve as an important foundation for improving VOS techniques for low-light conditions. For our framework, we propose the adaptive cross-modal fusion (ACMF) module to combine complementary information by learning the adaptive filters to select useful information between two modalities. Besides, we propose an event-guided memory matching (EGMM) to solve the inaccurate matching mechanism. Our EGMM utilizes a joint approach of mask and event to guide the network in matching the target areas of memory, thereby enhancing the matching accuracy. We evaluate our method on the synthetic LLE-DAVIS dataset and LLE-VOS dataset. Experiments show our significant effectiveness, setting new standards for performance on both the LLE-DAVIS (62.8%) and LLE-VOS (67.8%) datasets.

In brief, our contributions are summarized as follows:

- We propose the first event-based low-light VOS framework by utilizing the unique properties of event cameras.
- The first real-world event-based low-light VOS dataset is constructed, which contains event streams and frames captured under low-illumination scenarios, clear images and accurate annotations.
- We elaborately design two components, i.e., an adaptive cross-modal fusion module and an event-guided memory matching module, for adaptively fusing the information of both frame and event modalities and enhancing the motion features for the matching module.
- Both quantitative and qualitative results over synthetic and real-world datasets showcase that our proposed method outperforms existing state-of-the-art methods.

2. Related Work

2.1. Video Object Segmentation

In the field of VOS, there are three categories of methods: online fine-tuning-based methods, propagation-based methods, and matching-based methods. Online fine-tuning-based methods [3, 21, 31] concentrate on adjusting pre-existing segmentation networks during the evaluation phase to align them with the specific object being targeted. Propagation-based methods [4, 9, 18, 32, 35] aim to expedite testing time by employing the mask from the previous frame to predict the mask for the current frame. Meanwhile, Matching-based methods, highlighted in [5, 6, 16, 20, 23, 24, 33, 34], ascertain pixel classification by evaluating its resemblance to the target object across memory frames.

2.2. Event Segmentation

There has been a growing interest in tailoring segmentation methods for event cameras, considering their advantageous features. Specifically, in event-based motion segmentation and event-based semantic segmentation. Stoffregen *et al.* [26] proposed a distinctive per-event segmentation method aimed at estimating event-object associations. Mitrokhin *et al.* [17] introduced a graph convolutional neural network to address the challenge of accurately analyzing dynamic scene evolution over time. Zhou *et al.* [40] developed an approach to tackle motion segmentation in event-based camera data by minimizing energy and fitting multiple motion models. On the other hand, Alonso *et al.* [1] presented a novel representation of event data and released a new semantic dataset for event-based semantic segmentation. Additionally, Sun *et al.* [27] proposed an unsupervised domain adaptation method for transferring semantic segmentation tasks from labeled image datasets to unlabeled event data. Moreover, Xia *et al.* [30] introduced an unsupervised cross-domain framework for effective nighttime semantic segmentation. Despite these advancements, a prevailing challenge lies in the inability of these methods to consistently track specific objects throughout entire video sequences. This limitation signifies a significant area necessitating further enhancement and future research endeavors.

2.3. Low-Light Event Application

Numerous studies [13, 14, 25, 38, 39] have delved into the potential of event cameras under low-light conditions. Zhang *et al.* [36] proposed an unsupervised domain adaptation network aimed at reconstructing images captured by event cameras in low-light conditions to resemble those taken during daylight. Jiang *et al.* [10] introduced a framework utilizing event cameras' superior dynamic range to produce clear images in near-darkness by integrating underexposed frames and event streams. Liu *et al.* [14] proposed a novel method for enhancing low-light videos using synthetic events from multiple frames, addressing the artifacts in extreme low light or fast-motion scenarios. Liang *et al.* [13] suggested a video enhancement approach for low-light conditions, establishing spatiotemporal coherence from frame-based and event cameras through a neural network. Zhou *et al.* [39] introduced a two-stage approach to enhance the deblurring of low-light images, leveraging the high dynamic range and low latency of event cameras. Nonetheless, prevailing research on event-based low-light scenarios primarily focuses on foundational tasks, leaving the domain of VOS largely unexplored.

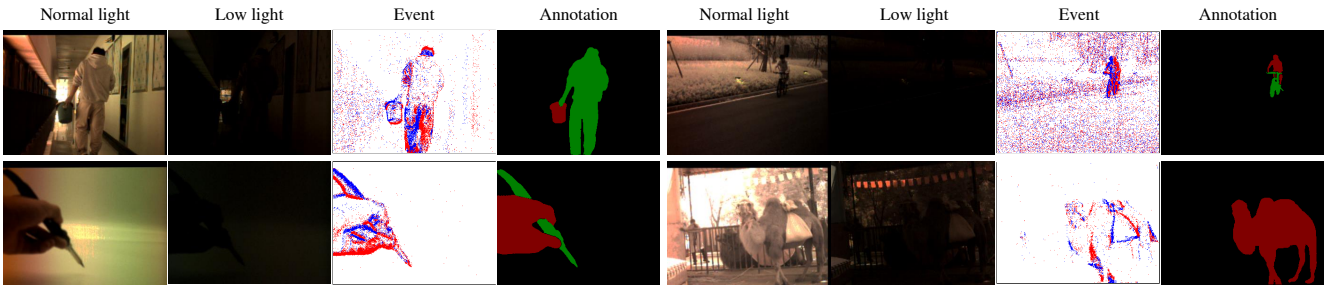

Figure 1. Examples of our LLE-VOS dataset. The dataset contains paired normal/low-light APS images, event stream and annotations.

	Synthetic LLE-DAVIS Dataset			Real-world LLE-VOS Dataset				
	Train	Validation	Total	Train	Validation	Val-Indoor	Val-Outdoor	Total
# Seq.	60	30	90	50	20	10	10	70
# Frm.	4149	1969	6118	3777	1823	993	830	5600
# Evt.	1406.19M	626.70M	2032.89M	575.61M	202.93M	70.83M	132.10M	778.54M
Mean # Frm. per Seq.	69	66	68	76	91	99	83	80
Mean # Evt. per Seq.	23.44M	20.89M	22.59M	11.51M	10.15M	7.08M	13.21M	11.12M
Mean # Objs. per Seq.	2.4	2.0	2.3	1.8	1.6	1.7	1.5	1.7

Table 1. The summary of our synthetic LLE-DAVIS dataset and real-world LLE-VOS dataset, including the number of sequences (#Seq.), frames (#Frm.), events (#Evt.), the mean number of frames, events and objects (#Objs.) per sequence.

(a) (b) (c)

Figure 2. (a) A hybrid camera system for building real-world dataset. We configure two identical cameras with different exposure time for generating normal-light (b) and low-light (c) pairs.

3. Benchmark Dataset

To the best of our knowledge, there has been no event-based VOS dataset in low-light scenarios. In this work, we build two low-light event-based VOS datasets, consisting of a synthetic LLE-DAVIS dataset and a real-world LLE-VOS dataset. Tab. 1 presents the summary of these two datasets, including the number of sequences, frames, objects, and events on each of the dataset splits. Fig. 1 shows some examples in the LLE-VOS dataset.

3.1. Synthetic Dataset

The DAVIS 2017 dataset is a widely used benchmark for the VOS task [5, 6, 12, 19]. Thus based on this dataset, we construct a synthetic event dataset, named LLE-DAVIS, specifically tailored for the Low-Light Event-Based VOS task. Since the original DAVIS 2017 videos are recorded under normal-light conditions, we employ a devised technique [15] to generate low-light videos. Specifically, for a given normal-light frame I_t, we introduce random adjustments to the gamma correction factor and linear scaling factor to synthesize the corresponding low-light frame L_t. This process is mathematically expressed as:

$$L_t = \beta \times (\alpha \times I_t)^{\gamma} + N_{\sigma}, \tag{1}$$

Here, α, β, γ are randomly sampled from the uniform distributions $U(0.9, 1)$, $U(0.5, 1)$ and $U(7, 9)$, respectively. N_{σ} represents Gaussian noise $N(0, \sigma)$, with σ values drawn from a uniform distribution $U(0, 0.05)$. Subsequently, we employ FILM [22] to interpolate DAVIS videos to 100fps. Then, these frames are fed into the ESIM model [7] to generate events. In total, we have assembled 90 low-light video sequences, each accompanied by temporally-synchronized event streams.

3.2. Real-World Dataset

To collect the real-world LLE-VOS dataset, we build a hybrid camera system as shown in Fig. 2(a). The hybrid camera system is equipped with two DAVIS346 event cameras [28], each adept at capturing temporally synchronized APS frames and event streams. A beam splitter is integrated into the system to uniformly distribute incoming light, guaranteeing that both cameras record identical scenes. Additionally, we conduct accurate geometric calibration and temporal synchronization between two cameras. More details are provided in the Supplementary Material. We adjust the exposure times of these two cameras to collect a pair of normal-light and low-light videos in one shot. We employ a group of 20 volunteers to accomplish the annotation work.

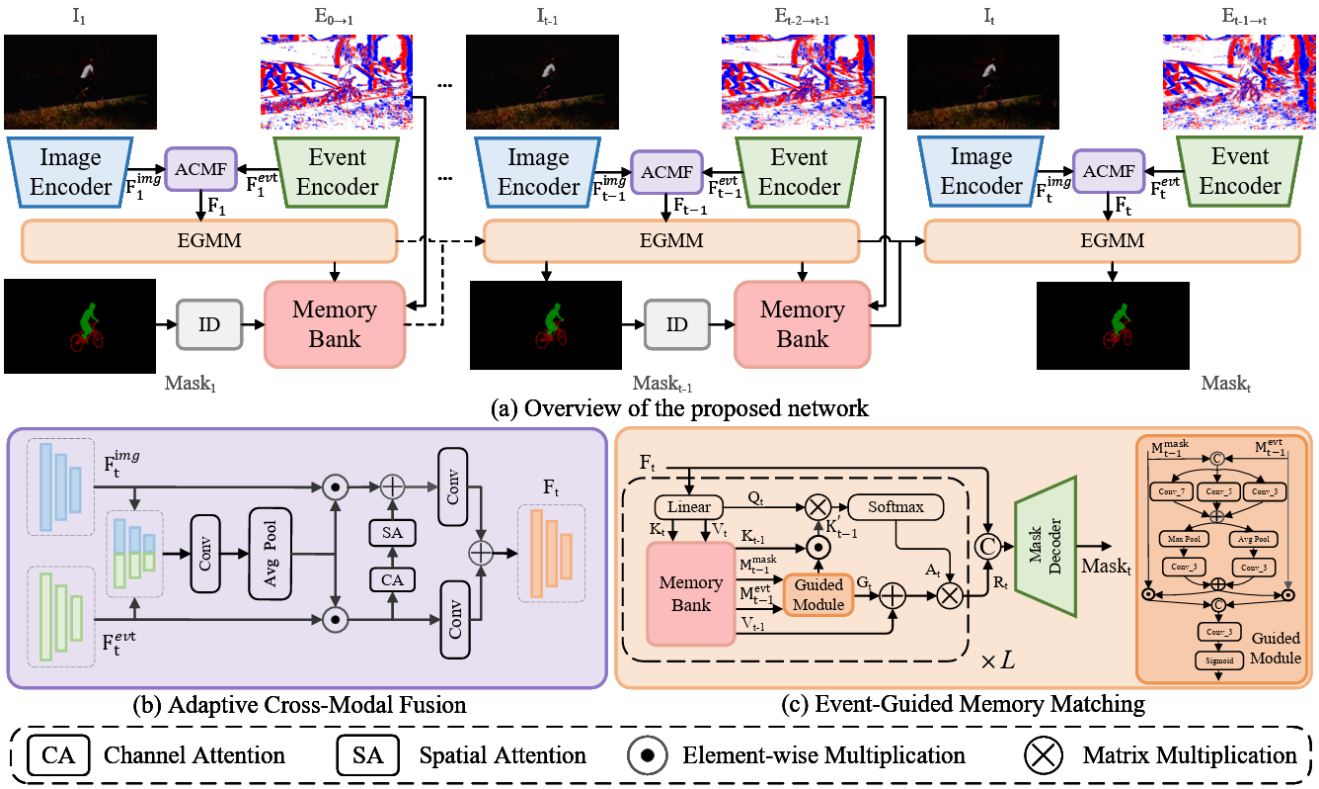

Figure 3. (a) Overview of the proposed method for event-assisted low-light video object segmentation. (b) The structure of Adaptive Cross-Modal Fusion (ACMF) module. (c) The structure of Event-Guided Memory Matching (EGMM) module.

Our final dataset includes 70 videos, consisting of paired normal and low-light videos, along with their corresponding segmentation annotations and event streams. The videos are recorded at a diverse range of locations, including gyms, playgrounds, classrooms, meeting rooms, and zoos. To enhance the robustness of the dataset, it includes varying lighting conditions and contains a rich spectrum of motion information. We randomly select 50 video clips as the training set and 20 as the testing set. To further evaluate the robustness of our proposed method across various scenes, we divide the test set into indoor and outdoor scenarios, with 10 clips each. To our best, this is the first real-world dataset for low-light event-based VOS. It should be noted that this dataset can also be applied to other tasks such as event motion segmentation and low-light video enhancement.

4. Method

4.1. Overview

We propose a novel event-assisted VOS framework, designed specifically for low-light conditions, as depicted in Fig. 3(a). The core components of our architecture comprise an event encoder, an image encoder, an Event-Guided Memory Matching (EGMM) module, an Adaptive Cross-Modal Fusion (ACMF) module, and an Identity Assignment (ID) module [34]. The event encoder and image encoder extract the event feature and image feature, respectively. Then the multi-scale features of events and images are processed by the ACMF module, which generates complementary features. At the first time step, the EGMM module uses these features to produce query Q_1, key K_1 and value V_1. Subsequently, the ID module processes an initial segmentation annotation to create a mask feature. This generated mask feature, along with the predicted mask from the EGMM and the event features, are then systematically archived in the Memory Bank. From the second time step until the final moment, the Memory Bank provides stored representations to the EGMM, along with the representations from image and event encoder. The EGMM then continually generates the mask at each time step, which is preserved in the Memory Bank for subsequent iterative refinement.

4.2. Event Representation

Our proposed framework transforms an asynchronous event streams from $t-1$ to t into corresponding voxel grids, denoted by $E_{t-1\to t} \in \mathbb{R}^{B\times H\times W}$, where B, H and W represent the number of temporal bins, the height and width of the grids, respectively.

4.3. Adaptive Cross-Modal Fusion

We introduce the Adaptive Cross-Modal Fusion (ACMF) to adaptively select the effect information from event and image in low-light environments, which is depicted in Fig. 3(b). Under low-light conditions, the image modality provides inadequate texture and color information of objects while the event modality offers richer edge and motion cues. These distinct attributes of events are advantageous for segmentation tasks.

Initially, ACMF merges the image feature F_t^{img} and event feature F_t^{evt} to create a more comprehensive feature set. This combined feature set is then processed to capture coupled information. The processed features act as a filter that selectively integrates information from two features through element-wise multiplications, reducing noise and enhancing relevant details. Diminished image quality under low-light conditions often results in lost structural details. Therefore, a channel-attention (CA) and a spatial-attention (SA) extract edge information from event feature to complete such details. This information helps to restore invisible object contours. Two convolutional layers further enhance two features. The final output is yielded by summing two features, acting as a robust representation for VOS.

4.4. Event-Guided Memory Matching

Our proposed Event-Guided Memory Matching (EGMM) introduces a novel mechanism designed to enhance the accuracy of object matching between the current feature F_t and the prior feature K_{t-1}, specifically under circumstances of imprecise mask predictions. The EGMM architecture, illustrated in Fig. 3(b), integrates event feature with mask feature from memory to concentrate on motion-specific areas within a scene, consequently refining feature matching.

Initially, the current feature F_t generates the Q_t, K_t, V_t through linear layer. K_t and V_t are saved in the Memory Bank for the next time step. Then, the event feature M_{t-1}^{evt} and mask feature M_{t-1}^{mask} from Memory Bank are sent to the Guided Module, showing in the right of Fig. 3(c). The purpose of the guided branch is to improve unreliable mask predictions. It first integrates the multi-scale information from the concatenated feature through three parallel convolutional layers with various kernel sizes. Then, pooling layers and convolutional layers further reform its channel context, resulting in a guided signal for more accurate matching. Subsequently, the guided signal after passing a Sigmoid function is multiplicated with the past feature K_{t-1}, filtering misaligned features of the object region. The filtering process is described by the equation:

$$G_t = \text{Guide}(M_{t-1}^{\text{evt}}, M_{t-1}^{\text{mask}}) \tag{2}$$

$$K'_{t-1} = K_{t-1} \cdot G_t, \tag{3}$$

where $\text{Guide}(.,.)$ represents the Guided module. $K_{t-1}, M_{t-1}^{\text{evt}}$ and M_{t-1}^{mask} represent the previous key, event and mask features derived from memory, respectively. K'_{t-1} and G are the filtered key and guided branch output, respectively.

The filtered key K'_{t-1} interacts with Q_t derived from F_t through multiplication and softmax, formulated by:

$$A_t = \text{Softmax}\left(\frac{Q_t(K'_{t-1})^T}{\sqrt{d_k}}\right), \tag{4}$$

where A_t is the attention map and d_k is the scaling factor corresponding to the dimension of the key vectors. The matching result R_t is obtained by multiplication between the attention map A_t and the summation of G_t and V_{t-1}:

$$R_t = A_t(G_t + V_{t-1}). \tag{5}$$

Finally, we concat the current feature and matching result to combine the current and memory information. Then the concatenation result is sent to the mask decoder to generate the current mask.

$$Mask_t = Decoder(Concat(R_t, F_t)). \tag{6}$$

This EGMM module thereby provides a robust solution for enhancing the consistency of VOS, proving particularly beneficial in scenarios with unreliable mask predictions.

4.5. Loss function in VOS

To effectively train our VOS framework, we implement a composite loss function, denoted as $\mathcal{L}$. This function combines the Binary Cross-Entropy (BCE) loss and the Soft Jaccard (SJ) loss. The overall loss is formalized as:

$$\mathcal{L} = \sum_{t=1}^{T}\left(\alpha\sum_{o=1}^{N}\mathcal{L}_{\text{BCE}}^{(t,o)} + \beta\sum_{o=1}^{N}\mathcal{L}_{\text{SJ}}^{(t,o)}\right), \tag{7}$$

where $\mathcal{L}_{\text{BCE}}^{(t,o)}$ and $\mathcal{L}_{\text{SJ}}^{(t,o)}$ represent the BCE loss and SJ loss for the object o at time step t, respectively. N denotes the total number of segmented objects. The parameters α and β are the weights attributed to the BCE loss and IoU loss, respectively. T, α, β are set as 5, 0.5 and 0.5, respectively.

5. Experiment

5.1. Comparison Methods

We compare our method with state-of-the-art VOS methods, which include: STCN [6], XMem [5], AOT [34], and DeAOT [33]. In addition to these direct VOS methods, we also compare our method with two-step approachs, which first utilize Zero-DCE [8] for low-light video enhancement and then apply above mentioned methods for VOS.

Following [5, 6], the evaluation of the VOS task employs the $\mathcal{J}$ score for segmentation accuracy, reflecting the average IoU between predictions and ground truth, and the $\mathcal{F}$

Method	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] [6]	0.486	0.321	0.403	0.400	0.309	0.354	0.445	0.316	0.380
XMem [ECCV2022] [5]	0.664	0.528	0.596	0.507	0.456	0.481	0.590	0.494	0.542
AOT [NIPS2021] [34]	0.699	0.618	0.659	0.592	0.571	0.581	0.649	0.596	0.623
DeAOT [NIPS2022] [33]	0.716	0.643	0.680	0.580	0.580	0.580	0.653	0.614	0.633
Zero-DCE [8] + STCN [6]	0.513	0.334	0.424	0.415	0.313	0.364	0.467	0.324	0.396
Zero-DCE [8] + XMem [5]	0.681	0.527	0.604	0.535	0.497	0.516	0.612	0.513	0.563
Zero-DCE [8] + AOT [34]	0.694	0.601	0.647	0.568	0.555	0.562	0.635	0.579	0.607
Zero-DCE [8] + DeAOT [33]	0.648	0.595	0.621	0.594	0.586	0.590	0.622	0.590	0.606
Ours	**0.789**	**0.710**	**0.749**	**0.604**	**0.588**	**0.596**	**0.702**	**0.653**	**0.678**

Table 2. Quantitative comparisons of various VOS methods on the real-world LLE-VOS dataset. The best results are marked in bold.

score for boundary preciseness, comparing the similarity of segmentation edges to actual contours. The mean of these scores denoted as $\mathcal{J}\&\mathcal{F}$, offers a comprehensive and balanced measure of overall performance.

5.2. Implementation Details

For model optimization, we deploy the AdamW optimizer with an initial learning rate of 2×10^{-4} and a weight decay of 0.07. We set our batch size to 8 and train our model over 50,000 iterations. For a fair comparison, we maintain the original training strategies of the models to obtain the best models. We apply standard data augmentation techniques: random scaling, random cropping, random horizontal flipping, and resizing. Besides, we randomly reverse the video and event sequences. The crop size of random cropping is (465, 465) for the LLE-DAVIS dataset and (256, 256) for the LLE-VOS dataset. All models are trained from scratch on both synthetic and real-world datasets under the same settings to ensure equitable comparisons. All the training experiments are conducted on four NVIDIA A800 GPUs.

5.3. Experimental Results

5.3.1 Quantitative Results

Synthetic Dataset. Tab. 3 provides a detailed quantitative analysis of our framework in comparison with other methods on the LLE-DAVIS dataset. The comparative results clearly demonstrate that our end-to-end approach combined with event and image consistently outperforms the existing methods in terms of standard VOS metrics. Notably, our method exhibits a substantial performance increase with an improvement of 6.9% over the AOT method in terms of $\mathcal{J}\&\mathcal{F}$. Additionally, our method significantly improves upon the two-stage method Zero-DCE+AOT, with increments of 6.7% for $\mathcal{J}\&\mathcal{F}$. These improvements are due to the robust feature extraction capabilities of our framework and its effective integration of both image and event data in our VOS framework.

Method	LLE-DAVIS		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
STCN [NIPS2021] [6]	0.424	0.453	0.438
XMem [ECCV2022] [5]	0.465	0.477	0.471
AOT [NIPS2021] [34]	0.540	0.578	0.559
DeAOT [NIPS2022] [33]	0.541	0.571	0.556
Zero-DCE [8] + STCN [6]	0.440	0.469	0.455
Zero-DCE [8] + XMem [5]	0.494	0.512	0.503
Zero-DCE [8] + AOT [34]	0.544	0.577	0.561
Zero-DCE [8] + DeAOT [33]	0.541	0.577	0.559
Ours	**0.602**	**0.654**	**0.628**

Table 3. Quantitative comparisons of various VOS methods on the synthetic LLE-DAVIS dataset. The best results are marked in bold.

Real-world Dataset. Tab. 2 presents the quantitative results of one-stage and two-stage VOS methods on the LLE-VOS dataset. Given that real-world scenarios often involve extremely low-light conditions, the Zero-DCE method struggles to perform effectively. Therefore, the two-stage method Zero-DCE+AOT does not perform better than the AOT method. However, our method can improve the $\mathcal{J}\&\mathcal{F}$ metric over the existing one-stage VOS methods, with an increment when compared to DeAOT. Furthermore, our method outperforms Zero-DCE+AOT. In particular, it increases the $\mathcal{J}\&\mathcal{F}$ score from 0.607 to 0.678 in comparison to Zero-DCE+AOT. Our approach is measured against other one-stage and two-stage methods.

5.4. Qualitative Results

Fig. 4 showcases a qualitative comparison of our proposed method against both one-stage and two-stage methods on the synthetic LLE-DAVIS dataset under low-light conditions. A key observation is that our method produces precise object masks that closely match the groundtruth. Instead, AOT either misses details or overlays redundant masks onto areas without objects. The Zero-DCE+AOT

Figure 4. Qualitative comparisons with other methods on the synthetic LLE-DAVIS dataset.

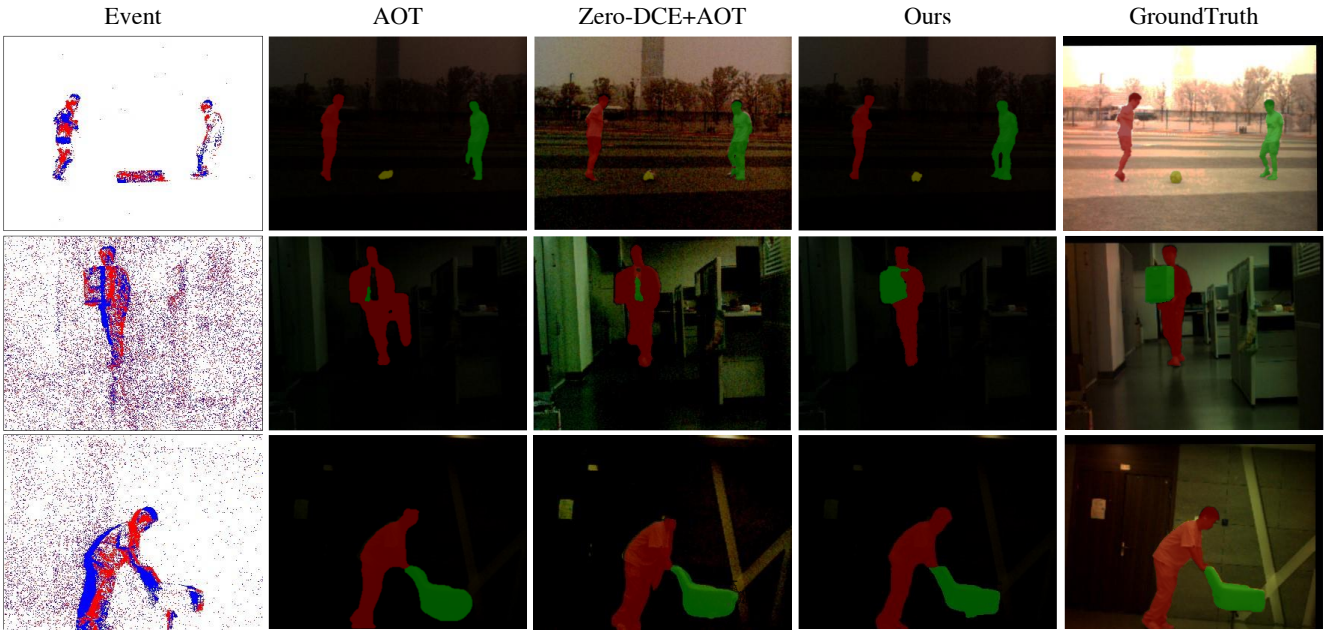

Figure 5. Qualitative comparisons with other methods on the real-world LLE-VOS dataset.

method offers enhancements over AOT alone, yet it fails to achieve the same level of clarity and precision provided by our method. These examples illustrate our method's superior ability to distinguish and outline objects in conditions where light is limited, demonstrating the practical advantage of integrating image and event data. This integration proves especially beneficial in difficult lighting, ensuring our VOS framework remains effective and reliable.

Real-world Dataset. Fig. 5 illustrates qualitative results between our method and other approaches AOT and Zero-DCE+AOT on the real-world LLE-VOS dataset. The results of AOT method often miss parts of the objects, but our method is much better at finding the full shape, even when objects are hard to see in the frame. Our method stands out against Zero-DCE+AOT by being more precise and not mixing up different objects. This capability demonstrates that our approach is not only effective in simulated conditions but also maintains its reliability on real-world data, proving its practical applicability in low-light conditions.

5.5. Ablation Study

Impact of input modalities. In the first part of Tab. 4, we show the comparative performance of the VOS network with different inputs. Utilizing only the image modality as input, the network achieves a $\mathcal{J}\&\mathcal{F}$ score of 0.559. In contrast, relying solely on the event modality yields a

	Method	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Fwk Input	Image Only	0.540	0.578	0.559
	Event Only	0.532	0.563	0.547
	Image + Event	**0.602**	**0.654**	**0.628**
Fwk Mod.	Baseline	0.555	0.614	0.584
	+ ACMF	0.578	0.623	0.601
	+ ACMF + EGMM	**0.602**	**0.654**	**0.628**
EGMM Input	None	0.578	0.623	0.601
	Mask	0.461	0.468	0.464
	Event	0.540	0.563	0.552
	Event + Mask	**0.602**	**0.654**	**0.628**

Table 4. Ablation results of different configurations on the VOS task. 'Fwk Input', 'Fwk Module' and 'EGMM Input' represent the input type of framework input, framwork module and the input type of EGMM module, respectively. The best results are marked in bold.

L	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	FPS
1	0.577	0.628	0.603	22.50
2	0.595	0.642	0.618	21.32
3	**0.602**	**0.654**	**0.628**	20.26
4	0.594	0.649	0.622	19.16

Table 5. Ablation results of different numbers of EGMM blocks on the LLE-DAVIS dataset. The best results are marked in bold.

$\mathcal{J}\&\mathcal{F}$ score of 0.547. These results are inferior compared to the combined input model, which attains a $\mathcal{J}\&\mathcal{F}$ score of 0.628. This marked improvement through fusing both image and event modalities validates the necessity of multi-modality fusion for VOS under low-light conditions.

The effectiveness of ACMF and EGMM. The second part of Tab. 4 validates the effectiveness of the ACMF and EGMM modules within our framework. Initially, our baseline method employs a straightforward fusion approach by concatenating event and image features. The integration of the ACMF module enhances performance, improving the $\mathcal{J}$ score to 0.578, the $\mathcal{F}$ score to 0.623, and the composite $\mathcal{J}\&\mathcal{F}$ score to 0.601. Incorporating both ACMF and EGMM modules further boosts the $\mathcal{J}\&\mathcal{F}$ score to 0.628, indicating a significant contribution to the VOS performance on the synthetic LLE-DAVIS dataset.

Fig. 6 presents a qualitative assessment of the individual and combined impacts of ACMF and EGMM. The baseline method exhibits inaccuracies in contour delineation and mask separation. The ACMF module refines the contour accuracy for the yellow human. However, the network still struggles to differentiate between the green human and the vehicle. The integration of the EGMM module effectively enables the network to distinguish between car and human.

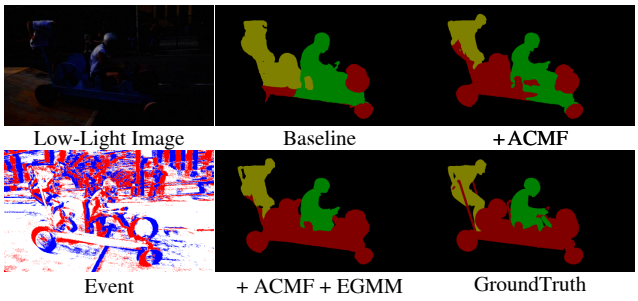

Figure 6. Visual results of ablation on different modules.

The effect of event and mask in EGMM. To evaluate the individual and combined contributions of event and mask priors within the EGMM module, we present a quantitative analysis in the third part of Tab. 4. Without EGMM, the $\mathcal{J}\&\mathcal{F}$ score is 0.601. Utilizing mask information alone yields a $\mathcal{J}\&\mathcal{F}$ score of 0.464 due to imprecise prediction to disturb matching. When the model processes only event data, the $\mathcal{J}\&\mathcal{F}$ is 0.552 due to noise and background motion. However, the fusion of both event and mask priors significantly enhances model performance, as evidenced by the improved $\mathcal{J}\&\mathcal{F}$ of 0.628. These results clearly illustrate the collaborate effect of combining mask and event information, which leads to a more robust and accurate performance in the EGMM.

The choice of EGMM block number. Tab. 5 examines the impact of changing the number of EGMM blocks on our model's performance. As we increase the number of EGMM blocks, the performance of our model improves, but it leads to more redundancy. When the number of EGMM blocks is four, the $\mathcal{J}\&\mathcal{F}$ score decreases. Therefore, we use three EGMM blocks in our experimental setup.

6. Conclusion

In this paper, we present a new approach to Video Object Segmentation (VOS) in low-light conditions. Unlike traditional VOS methods that rely heavily on high-quality video, our innovative framework uses the unique features of event cameras to improve segmentation accuracy. Our introduction of Adaptive Cross-Modal Fusion and Event-Guided Memory Matching has notably enhanced VOS performance. Our thorough testing on specially created Low-Light Event DAVIS (LLE-DAVIS) and Low-Light Event Video Object Segmentation (LLE-VOS) datasets proves that our method is superior, setting new standards for low-light VOS. This study paves the way for further breakthroughs in VOS and related fields.

7. Acknowledgments

This work was in part supported by the National Natural Science Foundation of China (NSFC) under grants 62032006 and 62021001.

References

[1] Inigo Alonso and Ana C Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 2

[2] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240 $\times$ 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014. 1

[3] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 221–230, 2017. 2

[4] Xi Chen, Zuoxin Li, Ye Yuan, Gang Yu, Jianxin Shen, and Donglian Qi. State-aware tracker for real-time video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9384–9393, 2020. 2

[5] 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. 1, 2, 3, 5, 6

[6] 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. 1, 2, 3, 5, 6

[7] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020. 3

[8] Chunle Guo Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1780–1789, 2020. 5, 6

[9] Varun Jampani, Raghudeep Gadde, and Peter V Gehler. Video propagation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 451–461, 2017. 2

[10] Yu Jiang, Yuehang Wang, Siqi Li, Yongji Zhang, Minghao Zhao, and Yue Gao. Event-based low-illumination image enhancement. *IEEE Transactions on Multimedia*, pages 1–12, 2023. 2

[11] Yu Jiang, Yuehang Wang, Siqi Li, Yongji Zhang, Minghao Zhao, and Yue Gao. Event-based low-illumination image enhancement. *IEEE Transactions on Multimedia*, 26:1920–1931, 2023. 1

[12] Mingxing Li, Li Hu, Zhiwei Xiong, Bang Zhang, Pan Pan, and Dong Liu. Recurrent dynamic embedding for video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1332–1341, 2022. 1, 3

[13] Jinxiu Liang, Yixin Yang, Boyu Li, Peiqi Duan, Yong Xu, and Boxin Shi. Coherent event guided low-light video enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10615–10625, 2023. 1, 2

[14] Lin Liu, Junfeng An, Jianzhuang Liu, Shanxin Yuan, Xiangyu Chen, Wengang Zhou, Houqiang Li, Yan Feng Wang, and Qi Tian. Low-light video enhancement with synthetic event guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1692–1700, 2023. 2

[15] Feifan Lv, Yu Li, and Feng Lu. Attention guided low-light image enhancement with a large scale low-light simulation dataset. *International Journal of Computer Vision*, 129(7): 2175–2193, 2021. 3

[16] Yunyao Mao, Ning Wang, Wengang Zhou, and Houqiang Li. Joint inductive and transductive learning for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9670–9679, 2021. 2

[17] Anton Mitrokhin, Zhiyuan Hua, Cornelia Fermuller, and Yiannis Aloimonos. Learning visual motion segmentation using event surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14414–14423, 2020. 2

[18] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7376–7385, 2018. 2

[19] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3

[20] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9226–9235, 2019. 2

[21] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2663–2672, 2017. 2

[22] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *European Conference on Computer Vision*, pages 250–266. Springer, 2022. 3

[23] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 629–645. Springer, 2020. 2

[24] Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. Hierarchical memory matching network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12889–12898, 2021. 2

[25] Peilun Shi, Jiachuan Peng, Jianing Qiu, Xinwei Ju, Frank Po Wen Lo, and Benny Lo. Even: An event-based framework

for monocular depth estimation at adverse night conditions. *arXiv preprint arXiv:2302.03860*, 2023. 2

[26] Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, and Davide Scaramuzza. Event-based motion segmentation by motion compensation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7244–7253, 2019. 2

[27] Zhaoning Sun, Nico Messikommer, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. In *European Conference on Computer Vision (ECCV)*, pages 341–357. Springer, 2022. 2

[28] 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. 3

[29] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Selective video object cutout. *IEEE Transactions on Image Processing*, 26(12):5645–5655, 2017. 1

[30] Ruihao Xia, Chaoqiang Zhao, Meng Zheng, Ziyan Wu, Qiyu Sun, and Yang Tang. Cmda: Cross-modality domain adaptation for nighttime semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21572–21581, 2023. 2

[31] Huaxin Xiao, Jiashi Feng, Guosheng Lin, Yu Liu, and Maojun Zhang. Monet: Deep motion exploitation for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1140–1148, 2018. 2

[32] Xiaohao Xu, Jinglu Wang, Xiao Li, and Yan Lu. Reliable propagation-correction modulation for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2946–2954, 2022. 1, 2

[33] Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. *Advances in Neural Information Processing Systems*, 35:36324–36336, 2022. 1, 2, 5, 6

[34] 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. 1, 2, 4, 5, 6

[35] Lu Zhang, Zhe Lin, Jianming Zhang, Huchuan Lu, and You He. Fast video object segmentation via dynamic targeting network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5582–5591, 2019. 2

[36] Song Zhang, Yu Zhang, Zhe Jiang, Dongqing Zou, Jimmy Ren, and Bin Zhou. Learning to see in the dark with events. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 666–682. Springer, 2020. 2

[37] Ziyu Zhang, Sanja Fidler, and Raquel Urtasun. Instance-level segmentation for autonomous driving with deep densely connected mrfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 669–677, 2016. 1

[38] Chu Zhou, Minggui Teng, Jin Han, Chao Xu, and Boxin Shi. Delieve-net: Deblurring low-light images with light streaks and local events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1155–1164, 2021. 2

[39] Chu Zhou, Minggui Teng, Jin Han, Jinxiu Liang, Chao Xu, Gang Cao, and Boxin Shi. Deblurring low-light images with events. *International Journal of Computer Vision*, 131(5): 1284–1298, 2023. 1, 2

[40] Yi Zhou, Guillermo Gallego, Xiuyuan Lu, Siqi Liu, and Shaojie Shen. Event-based motion segmentation with spatio-temporal graph cuts. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):4868–4880, 2021. 2