Tracking through Severe Occlusion via Event-Derived Transient Cues

Supplementary Material

Summary

The supplementary material is organized as follows:

- Section 1 introduces the details of the FEOT dataset.
- Section 2 discusses more ablation studies of EvoTrack.
- Section 3 shows more tracking results.
- Section 4 discusses the limitation of EvoTrack.

1. The FEOT Dataset

1.1. Data Collection System

We construct a coaxial data collection system comprising a conventional frame camera (FLIR BFS-U3-32S4C, 2048×1536) and an event camera (Prophesee EVK4, 1280×720). A beam splitter (Thorlabs BSW26R) divides the incoming light into two equal parts and directs them to the event camera and the frame camera, respectively. Additionally, we provide external trigger signals to the cameras through a programmable synchronous circuit (STM32 F103ZET6), enabling precise synchronization of the timestamps of both cameras. Finally, we achieve pixel alignment between the two cameras through stereo rectification, as illustrated in Fig. 1. After alignment and cropping, the resulting frame–event pairs achieve a spatial resolution of 1070×610. Based on this setup, we introduce a Frame–Event based Occluded Tracking dataset (FEOT), specifically designed for occlusion scenarios. Compared with existing datasets, FEOT not only provides higher spatial resolution but also includes detailed occlusion labels covering a variety of occlusion types, including static and dynamic occlusions. The FEOT dataset contains 354 video sequences, comprising 72K images and related events.

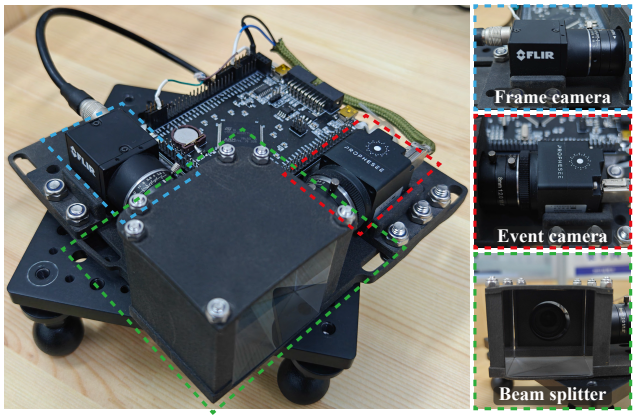


Figure 1. Implementation of the coaxial data-collection system.

1.2. Comparison of Frame-Event Tracking Dataset

Event-based visual tracking has garnered increasing attention in recent years, highlighting the growing importance of event-based benchmark datasets. Tab. 1 shows a comparison of our proposed FEOT dataset against other frame–event tracking datasets. Note that event-only datasets and test-only datasets are not included in this table.

FE108 [1] is a frame–event tracking dataset collected in indoor environments, featuring challenging conditions such as low illumination, high dynamic range, and high-speed nonlinear motion. It contains 108 sequences in total, with 76 used for training and 32 for testing. All data are captured using a DAVIS sensor at a resolution of 346×260.

VisEvent [2] is the first large-scale event-based VOT dataset and remains the most widely adopted frame–event benchmark. It comprises 820 sequences (500 for training and 320 for testing) spanning 17 challenging attributes, captured with a DAVIS camera at 346×260 resolution.

COESOT [3] is a large-scale VOT dataset designed for general-purpose scenarios. It contains 1,354 sequences covering 90 object categories and 17 challenging attributes, with 827 sequences for training and 527 for testing. All sequences are recorded using a DAVIS sensor with a resolution of 346×260.

FELT [4] is the first frame–event dataset tailored for long-term tracking. It includes 742 sequences with 14 challenging attributes, of which 520 are used for training and 220 for testing. The dataset is captured with a DAVIS camera at a resolution of 346×260.

CRSOT [5] is the first high-resolution frame–event tracking dataset; however, the two modalities are not spatially aligned. The images have a resolution of 1440×1080, while the event data are 1280×800. It contains 1,030 sequences (836 for training and 194 for testing). Despite its high spatial resolution, misalignment between modalities makes effective feature fusion particularly challenging.

FEOT is the first pixel-aligned, high-resolution frame–event dataset explicitly designed for evaluating tracking under occlusion. It provides 11 occlusion levels covering diverse types, including dynamic/static and hard/soft occlusions. FEOT contains 354 sequences and is used solely for evaluation, without a predefined train/test split, to benchmark tracker robustness under occlusion.

1.3. Challenging Attribute Definition

In constructing the FEOT dataset, we define ten challenging attributes to comprehensively assess state-of-the-art trackers. These attributes are categorized into three levels:

Table 1. Comparison of existing frame-event datasets for visual object tracking.

Datasets	Year	Sequences	Resolution	Alignment	Occlusion Rate	Occlusion GT	Occlusion Level	Focus
FE108 [1]	2021	108	346×260	✓	×	×	×	Indoor
VisEvent [2]	2021	820	346×260	✓	9%	×	×	Large-scale
COESOT [3]	2022	1354	346×260	✓	2%	×	×	General
FELT [4]	2024	742	346×260	✓	40%	×	×	Long-term
CRSOT [5]	2024	1030	1280×800	×	9%	×	×	Cross-resolution
FEOT	2025	354	1070×610	✓	90%	✓	11	Occlusion

Table 2. Attributes and corresponding description.

Level	Attribute	Description
Environment	OE	*Overexposure*. The illumination is too high to distinguish the target.
	LI	*Low Illumination*. The videos are recorded in the dark environment.
	RN	*Rain*. Rainy weather conditions.
Platform	BI	*Background Influence*. The target is heavily influenced by the background.
	VC	*Viewpoint Change*. The viewpoint is not fixed because the capturing angle changes.
Target	OC	*Occlusion*. The target is partially or fully occluded.
	FM	*Fast Motion*. The target moves quickly.
	NM	*No Motion*. Target is stationary.
	SV	*Scale Variation*. Ratio change of target size between minimum and maximum is more than 50%.
	OV	*Out of View*. The target partially or fully leaves the camera view.

Table 3. Experiments under occlusion and nonlinear motion.

Module	Ratio (%)			Duration (frames)			Nonlinear
	[0,10]	[40,50]	[80,90]	[1,50]	[200,250]	[400,500]	acc./dec./turn.
EvoTrack	71.6	49.7	27.3	53.6	32.8	28.5	60.2
EMA only	54.0	38.3	25.1	41.0	29.2	27.6	51.3
TAM only	69.2	43.5	18.8	50.5	27.6	24.9	48.4

Table 4. Generalization experiments on public datasets.

Method	VisEvent PR / SR (%)	VisEvent-Occ PR / SR (%)	COESOT PR / SR (%)	COESOT-Occ PR / SR (%)
SeqTrack	76.9 / 60.7	74.5 / 58.4	82.2 / 71.8	81.7 / 66.0
SeqTrackV2	79.4 / **63.0**	76.2 / 60.2	85.0 / 75.9	82.8 / 71.1
EvoTrack	**80.1** / 62.1	**78.5 / 61.3**	**85.4 / 76.2**	**83.9 / 73.4**

Table 5. Ablation study of the sensor frame rate (10 test videos).

Temporal sampling	Baseline PR / SR (%)	w/o bidirectional consistency PR / SR (%)	w/o trajectory token PR / SR (%)
Lower (12.5 fps)	79.9 / 67.6	74.8 / 64.6	77.9 / 61.7
Normal (25 fps)	89.8 / 74.0	79.6 / 65.8	84.4 / 70.0
Higher (50 fps)	**93.2 / 78.7**	**81.2 / 69.4**	**86.6 / 75.3**

environment, platform, and target, each reflecting variations that trackers must robustly handle in real-world scenarios. Specifically, the attributes include Overexposure (OE), Low Illumination (LI), Rain (RN), Background Influence (BI), Viewpoint Change (VC), Occlusion (OC), Fast Motion (FM), No Motion (NM), Scale Variation (SV), and Out of View (OV), as summarized in Tab. 2. Each video sequence in the dataset is annotated with multiple challenge attributes. By incorporating such fine-grained annotations, FEOT provides a more comprehensive benchmark for evaluating the generalization ability of future trackers.

2. Ablation Study and Discussion

2.1. Fine-grained Analysis of the EMA

Tab. 3 reports EMA performance (SR metric) under fine-grained challenging conditions. EvoTrack and TAM results are included for reference, demonstrating the nonlinear motion modeling of the EMA. Failure cases will be provided in the final version.

2.2. Cross-dataset Generalization under Occlusion

FEOT is an occlusion-focused dataset, where EvoTrack clearly outperforms other methods, demonstrating occlusion robustness. We specifically select occlusion segments from the public datasets to eliminate dataset-specific risks. In Tab. 4, the performance gap between EvoTrack and other methods is larger on occlusion-only subsets than on original datasets, demonstrating its generalization and robustness.

2.3. Contribution Analysis of Sensor *vs* Algorithm

We break down performance gains in Tab. 5. Higher frame rates markedly improve accuracy due to "slower" appearance variation under partial occlusion. Bidirectional consistency supervision contributes more than the trajectory token, since transient cues better capture non-linear motion.

2.4. Comparison of Event Representations

Most existing event-based tracking methods rely on event frame representations to extract target features. While these approaches have demonstrated notable performance gains, they largely overlook the intrinsic timestamp information carried by events. Each event provides pixel-level temporal cues that encode the temporal firing order, which in turn implicitly reveals the target's motion direction. To characterize such temporal cues, current time-surface (TS) representations [11, 12] utilize an exponential kernel to emphasize differences in event timestamps, as follows:

$$TS(i,j) = \exp(-\frac{t_{ref} - t(i,j)}{\tau}), \tag{1}$$

where $TS(i,j)$ denotes the pixel value at coordinates (i,j), t_{ref} is the reference time, and τ is the predefined decay constant. This representation explicitly encodes the temporal evolution of events. However, its effectiveness **heavily depends on the choice of** τ. An overly large or

Table 6. Effectiveness of event cues on tracking performance. SR/PR are reported on FE108.

Module	w/o event	w/ event frame	w/ TS ($\tau = 5e4$)	w/ forward TS	w/ forward & backward TS
SR (%)	64.2	65.4	66.1	66.9	68.4
PR (%)	91.5	91.7	92.0	92.1	94.6

Table 7. Comparison of model parameters, speed, and tracking accuracy with SOTA frame-event trackers. SR is reported on FE108.

Method	MixFormer [6]	ORTrack [7]	ARTrack [8]	SeqTrack [9]	ViPT [10]	EvoTrack
Params (M)	97.2	8.0	173.1	89.1	93.0	189.1
Macs (G)	30.4	2.4	38.0	65.9	21.9	84.6
FPS	79.0	95.2	19.6	45.4	35.4	19.5
SR (%)	51.6	38.8	49.9	55.4	65.8	68.4

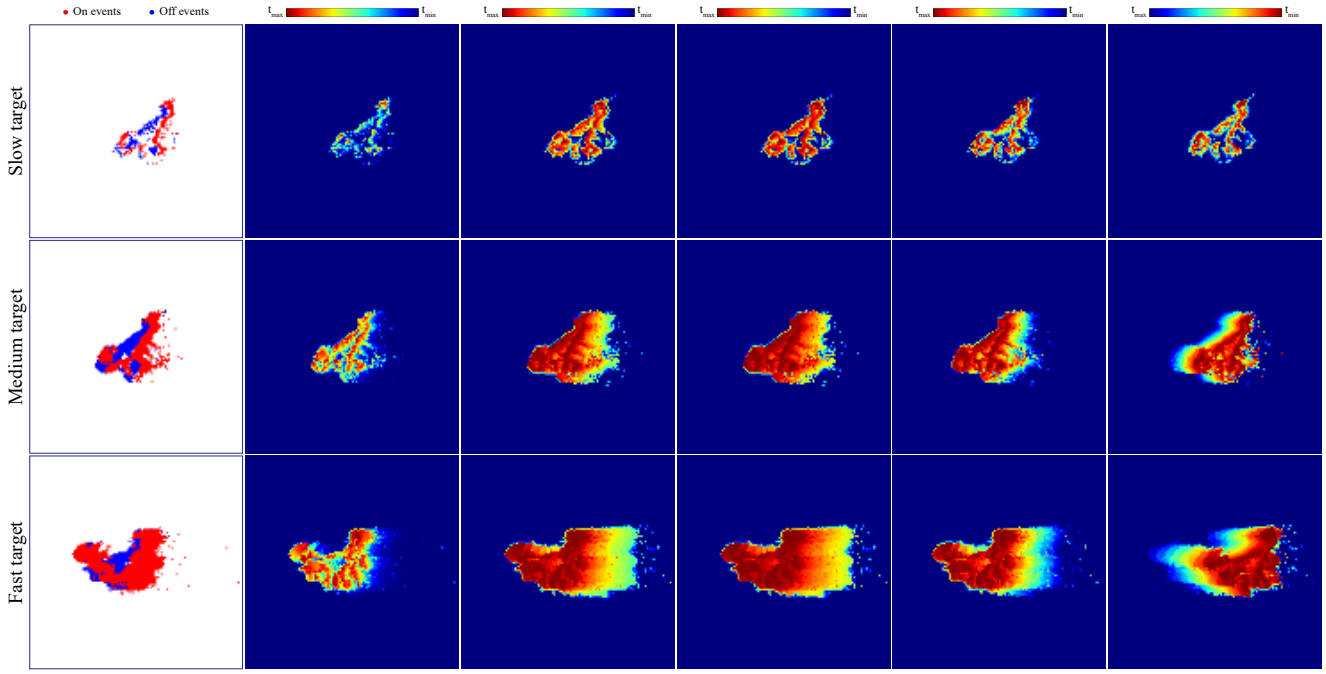


Figure 2. Visualization of different event representations for slow, medium, and fast targets. The columns show event frames, exponential-kernel time-surfaces ($\tau = 5e3/5e4/7e4$), and the proposed forward and backward time-surfaces (FTS / BTS), respectively.

small τ degrades the ability to indicate motion direction, as illustrated in Fig. 2(b–d). Moreover, the use of the exponential function makes it non-trivial to construct a backward time-surface based on the Eq. (1). To address this, we adopt a simple linear time-surface representation to encode the target motion state:

$$I_f(i,j) = \max\{t_e \mid e \in \xi^*, x_e = i, y_e = j\}, \quad (2)$$

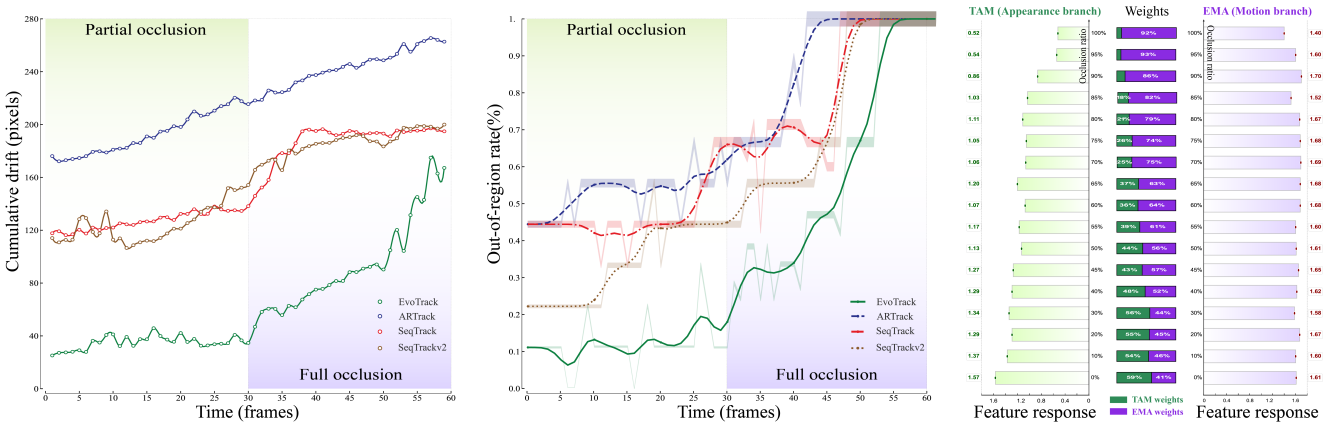
$$FTS(i,j) = H(I_f(i,j)), \quad (3)$$

where $I_f(i,j)$ denotes the latest event timestamp at pixel (i,j), $e = (x_e, y_e, p_e, t_e)$ is an event, and pixels without events are assigned zero. A histogram equalization transformation [13] $H(*)$ is then applied to the time image I_f to produce the forward time-surface (FTS). Specifically, we first compute the statistics of event timestamps and construct a histogram $h(r)$, where $h(r)$ denotes the number of events with timestamp r. Then, the cumulative distribution function $F(*)$ is defined as:

$$F(r) = \sum_{i=0}^{r} h(i), \quad (4)$$

where $F(r)$ represents the number of events whose timestamps are less than or equal to r. Subsequently, a histogram equalization function $H(*)$ is constructed as:

$$H(r) = (L-1)\frac{F(r) - F_{\min}}{F_{\max} - F_{\min}}, \quad (5)$$



(a) Drift accumulation curves (b) Out-of-region rate curves (c) Feature reponse and fusion gate weights

Figure 3. More experiments of EvoTrack.

where $H(r)$ denotes the reassigned timestamp for events with original timestamp r, $L = 256$ indicates that the original timestamp range is $[0, 255]$.

The above formulation avoids the dependency on the hyperparameter τ and provides a linear indication of motion direction. The backward time-surface (BTS) can be constructed by inverting the forward time-surface, as follows:

$$I_b(i,j) = \min\{t_e \mid e \in \xi^*,\, x_e = i,\, y_e = j\}, \quad (6)$$

$$BTS(i,j) = H(255 - I_b(i,j)). \quad (7)$$

FTS and BTS provide two opposite temporal views of events occurring within the same interval, offering an intrinsic motion consistency signal, as illustrated in Fig. 2(e-f).

2.5. Discussion of Events on Tracking Performance

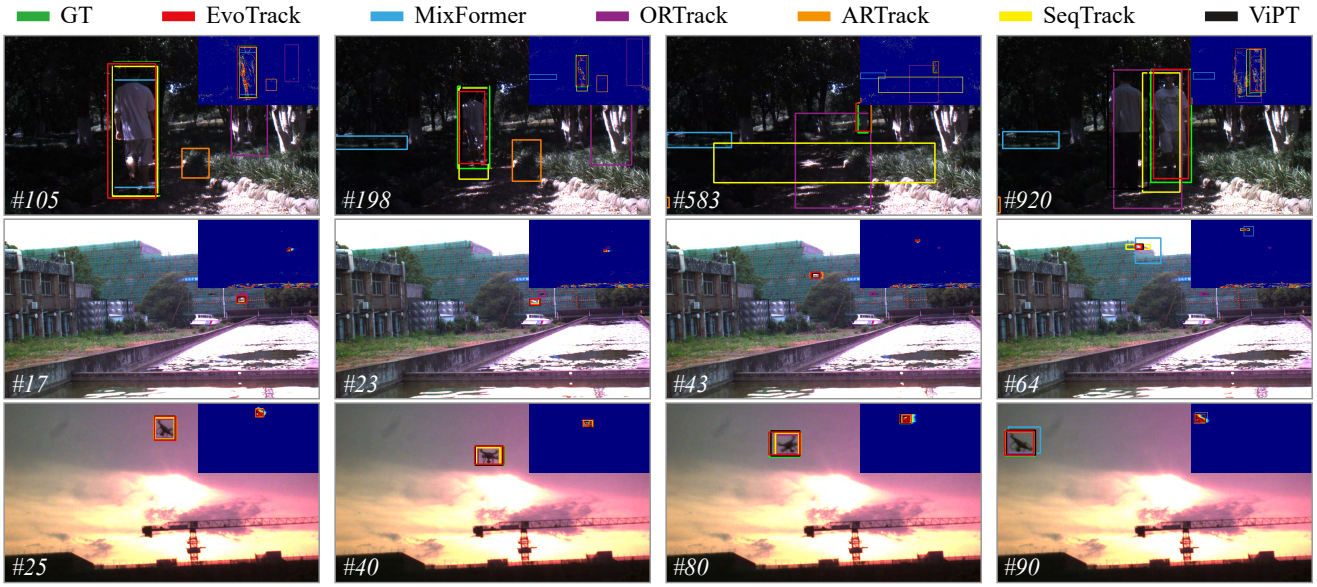
To assess the contribution of event cues to motion modeling, we conduct a series of ablation studies by progressively injecting different event representations into our motion module, as shown in Tab. 6. Using only frame information yields the lowest performance, as the tracker can merely rely on appearance cues and the global trajectory, making it vulnerable to rapid or nonlinear displacements. Introducing event frames brings a modest improvement, indicating that even coarse event aggregation already supplies beneficial local motion hints. More importantly, leveraging exponential-kernel time-surface (TS) provides a clear and consistent boost. The TS offers stronger guidance by explicitly encoding instantaneous motion direction, which is crucial for resolving ambiguous target locations under occlusion. Replacing the exponential-kernel TS with a linear forward TS further enhances performance, suggesting that building a linear correlation between the temporal evolution of events and target motion helps model the motion more accurately. The best results emerge when combining forward and backward TS, demonstrating that bidirectional motion consistency offers a more complete transient-motion description. These results demonstrate that transient motion cues from events are essential for robust prediction of nonlinear target motion, especially under appearance degradation caused by occlusion.

2.6. Comparison of Tracking Speed

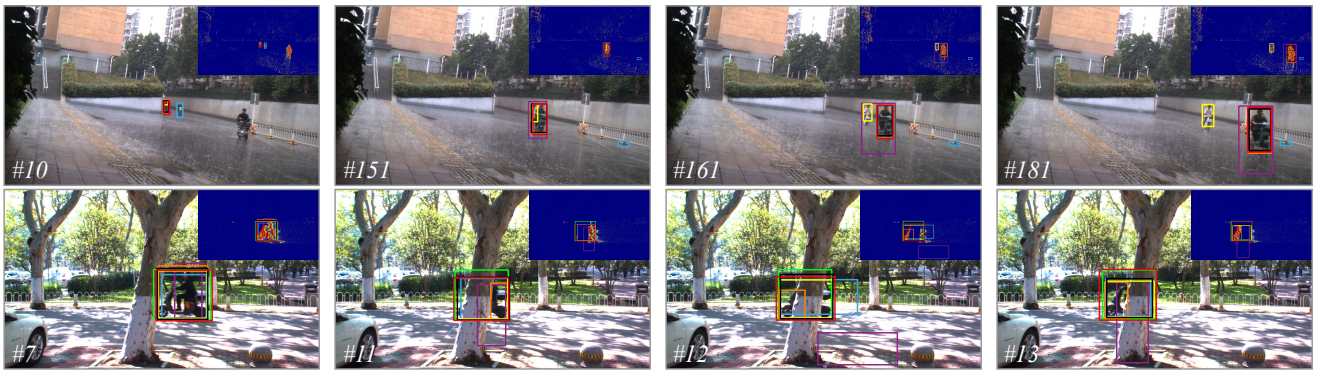
Tab. 7 presents a comprehensive comparison between EvoTrack and other representative tracking methods in terms of model parameters, computational cost, runtime efficiency, and overall tracking accuracy. All evaluations are performed on an NVIDIA RTX 3090 GPU platform. Across these dimensions, lightweight trackers maintain low computational overhead but consistently underperform in terms of robustness. In contrast, other heavier trackers offer stronger representational capacity, yet typically suffer from substantial computational cost and slower inference speed. EvoTrack belongs to this latter category in scale, but it clearly stands out by achieving markedly superior tracking performance compared with other high-capacity baselines. This suggests that the gains of EvoTrack stem from its unified motion–appearance design rather than mere increases in model size. Overall, the table illustrates that EvoTrack strikes a favorable balance between accuracy and model complexity: while not the smallest or fastest, it delivers the most reliable performance, validating the effectiveness of the proposed framework and its practical advantages in challenging occlusion scenarios.

2.7. Discussion of EvoTrack under Full Occlusion

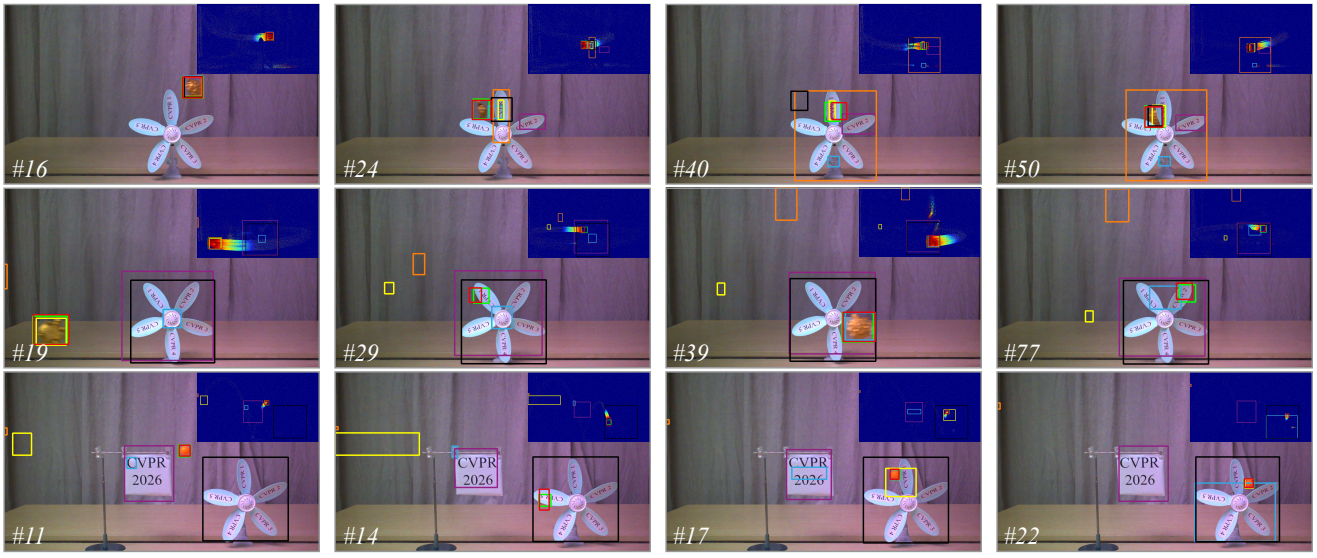
Under consecutive full occlusion, motion autoregression is unavailable and does not directly contribute. Instead, temporally dense events capture richer motion cues under partial occlusion, which reduces prediction drift during subsequent full occlusion. Consequently, events are beneficial under **partial** and **short-term full** occlusion. In Fig. 3(a), EvoTrack shows markedly lower drift than baselines and remains stable after brief full occlusion. In Fig. 3(b), it



(a) Tracking results on moving targets in non-occluded scenes



(b) Tracking results on linearly moving targets under occlusion



(c) Tracking results on nonlinearly moving targets under occlusion

Figure 4. Qualitative comparison of EvoTrack against state-of-the-art trackers across different motion patterns and occlusion scenarios.

drifts out of the search region on all videos around frame 55 (25 fully occluded frames), and survives longer than AR-based baselines. Additionally, Fig. 3(c) further shows EMA/TAM feature responses and fusion weights versus occlusion severity: TAM dominates under no/low occlusion, whereas EMA under severe occlusion. The above experiments demonstrate the effectiveness of EvoTrack in occlusion scenarios.

3. Additional Results

To further examine the performance of different trackers under diverse motion and occlusion conditions, Fig. 4 presents qualitative comparisons across three representative scenarios: (a) moving targets in non-occluded scenes, (b) linearly moving targets under occlusion, and (c) nonlinearly moving targets under occlusion. In non-occluded cases, all methods maintain reasonable localization accuracy, while EvoTrack delivers consistently competitive performance, producing more stable bounding boxes around the target. The advantage of EvoTrack becomes more pronounced in the presence of occlusion. For linear motion under partial or heavy occlusion, competing trackers frequently drift toward distractors or fail to recover the target afterward. In contrast, EvoTrack preserves robust spatial consistency and rapidly re-locks the target, demonstrating strong resilience to visual degradation. The superiority is even more evident in nonlinear-motion sequences where rapid trajectory changes and intermittent occlusions co-occur. EvoTrack effectively maintains accurate predictions throughout the occlusion, whereas other methods show significant instability. These results highlight the strong robustness of EvoTrack in challenging occlusion scenarios while maintaining solid performance in regular tracking conditions.

4. Limitation

The proposed method may still struggle under long-term full occlusion. When the target becomes fully hidden, appearance cues vanish completely, forcing the tracker to rely solely on motion-based prediction. Although the method can maintain accurate short-term estimates, the absence of appearance correction inevitably causes motion predictions to drift during prolonged occlusion, ultimately leading to tracking failure.

References

[1] Zhang Jiqing, Yang Xin, Fu Yingkai, Wei Xiaopeng, Yin Baocai, and Dong Bo. Object tracking by jointly exploiting frame and event domain. In *Proc. ICCV*, 2021. 1, 2

[2] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. Visevent: Reliable object tracking via collaboration of frame and event flows. *IEEE Trans. Cybern.*, 54(3):1997–2010, 2024. 1, 2

[3] Tang Chuanming, Wang Xiao, Huang Ju, Jiang Bo, Zhu Lin, Zhang Jianlin, Wang Yaowei, and Tian Yonghong. Revisiting color-event based tracking: A unified network, dataset, and metric. *arXiv preprint*, 2022. 1, 2

[4] Wang Xiao, Lou Xufeng, Wang Shiao, Huang Ju, Chen Lan, and Jiang Bo. Long-term visual object tracking with event cameras: An associative memory augmented tracker and a benchmark dataset. *arXiv preprint*, 2024. 1, 2

[5] Zhu Yabin, Wang Xiao, Li Chenglong, Jiang Bo, Zhu Lin, Huang Zhixiang, Tian Yonghong, and Tang Jin. Crsot: Cross-resolution object tracking using unaligned frame and event cameras. *IEEE Trans. Multimedia*, 27:6529–6542, 2025. 1, 2

[6] Yutao Cui, Cheng Jiang, Gangshan Wu, and Limin Wang. Mixformer: End-to-end tracking with iterative mixed attention. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46:4129–4146, 2024. 3

[7] You Wu, Xucheng Wang, Xiangyang Yang, Mengyuan Liu, Dan Zeng, Hengzhou Ye, and Shuiwang Li. Learning occlusion-robust vision transformers for real-time uav tracking. In *Proc. CVPR*, 2025. 3

[8] Xing Wei, Yifan Bai, Yongchao Zheng, Dahu Shi, and Yihong Gong. Autoregressive visual tracking. In *Proc. CVPR*, 2023. 3

[9] Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *Proc. CVPR*, 2023. 3

[10] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proc. CVPR*, 2023. 3

[11] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(7):1346–1359, 2016. 2

[12] Ninghui Xu, Lihui Wang, Zhiting Yao, and Takayuki Okatani. Mets: Motion-encoded time-surface for event-based high-speed pose tracking. *Int. J. Comput. Vis.*, 133(7):4401–4419, 2025. 2

[13] WO Saxton, TJf Pitt, and M Horner. Digital image processing: the semper system. *Ultramicroscopy*, 4(3):343–353, 1979. 3