Tracking through Severe Occlusion via Event-Derived Transient Cues

Hao Dong, Yujin Liu, Haoyue Liu, Zhenyu Wang, Shihan Peng, Zhiwei Shi, Yi Chang*, Luxin Yan
State Key Laboratory of Multispectral Information Intelligent Processing Technology
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
{donghao0205, yujinliu, liuhy, yichang, yanluxin}@hust.edu.cn

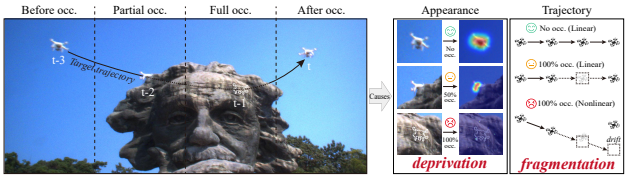
Abstract

Tracking targets with high-speed and nonlinear motion under occlusion remains challenging due to spatial appearance deprivation and temporal trajectory fragmentation caused by missing visual cues. Existing methods typically either dynamically update templates to maintain appearance similarity or employ autoregressive models to predict targets from historical trajectories. However, these methods are ineffective under severe occlusion owing to template contamination and limited frame rates for complex motion. In this work, we observe that occlusion inherently degrades the spatial matching mechanism, highlighting the importance of temporal cues. Meanwhile, event cameras with microsecond-level temporal resolution provide transient dynamic cues that facilitate modeling nonlinear motion. In light of this, we propose ***EvoTrack****, an occlusion-robust tracking framework via event-derived transient evolution, which comprises event-based motion autoregression and target-aware appearance matching. Specifically, for motion autoregression, the fine-grained timestamps of events naturally encode the target's direction and speed, motivating a bidirectional motion consistency that constrains inter-frame displacement prediction under nonlinear motion. For appearance matching, we adopt a Gaussian masking strategy to simulate occlusion degradation, guiding the model to focus on target regions and learn invariant representations. Furthermore, we build a pixel-aligned Frame-Event tracking dataset with higher spatial resolution and explicit occlusion labels. Extensive experiments demonstrate the effectiveness of EvoTrack in challenging occlusion scenes.*

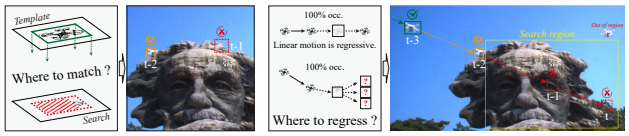
1. Introduction

Visual object tracking (VOT) seeks to localize arbitrary targets across video frames given their initial state. Despite the impressive advances of modern approaches [2, 7, 12, 13, 18, 27, 54], achieving reliable tracking under occlusion remains a long-standing challenge, particularly for targets exhibiting nonlinear motion. For example, agile drones momentarily hidden by obstacles or fast-moving vehicles temporarily occluded in traffic often cause catastrophic tracking failures. The fundamental difficulty arises from two coupled issues: (1) spatial appearance degradation that destroys template–search similarity, and (2) temporal trajectory interruption that complicates motion dynamics, as shown in Fig. 1(a). Tracking nonlinearly moving targets in occluded scenes remains a highly challenging problem.

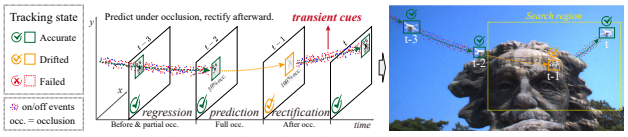
An intuitive solution is to dynamically update the template [31] or construct template libraries [24] to preserve appearance similarity between the template and the search region. However, such methods are prone to background



(a) Problem of visual object tracking in occluded scenarios



(b) Appearance-based matching (c) Trajectory-based autoregression



(d) Event-based motion autoregression

Figure 1. Illustration of three tracking paradigms. (a) Occlusion causes spatial appearance deprivation and temporal trajectory fragmentation. (b) Appearance-based matching fails under severe occlusions as similarity collapses. (c) Trajectory-based autoregression drifts under nonlinear motion during occlusion, hindering rectification afterward. (d) We explore event-based motion autoregression, which leverages inter-frame motion to achieve precise prediction during occlusion and effective rectification afterward.

*Corresponding author.

or occluder interference, leading to template contamination. More recently, several studies [50, 60] have explored learning occlusion-robust invariant representations through random masking strategies. Although these methods improve robustness, they remain constrained by the appearance-matching paradigm, which inherently fails under severe occlusions where target appearance is destroyed, especially during short-term full occlusion, as illustrated in Fig. 1(b).

Beyond spatial appearance modeling, recent works [1, 8, 29] have leveraged temporal cues to enhance tracking robustness. For instance, SeqTrack [9] and ARTrack [48] formulate tracking as a sequential autoregression problem, predicting the current target position from historical trajectories. While these approaches improve robustness against occlusion, they are sensitive to motion patterns and struggle with nonlinear motion of targets. This problem worsens under occlusion: the limited frame rate of conventional cameras fails to capture inter-frame dynamics, causing drift in autoregressive prediction. On the other hand, occlusion further fragments the sparse trajectories and amplifies prediction errors. Such errors may even push the target out of the search region, hindering recovery afterward, as shown in Fig. 1(c). The nonlinear motion of occluded targets poses a tricky challenge for autoregression-based trackers: *How can we achieve robust tracking of targets with severe appearance degradation and drastic trajectory variations*?

Fortunately, event cameras [30], with microsecond-level temporal resolution, provide a promising solution by capturing transient motion details lost by conventional cameras. Although existing event-based trackers [39, 42, 46, 59] perform well in high-speed and high dynamic scenes, they largely overlook the challenge of occlusion. To address this, our key insight is to leverage temporal motion prediction to alleviate appearance degradation under occlusion, while exploiting event streams to capture inter-frame dynamics for modeling nonlinear motion, as illustrated in Fig. 1(d).

Based on this idea, we propose EvoTrack, an occlusion-robust tracking framework that consists of Event-based Motion Autoregression (EMA) and Target-aware Appearance Matching (TAM). For motion autoregression, we observe that the fine-grained timestamps of events encode rich information about the target's direction and speed, benefiting the modeling of nonlinear motion. This inspires us to design a bidirectional motion consistency supervision to predict inter-frame displacements, enforcing physical constraints for motion prediction. For appearance matching, we further adopt a Gaussian-distributed masking strategy to particularly focus on the target region, thereby learning more invariant representations by simulating occlusion effects.

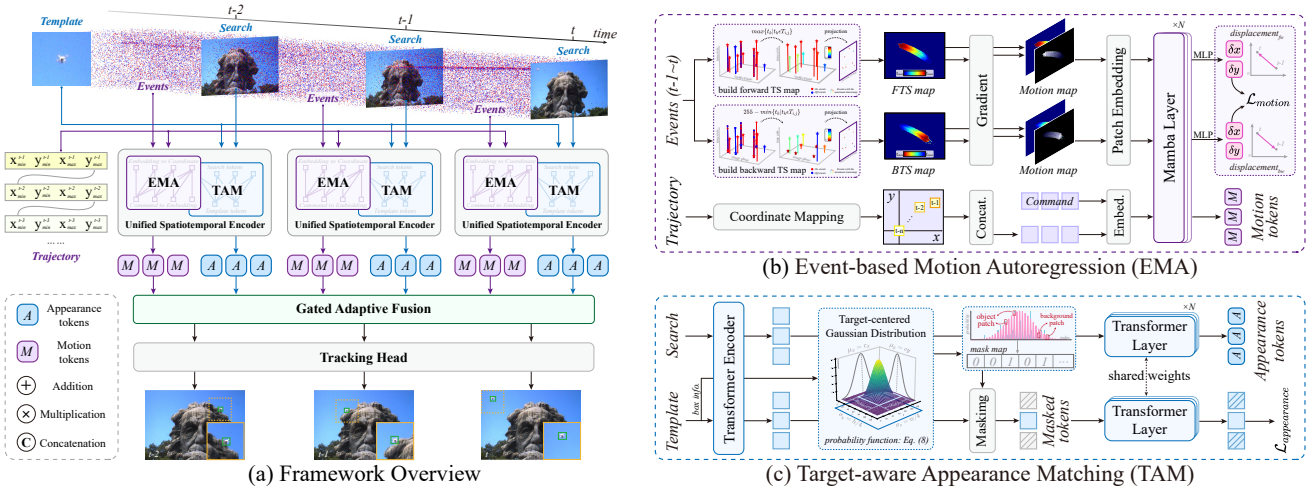
Specifically, in the EMA module, we first project inter-frame events to construct a time-surface representation that captures the transient dynamics of the target's motion. Target coordinates from previous frames are then transformed into the global coordinate system to form trajectory tokens. We combine these trajectory tokens with local TS cues to predict inter-frame displacements, supervised by bidirectional motion consistency. In the TAM module, a Gaussian mask is generated from the template state, and Transformer layers are employed to reconstruct template features, enhancing invariant target representation learning. Finally, an adaptive gating unit dynamically fuses motion and appearance features to handle varying occlusions.

In addition, we construct a coaxial imaging system to collect a real-world object tracking dataset (FEOT) with high spatial resolution and challenging occlusions, serving as an evaluation benchmark. Compared to existing datasets VisEvent [44], FE108 [55], COESOT [43], FELT [45], and CRSOT [63], FEOT contains occlusion sequences with varying durations and ratios, along with occlusion-level annotations. Our contributions are summarized as follows:

- We propose a novel occlusion-robust visual object tracking framework based on event cameras, named EvoTrack. This approach leverages temporal motion prediction to alleviate spatial appearance degradation, enabling stable tracking under challenging occlusion scenarios.
- We introduce an event-based motion autoregression that captures nonlinear motion dynamics from transient inter-frame events. Bidirectional motion consistency supervision imposes physical constraints, ensuring precise prediction during occlusion and rapid recovery afterward.
- We present a high spatial resolution Frame-Event tracking dataset with pixel-level alignment and occlusion-level annotations, serving as a reliable benchmark for evaluating occlusion-robust trackers. Extensive experiments demonstrate the superiority of the proposed method.

2. Related Work

Frame-based Visual Object Tracking. Frame-based VOT methods can be broadly categorized into two main approaches: appearance-based matching and trajectory-based autoregression. Appearance-based matching methods [3, 10, 19, 22, 49, 53, 54, 61] treat tracking as a similarity matching problem between the template and search region, locating the target via similarity computation. However, these methods heavily rely on appearance cues, making them unreliable under severe deformation or occlusion. Trajectory-based autoregression methods [1, 8, 9, 29, 48] model tracking as a sequential autoregressive problem, inferring the current target position from historical trajectories while incorporating appearance cues from the template and search image for enhanced robustness. Nevertheless, without explicit motion modeling, they tend to drift under nonlinear motion when appearance cues are unavailable (e.g., occlusion scenario). In this work, we aim to strengthen autoregressive modeling for nonlinear motion and improve tracking robustness under appearance degradation scenes.



(a) Framework Overview (b) Event-based Motion Autoregression (EMA) (c) Target-aware Appearance Matching (TAM)

Figure 2. The overall architecture of EvoTrack includes the Event-based Motion Autoregression (EMA) module and the Target-aware Appearance Matching (TAM) module. The EMA module predicts target positions during occlusions by leveraging both local transient motion from events and the global historical trajectory from frames. The TAM module extracts occlusion-robust, invariant features using a Gaussian-based masking strategy. Finally, a gated adaptive fusion mechanism is employed to integrate motion and appearance features.

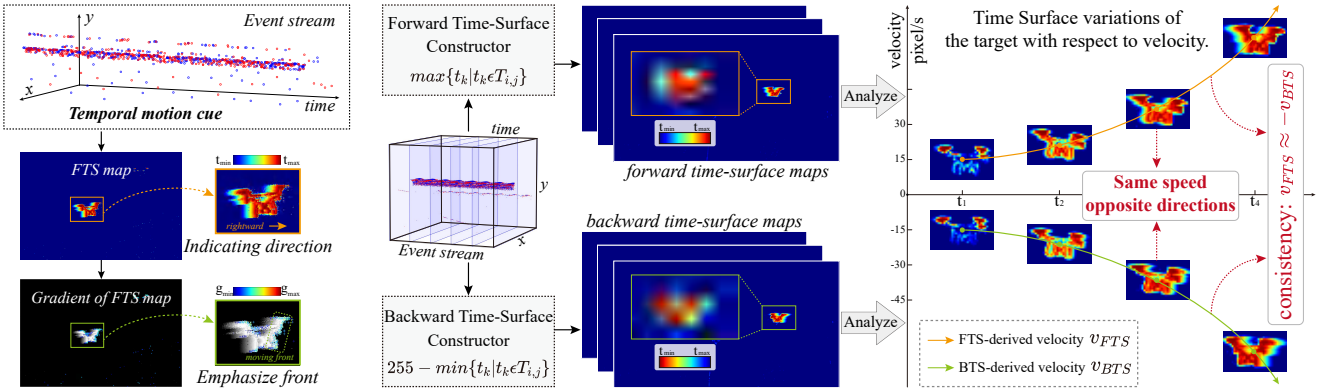
Event-based Visual Object Tracking. Event cameras [15, 16], with their superior properties such as high temporal resolution and high dynamic range, have greatly advanced the development of event-based VOT methods. Existing event-based works can be divided into event-only and frame–event fusion methods. Event-only methods [4, 6, 14, 46, 56, 64] mainly focus on designing effective event representations tailored for tracking. For instance, spiking or graph neural networks [56, 64] are adopted to process asynchronous event streams. However, constrained by the sparsity of events, such methods often achieve limited tracking accuracy. In contrast, frame–event fusion methods [21, 23, 39, 51, 55, 57, 58, 62, 65] have been more extensively explored, primarily due to the complementary advantages of images and events in terms of appearance texture and dynamic range. Although existing event-based methods have achieved landmark progress in challenging high-speed and high dynamic scenarios, they largely overlook the long-standing challenge of occlusion. In this work, we leverage events to capture inter-frame transient motion, mitigating spatial appearance degradation caused by occlusion.

Occlusion Scene Object Tracking. Occlusion remains a major challenge in computer vision [17, 33, 41, 47], as it suppresses the discriminative appearance of targets and leads to significant performance degradation. Early traditional trackers attempted to address this issue by dynamically updating templates [31, 32] or constructing occlusion-aware templates using spatial masks [35]. With the rise of deep learning-based approaches, several methods [5, 28, 37, 40, 50] have explored temporal modeling to mitigate occlusion degradation. LTOP [5] leverages recurrent neural networks to propagate target appearance features over time, modeling the occlusion process in a data-driven manner. DOCPF [28] and MTOA [40] maintain a template library to adaptively select reliable representations under different occlusion conditions. More recently, ORTrack [50] introduces random masking operations [36, 60] to enhance target representations under frequent occlusions. Nevertheless, most approaches still rely on spatial appearance cues, which are impaired under severe or complete occlusions, making tracking difficult. Building on autoregressive tracking, we move beyond appearance modeling toward temporal motion prediction, leveraging transient motion cues from event streams to enhance robustness against occlusion.

3. Unified Motion-Appearance Network

3.1. Overall Architecture

Visual object tracking under severe occlusion suffers from significant spatial appearance deprivation and temporal trajectory fragmentation. To address this, we leverage temporal motion cues to compensate for damaged spatial information, which requires precise modeling of complex nonlinear motion. We introduce event cameras to recover transient inter-frame dynamics missed by conventional cameras and propose an occlusion-robust autoregressive tracking framework in Fig. 2, including event-based motion autoregression and target-aware appearance matching. The motion autoregression module constructs time-surface representations from events and trajectory tokens from historical coordinates to locate the target under occlusion. A bidirectional time-surface motion consistency supervision enforces physical constraints for accurate motion learning. The appearance matching module builds a Gaussian-based



(a) FTS representation (b) Motion consistency based on bidirectional time-surface

Figure 3. Illustration of motion cues contained in TS representation. (a) The FTS encodes temporal information, revealing the target's motion direction, with its gradient highlighting the front region. (b) The FTS/BTS reflects the target's motion speed through tailing, motivating the design of a bidirectional time-surface motion consistency supervision to improve the modeling of nonlinear target motion.

spatial distribution to generate a target mask, simulating occlusion and enabling robust feature reconstruction. Finally, an adaptive gating unit fuses motion and appearance cues. This design ensures accurate target prediction and rapid recovery under severe occlusion, while maintaining high tracking accuracy in non- or mildly occluded scenarios.

3.2. Event-based Motion Autoregression

Time-surface Construction. The time-surface (TS) offers a compact representation of event streams, assigning each pixel with the maximum timestamp of recent events to maintain temporal information. This representation, coupled with the high temporal resolution of events, effectively captures transient inter-frame motion cues: **temporal increments indicate motion direction, while event trails imply motion speed**. However, existing TS formulations based on the exponential decay kernel [26, 52] fail to exhibit clear motion boundaries, hindering the indication of motion speed from event trails (details are provided in the supplementary material). To address this issue, we propose a novel forward time-surface (FTS) formulation that strengthens the correlation between temporal evolution and target motion. Specifically, FTS is constructed from an event set $\xi = \{e = (p_k, t_k, x_k, y_k)\}_{k=1}^{N}$ recorded within $[st, et]$, where (x_k, y_k), p_k, and t_k represent the pixel coordinates, polarity, and timestamp, respectively, and N is the total event count. All timestamps are first normalized to $[0, 255]$, yielding the normalized event set ξ^*. From this set, we then construct a time image I_f:

$$I_f(i,j) = \max\{t_e \mid e \in \xi^*,\ x_e = i,\ y_e = j\}, \quad (1)$$

where $I_f(i,j)$ denotes the latest event timestamp at pixel (i,j), and pixels without events are assigned a value of zero. Next, we introduce a histogram equalization transformation [38] to ensure the FTS accurately reflects target motion, alleviating the uneven distribution of triggered events: $FTS(i,j) = H(I_f(i,j))$. The transformation $H(*)$ is explained in the supplementary material for brevity.

The linear forward time-surface encodes the temporal information of events into a 2D spatial distribution, directly recording transient motion dynamics. To accentuate motion fronts, we compute its gradient map. The FTS and its gradient are then stacked channel-wise to form a consolidated motion map for feature learning, as shown in Fig. 3(a).

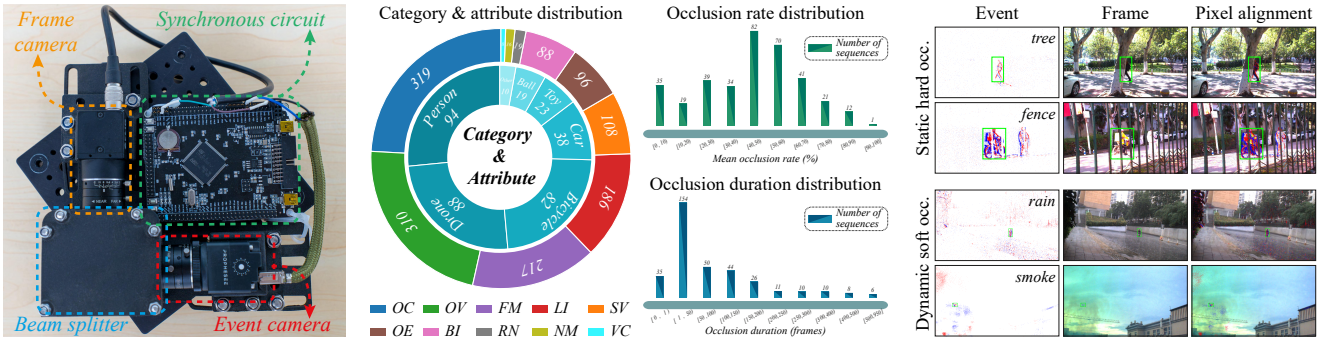
Motion Autoregression Paradigm. Trajectory-based autoregressive tracking paradigm [48] can be formulated as:

$$P(Y^t | Y^{t-1-N:t-1}, (C, Z, X^t)), \quad (2)$$

where Z and X^t denote the template and search image, respectively, C represents the command token, and Y^t is the target position at t. This paradigm models tracking as a sequential autoregressive process [25], where the current position Y^t is inferred from the latest N positions by leveraging temporal trajectory dependencies, thereby enhancing robustness in long-term tracking. However, under severe occlusion, the appearance cues in X^t become unreliable and the similarity between Z and X^t degrades significantly, causing drift accumulation, especially in complex nonlinear motion, after which re-localization becomes highly challenging. To mitigate this issue, we extend the trajectory-autoregression into a motion-autoregressive formulation. An event camera is introduced to capture fine-grained transient inter-frame motion, enabling accurate position regression even when appearance cue is severely degraded. The motion-autoregressive process can be expressed as:

$$P(Y^t | (Y^{t-1-N:t-1}, M^{t-1:t}, C), (Z, X^t)), \quad (3)$$

where M^t denotes the motion map obtained by projecting inter-frame events, providing localized transient cues for position prediction. Meanwhile, appearance matching be-



(a) Collection device (b) Statistics analysis of the FEOT (c) Occlusion samples

Figure 4. Illustration of the proposed FEOT dataset. (a) The collection device consists of a frame camera and an event camera, forming a coaxial system via a beam splitter. (b) Statistical analysis of the FEOT dataset, covering categories, attributes, occlusion ratio, and duration. (c) Representative examples of various occlusions, including static hard occlusions (e.g., fences) and dynamic soft occlusions (e.g., smoke).

tween Z and X^t is preserved to ensure higher tracking accuracy under normal (non-occluded) conditions. Concretely, we first project target boxes from previous frames into a global coordinate system to construct a unified trajectory representation. These unified boxes are then transformed into trajectory tokens together with command tokens. The event-based motion map is partitioned into patches and concatenated with the trajectory tokens to form the token embeddings. Finally, a Mamba module [20] is employed to extract motion features for target regression.

Bidirectional motion consistency. During construction of the forward time-surface, each pixel is continuously updated by subsequent events, producing a temporal evolution from old to new. This naturally leads to the question: Can we construct a complementary "new-to-old" view? To this end, we build a backward time image I_b based on Eq. (1):

$$I_b(i,j) = \min\{t_e \mid e \in \xi^*,\, x_e = i,\, y_e = j\}, \quad (4)$$

where at each pixel the earliest available event is selected. Pixels containing no events are assigned a value of 255. Histogram equalization is then applied to obtain the final backward time-surface (BTS) representation as follows:

$$BTS(i,j) = H(255 - I_b(i,j)). \quad (5)$$

FTS and BTS provide two opposite temporal views of events occurring within the same interval: forward (old-to-new) and backward (new-to-old). This dual perspective enables modeling the same physical motion both as forward and time-reversed motion. The two processes have identical speeds but opposite directions, offering an intrinsic motion consistency signal, as illustrated in Fig. 3(b). Specifically, motion maps from FTS and BTS are concatenated with trajectory tokens and fed into the Mamba module to extract motion features. A shared-weight MLP then predicts the inter-frame displacements δ_{forward} and δ_{backward} from the respective features. By leveraging this dual relationship, we impose a bidirectional motion consistency supervision that serves as an explicit physical constraint during training.

3.3. Target-aware Appearance Matching

Although we leverage temporal motion prediction to handle occlusion, this does not imply completely discarding appearance cues, which remain essential for accurate tracking under non-occluded or mildly occluded conditions. Existing works [50, 60] often adopt random masking to reconstruct the template, enhancing the model's ability to extract invariant appearance features. However, these methods apply random masks over the entire template, which typically includes both target and background regions. Such indiscriminate masking can cause the model to learn background interference, particularly under occlusion, leading to erroneous matching responses. To address this limitation, we propose a target-aware Gaussian masking strategy. Using prior knowledge of the target box in the template, we construct a Gaussian distribution centered on the target, guiding the mask to focus primarily on the target region. The probability density function f_g is defined as:

$$f_g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp(-[\frac{(x-c_x)^2}{\sigma_x^2} + \frac{(y-c_y)^2}{\sigma_y^2}]), \quad (6)$$

where (c_x, c_y) denote the center of the target box in the template, and the standard deviations $[\sigma_x, \sigma_y]$ are set to one-fourth of the target box width and height, respectively.

This target-centered Gaussian mask effectively simulates occlusion on the target, encouraging the model to attend to discriminative parts during training and enhancing feature invariance. Meanwhile, reducing the masking probability in background regions significantly alleviates interference. Specifically, we first extract appearance features from the template and search region via cross-attention, generate a Gaussian-guided mask over the target, and reconstruct the template features through a shared self-attention module.

3.4. Gated Adaptive Fusion and Loss Function

Occlusion severity in real-world tracking varies: appearance features are reliable under mild occlusion, while mo-

Table 1. Comparison with SOTA trackers on the FE108, VisEvent, COESOT, and FEOT datasets. Best results are **bolded**.

Method	Type	FE108		VisEvent		COESOT		FEOT	
		PR(%)	SR(%)	PR(%)	SR(%)	PR(%)	SR(%)	PR(%)	SR(%)
MixFormer [11]	Frame	75.7	51.6	69.9	53.3	75.3	65.1	35.8	28.3
ORTrack [50]	Frame	59.4	38.8	55.1	40.8	59.2	46.8	21.4	19.0
SeqTrack [9]	Frame	80.5	55.4	76.9	60.7	82.2	71.8	50.1	38.2
ARTrack [48]	Frame	74.1	49.9	70.0	54.3	75.1	64.6	39.1	30.6
STNet [56]	Event	89.6	58.5	49.2	35.5	62.3	50.6	47.6	34.5
HDETrack [46]	Event	92.2	59.8	54.6	37.3	64.1	53.1	53.1	40.1
AFNet [57]	Frame+Event	87.0	58.4	59.3	44.5	67.8	59.2	49.5	36.8
CEUTrack [43]	Frame+Event	84.5	55.6	69.1	53.1	76.0	62.7	50.8	39.3
ViPT [62]	Frame+Event	93.8	65.8	75.8	59.2	84.9	75.4	55.4	43.4
SDSTrack [23]	Frame+Event	92.0	64.6	76.7	59.7	84.5	74.9	58.0	45.1
SeqTrack v2 [8]	Frame+Event	92.8	65.5	79.4	**63.0**	85.0	75.9	56.1	43.1
EvoTrack	Frame+Event	**94.6**	**68.4**	**80.1**	62.1	**85.4**	**76.2**	**62.7**	**45.2**

tion features dominate under severe occlusion. Therefore, we introduce a Gated Adaptive Fusion (GAF) module that dynamically combines these cues. The model is trained with a cross-entropy loss for classification, a combined GIoU and L1 loss for bounding box regression, and MSE losses for both appearance reconstruction and inter-frame displacement prediction. The overall loss is as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ce} + \lambda_2 \mathcal{L}_{giou} + \lambda_3 \mathcal{L}_{l1} + \lambda_4 \mathcal{L}_{app.} + \lambda_5 \mathcal{L}_{mot.}, \quad (7)$$

where λ_i is a balancing weight. $\mathcal{L}_{app.}$ and $\mathcal{L}_{mot.}$ denote the reconstruction and prediction losses, respectively.

4. Frame-Event Occlusion Dataset

Data Collection and Annotation. Existing frame–event datasets are mostly captured with the DAVIS346 event camera, whose limited 346×260 resolution constrains model evaluation. Moreover, they rarely include occlusion-level annotations. To address these limitations, we propose a novel coaxial data collection system (Fig. 4(a)) combining a frame and an event camera. After alignment and cropping, the paired data achieve a spatial resolution of 1070×610. Based on this setup, we construct the **F**rame–**E**vent-based **O**ccluded **T**racking dataset (FEOT Link), tailored for occlusion scenes. Compared with existing datasets [43–45, 55], FEOT provides higher resolution and fine-grained occlusion annotations covering dynamic (e.g., moving vehicles) and static (e.g., buildings) as well as hard (e.g., trees) and soft (e.g., smoke) occlusion types. It contains 354 videos with 73K frames and synchronized event streams. All bounding boxes are annotated by a professional data-labeling organization and verified through multiple quality checks. In addition, FEOT defines 11 occlusion levels, establishing a new benchmark for visual object tracking under occlusion.

Data Analysis. We conduct a comprehensive statistical analysis of the FEOT dataset, as shown in Fig. 4(b). FEOT contains ten object categories, covering six common types such as persons, cars, drones, bicycles, and others. Ten challenging tracking attributes are defined, among which occlusion is the most representative. Detailed explanations of all attributes are provided in the supplementary material. As FEOT mainly focuses on occlusion scenarios, we further analyze the distributions of occlusion ratio and duration. The occlusion ratio of each sequence is calculated as the average ratio of all occluded frames and divided into 11 levels ranging from no occlusion to complete occlusion, with most sequences falling between 20% and 70%. For occlusion duration, we count the number of occluded frames per sequence, showing that most occlusions last for 1–50 frames, while a few persist for 50–1000 frames.

5. Experiments

5.1. Experiments Setup

Implementation Details. We implement EvoTrack in PyTorch and train it on 8 NVIDIA RTX 3090 GPUs with a batch size of 8. The AdamW optimizer is used with a weight decay of 5×10^{-4} and a learning rate of 8×10^{-5}. The motion branch employs the pre-trained Mamba [20] module, and the appearance branch uses ViT-B with pre-trained DINOv2 [34] weights. The search region and template sizes are 224×224 and 112×112, respectively. EvoTrack is fine-tuned for 200 epochs on the training set.

Datasets and Compared Methods. To demonstrate the effectiveness of our approach, we compare our tracker with representative state-of-the-art methods on widely used benchmarks: VisEvent [44], FE108 [55], and COESOT [43]. The proposed FEOT dataset is used exclusively for evaluating the robustness of trackers under occlusions and is not intended for training. For evaluation, we adopt the widely used Precision Rate (PR) and Success Rate (SR) metrics to quantitatively assess tracking performance. Representative methods from frame-based, event-based, and frame-event approaches are selected for comparison.

5.2. Comparison with State-of-the-art

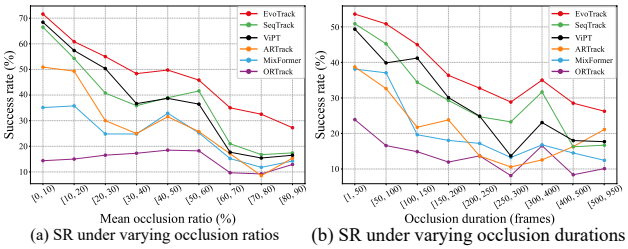
We compare our method with recent SOTA trackers on four datasets, including three public benchmarks and our

Table 2. Ablation study on key components of EvoTrack.

TAM	EMA_{base}	EMA_{bmc}	PR(%)	SR(%)
✓			91.4	62.8
	✓		84.1	50.2
		✓	87.3	56.4
✓	✓		92.1	66.9
✓		✓	**94.6**	**68.4**

Table 3. Ablation study on different masking strategies.

Strategy	w/o Masking PR / SR (%)	w/ Random PR / SR (%)	w/ Gaussian PR / SR (%)
Metric	75.7 / 59.6	79.7 / 60.2	**80.1 / 62.1**



(a) SR under varying occlusion ratios (b) SR under varying occlusion durations

Figure 5. Analysis of occlusions on tracking performance.

Table 4. Ablation study on different fusion strategies.

Strategy	Add PR / SR (%)	Concatenate PR / SR (%)	Gated adaptive PR / SR (%)
Metric	77.5 / 61.8	78.9 / 62.0	**80.1 / 62.1**

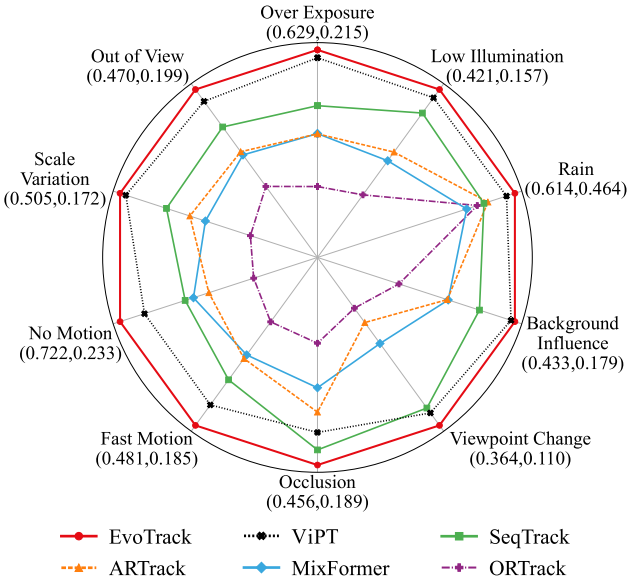


Figure 6. Comparison of SR across different attributes on FEOT.

constructed FEOT dataset. As shown in Tab. 1, our method achieves the best overall performance across all datasets, demonstrating generalization to diverse tracking scenarios.

FE108 is an indoor dataset that emphasizes fast and nonlinear motion scenarios. Our EvoTrack achieves state-of-the-art performance with 68.4% SR and 94.6% PR, demonstrating its competitiveness in complex motion challenges.

VisEvent is a large-scale frame-event dataset and the most widely used bimodal benchmark. Our method outperforms the previous best tracker by 0.7% PR, while achieving the suboptimal SR. We attribute this to the lack of raw event files in some videos, which hindered the training process.

COESOT is a general-purpose dataset covering 90 object categories and 17 attributes. Our method obtains superior performance with 76.2% SR and 85.4% PR, highlighting the capability of EvoTrack in general object tracking scenes.

FEOT is a high-resolution tracking dataset collected for occlusion scenarios. It includes 11 levels of occlusion labels and covers diverse occlusions. Our EvoTrack surpasses other top-performing trackers by a clear margin, verifying the effectiveness of motion cues under appearance degradation. Additionally, we further evaluate tracker performance across different challenging attributes in Fig. 6. Our method consistently outperforms previous trackers, underscoring the adaptability and robustness of the proposed approach.

5.3. Ablation Study and Discussion

What role does each component of EvoTrack play? Tab. 2 demonstrates the roles of the two core components in EvoTrack: motion autoregression (*EMA*) and appearance matching (*TAM*). EMA_{base} employs only the forward time-surface for motion autoregression, while EMA_{bmc} leverages both forward and backward time-surfaces with motion-consistency supervision. Removing the EMA module degrades performance, mainly due to the appearance discrepancy between the search region and the template. Removing the TAM module forces the tracker to rely solely on the motion branch, leading to degraded accuracy due to the lack of appearance guidance. By integrating appearance and motion information, EvoTrack achieves a notable performance gain, as the two modalities are strongly complementary: motion information provides short-term positional compensation when appearance cues deteriorate, whereas appearance cues correct localization errors when motion prediction deviates. Furthermore, incorporating the bidirectional time-surface brings improvements of 2.5% PR and 1.5% SR, validating that motion-consistency supervision facilitates accurate modeling of the target's motion.

Analysis of Occlusion Degradation. We analyze the effect of occlusion ratio and duration on tracking performance in Fig. 5. In each analysis, the occlusion duration and ratio are set to 1–50 frames and 30–60%, respectively. As expected, performance gradually drops with increasing occlusion severity and duration. When the occlusion ratio exceeds 60%, the SR metric exhibits significant degradation. Notably, our tracker maintains higher SR across all conditions, demonstrating strong robustness against occlusion.

Effectiveness of Gaussian-based Masking. We evaluate the effect of different masking strategies on appearance matching in Tab. 3. Applying random masking brings a

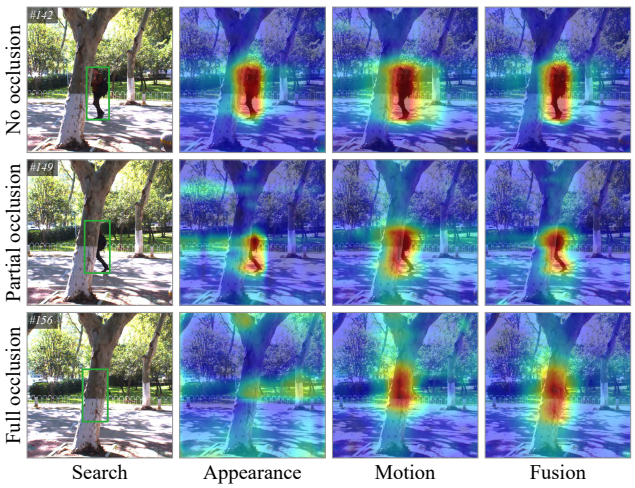
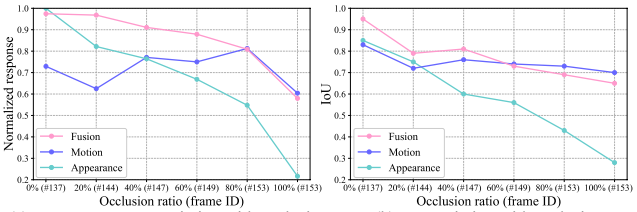


Figure 7. Attention visualizations of EvoTrack in the appearance, motion, and fusion branches. The green box represents the GT.



(a) Feature response variation with occlusion (b) IoU variation with occlusion

Figure 8. Evolution of different branches over occlusion process.

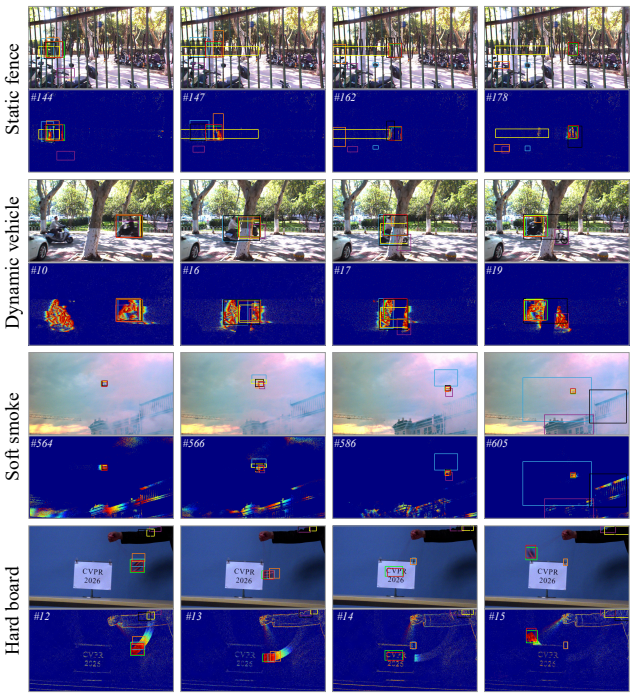


Figure 9. Qualitative comparison of EvoTrack with other SOTA trackers under various occlusions on the FEOT dataset. The upper row shows frames, the bottom row shows forward time-surfaces.

clear performance gain, indicating that simulating occlusion through masking encourages the model to learn more robust appearance representations. Further introducing Gaussian masking yields the best performance, demonstrating that spatially target-aware masking better guides the model to focus on target-related cues under visual degradation and facilitates the extraction of invariant representations.

Effectiveness of Gated Adaptive Fusion. We investigate different fusion strategies to validate the gated-adaptive fusion design, as shown in Tab. 4. Concatenating motion and appearance features moderately improves over direct addition, indicating that richer joint representations enhance the target discrimination. The gated adaptive fusion performs best, showing that dynamically weighting motion and appearance features improves robustness under varying occlusions through complementary integration.

Discussion of Attention Evolution. We visualize the attention maps of the appearance, motion, and fusion branches under occlusions in Fig. 7. As occlusion grows, appearance responses degrade sharply, while motion activations remain stable. Their fusion yields robust representations, showing that motion cues effectively compensate for appearance loss. To further quantify this, we analyze the full occlusion progression in Fig. 8. Appearance responses decline with occlusion, while motion remains stable, causing fusion to favor motion. IoU curves follow the same trend: tracking accuracy drops as appearance deteriorates, whereas motion-based predictions effectively mitigate occlusion failure.

Qualitative Results. We qualitatively compare EvoTrack with SOTA methods under various occlusions in Fig. 9. Existing trackers can track targets under partial occlusion with minor drift but fail under severe cases. In contrast, EvoTrack achieves stable tracking of both linear and nonlinear targets, even during short-term full occlusion, thanks to the EMA with transient motion cues. These results demonstrate our philosophy that temporal motion prediction effectively mitigates appearance degradation under occlusion.

6. Conclusion

In this work, we propose an occlusion-robust visual tracking framework that leverages temporal motion prediction to alleviate spatial appearance degradation. We introduce an event-based motion autoregression module that models nonlinear target motion by jointly exploiting global trajectories from frames and local dynamics from events, enabling accurate prediction during occlusion and rapid recovery afterward. Additionally, we employ a Gaussian-distributed masking strategy to extract occlusion-robust invariant representations of targets. Moreover, we construct a high-resolution frame-event tracking dataset with pixel-level alignment and occlusion-level annotations, on which the proposed framework demonstrates superior performance.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant U24B20139 and 62371203.

References

[1] Yifan Bai, Zeyang Zhao, Yihong Gong, and Xing Wei. Artrackv2: Prompting autoregressive tracker where to look and how to describe. In *Proc. CVPR*, 2024. 2

[2] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Proc. ECCV*, 2016. 1

[3] Wenrui Cai, Qingjie Liu, and Yunhong Wang. Spmtrack: Spatio-temporal parameter-efficient fine-tuning with mixture of experts for scalable visual tracking. In *Proc. CVPR*, 2025. 2

[4] Yujeong Chae, Lin Wang, and Kuk-Jin Yoon. Siamevent: Event-based object tracking via edge-aware similarity learning with siamese networks. *arXiv preprint*, 2021. 3

[5] Satyaki Chakraborty and Martial Hebert. Learning to track object position through occlusion. *arXiv preprint*, 2021. 3

[6] Haosheng Chen, David Suter, Qiangqiang Wu, and Hanzi Wang. End-to-end learning of object motion estimation from retinal events for event-based object tracking. In *Proc. AAAI*, 2020. 3

[7] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proc. CVPR*, 2021. 1

[8] Xin Chen, Ben Kang, Jiawen Zhu, Dong Wang, Houwen Peng, and Huchuan Lu. Unified sequence-to-sequence learning for single- and multi-modal visual object tracking. *arXiv preprint*, 2023. 2, 6

[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. 2, 6

[10] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proc. CVPR*, 2022. 2

[11] 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. 6

[12] Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. In *Proc. CVPR*, 2020. 1

[13] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *Proc. CVPR*, 2019. 1

[14] Yingkai Fu, Meng Li, Wenxi Liu, Yuanchen Wang, Jiqing Zhang, Baocai Yin, Xiaopeng Wei, and Xin Yang. Distractor-aware event-based tracking. *IEEE Trans. Image Process.*, 32:6129–6141, 2023. 3

[15] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):154–180, 2020. 3

[16] Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014):1034–1040, 2024. 3

[17] Juan Luis Gonzalez, Xu Yao, Alex Whelan, Kyle Olszewski, Hyeongwoo Kim, and Pablo Garrido. Videospats: Video spatiotemporal splines for disentangled occlusion, appearance and motion modeling and editing. In *Proc. CVPR*, 2025. 3

[18] Dongyan Guo, Jun Wang, Ying Cui, Zhenhua Wang, and Shengyong Chen. Siamcar: Siamese fully convolutional classification and regression for visual tracking. In *Proc. CVPR*, 2020. 1

[19] Mingzhe Guo, Weiping Tan, Wenyu Ran, Liping Jing, and Zhipeng Zhang. Dreamtrack: Dreaming the future for multimodal visual object tracking. In *Proc. CVPR*, 2025. 2

[20] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *Proc. CVPR*, 2025. 5, 6

[21] Lingyi Hong, Shilin Yan, Renrui Zhang, Wanyun Li, Xinyu Zhou, Pinxue Guo, Kaixun Jiang, Yiting Chen, Jinglun Li, Zhaoyu Chen, and Wenqiang Zhang. Onetracker: Unifying visual object tracking with foundation models and efficient tuning. In *Proc. CVPR*, 2024. 3

[22] Lingyi Hong, Jinglun Li, Xinyu Zhou, Shilin Yan, Pinxue Guo, Kaixun Jiang, Zhaoyu Chen, Shuyong Gao, Runze Li, Xingdong Sheng, Wei Zhang, Hong Lu, and Wenqiang Zhang. General compression framework for efficient transformer object tracking. In *Proc. ICCV*, 2025. 2

[23] Xiaojun Hou, Jiazheng Xing, Yijie Qian, Yaowei Guo, Shuo Xin, Junhao Chen, Kai Tang, Mengmeng Wang, Zhengkai Jiang, Liang Liu, and Yong Liu. Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In *Proc. CVPR*, 2024. 3, 6

[24] Yuqing Huang, Xin Li, Zikun Zhou, Yaowei Wang, Zhenyu He, and Ming-Hsuan Yang. Rtracker: Recoverable tracking via pn tree structured memory. In *Proc. CVPR*, 2024. 1

[25] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proc. ICML*, 2020. 4

[26] 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. 4

[27] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proc. CVPR*, 2019. 1

[28] Yangfan Li, Nan Wang, Wei Li, Xiong Li, and Mengbin Rao. Object tracking in satellite videos with distractor–occlusion-aware correlation particle filters. *IEEE Trans. Geosci. Remote Sens.*, 62:1–12, 2024. 3

[29] Shiyi Liang, Yifan Bai, Yihong Gong, and Xing Wei. Autoregressive sequential pretraining for visual tracking. In *Proc. CVPR*, 2025. 2

[30] Songnan Lin, Ye Ma, Jing Chen, and Bihan Wen. Compressed event sensing (ces) volumes for event cameras. *Int. J. Comput. Vis.*, 133(1):435–455, 2025. 2

[31] Hieu Tat Nguyen and Arnold WM Smeulders. Fast occluded object tracking by a robust appearance filter. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(8):1099–1104, 2004. 1, 3

[32] Hieu Tat Nguyen, Marcel Worring, and Rein Van Den Boomgaard. Occlusion robust adaptive template tracking. In *Proc. ICCV*, 2001. 3

[33] Khanh Nguyen, Ghulam Mubashar Hassan, and Ajmal Mian. Occlusion-aware text-image-point cloud pretraining for open-world 3d object recognition. In *Proc. CVPR*, 2025. 3

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint*, 2023. 6

[35] Jiyan Pan and Bo Hu. Robust occlusion handling in object tracking. In *Proc. CVPR*, 2007. 3

[36] Haolin Qin, Tingfa Xu, Tianhao Li, Zhenxiang Chen, Tao Feng, and Jianan Li. Must: The first dataset and unified framework for multispectral uav single object tracking. In *Proc. CVPR*, 2025. 3

[37] Gozde Sahin and Laurent Itti. Multi-task occlusion learning for real-time visual object tracking. In *Proc. ICIP*, 2021. 3

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

[39] Chuanyu Sun, Jiqing Zhang, Yang Wang, Huilin Ge, Qianchen Xia, Baocai Yin, and Xin Yang. Exploring historical information for rgbe visual tracking with mamba. In *Proc. CVPR*, 2025. 2, 3

[40] Lifan Sun, Jiayi Zhang, Dan Gao, Bo Fan, and Zhumu Fu. Occlusion-aware visual object tracking based on multi-template updating siamese network. *Digit. Signal Process.*, 148:104440, 2024. 3

[41] Tao Tan and Qiulei Dong. Onda-pose: Occlusion-aware neural domain adaptation for self-supervised 6d object pose estimation. In *Proc. CVPR*, 2025. 3

[42] Yuedong Tan, Jiawei Shao, Eduard Zamfir, Ruanjun Li, Zhaochong An, Chao Ma, Danda Paudel, Luc Van Gool, Radu Timofte, and Zongwei Wu. What you have is what you track: Adaptive and robust multimodal tracking. In *Proc. ICCV*, 2025. 2

[43] Chuanming Tang, Xiao Wang, Ju Huang, Bo Jiang, Lin Zhu, Shifeng Chen, Jianlin Zhang, Yaowei Wang, and Yonghong Tian. Revisiting color-event based tracking: A unified network, dataset, and metric. *Pattern Recognition*, 172: 112718, 2026. 2, 6

[44] 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, 2023. 2, 6

[45] Xiao Wang, Xufeng Lou, Shiao Wang, Ju Huang, Lan Chen, and Bo Jiang. Long-term visual object tracking with event cameras: An associative memory augmented tracker and a benchmark dataset. *arXiv preprint*, 2024. 2, 6

[46] Xiao Wang, Shiao Wang, Chuanming Tang, Lin Zhu, Bo Jiang, Yonghong Tian, and Jin Tang. Event stream-based visual object tracking: A high-resolution benchmark dataset and a novel baseline. In *Proc. CVPR*, 2024. 2, 3, 6

[47] Yu-Hsiang Wang, Jun-Wei Hsieh, Ping-Yang Chen, Ming-Ching Chang, Hung-Hin So, and Xin Li. Smiletrack: Similarity learning for occlusion-aware multiple object tracking. In *Proc. AAAI*, 2024. 3

[48] Xing Wei, Yifan Bai, Yongchao Zheng, Dahu Shi, and Yihong Gong. Autoregressive visual tracking. In *Proc. CVPR*, 2023. 2, 4, 6

[49] Qiangqiang Wu, Yi Yu, Chenqi Kong, Ziquan Liu, Jia Wan, Haoliang Li, Alex C Kot, and Antoni B Chan. Temporal unlearnable examples: Preventing personal video data from unauthorized exploitation by object tracking. In *Proc. ICCV*, 2025. 2

[50] 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. 2, 3, 5, 6

[51] Zongwei Wu, Jilai Zheng, Xiangxuan Ren, Florin-Alexandru Vasluianu, Chao Ma, Danda Pani Paudel, Luc Van Gool, and Radu Timofte. Single-model and any-modality for video object tracking. In *Proc. CVPR*, 2024. 3

[52] 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. 4

[53] Chaocan Xue, Bineng Zhong, Qihua Liang, Yaozong Zheng, Ning Li, Yuanliang Xue, and Shuxiang Song. Similarity-guided layer-adaptive vision transformer for uav tracking. In *Proc. CVPR*, 2025. 2

[54] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *Proc. ECCV*, 2022. 1, 2

[55] Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong. Object tracking by jointly exploiting frame and event domain. In *Proc. ICCV*, 2021. 2, 3, 6

[56] Jiqing Zhang, Bo Dong, Haiwei Zhang, Jianchuan Ding, Felix Heide, Baocai Yin, and Xin Yang. Spiking transformers for event-based single object tracking. In *Proc. CVPR*, 2022. 3, 6

[57] Jiqing Zhang, Yuanchen Wang, Wenxi Liu, Meng Li, Jinpeng Bai, Baocai Yin, and Xin Yang. Frame-event alignment and fusion network for high frame rate tracking. In *Proc. CVPR*, 2023. 3, 6

[58] Jiqing Zhang, Bo Dong, Yingkai Fu, Yuanchen Wang, Xiaopeng Wei, Baocai Yin, and Xin Yang. A universal event-based plug-in module for visual object tracking in degraded conditions. *Int. J. Comput. Vis.*, 132(5):1857–1879, 2024. 3

[59] Tianlu Zhang, Kurt Debattista, Qiang Zhang, Guiguang Ding, and Jungong Han. Revisiting motion information for rgb-event tracking with mot philosophy. In *Proc. NeurIPS*, 2024. 2

[60] Haojie Zhao, Dong Wang, and Huchuan Lu. Representation learning for visual object tracking by masked appearance transfer. In *Proc. CVPR*, 2023. 2, 3, 5

[61] Jikai Zheng, Mingjiang Liang, Shaoli Huang, and Jifeng Ning. Exploring the feature extraction and relation modeling for light-weight transformer tracking. In *Proc. ECCV*, 2024. 2

[62] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proc. CVPR*, 2023. 3, 6

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

[64] Zhiyu Zhu, Junhui Hou, and Xianqiang Lyu. Learning graph-embedded key-event back-tracing for object tracking in event clouds. In *Proc. NeurIPS*, 2022. 3

[65] Zhiyu Zhu, Junhui Hou, and Dapeng Oliver Wu. Cross-modal orthogonal high-rank augmentation for rgb-event transformer-trackers. In *Proc. ICCV*, 2023. 3