International Journal of Computer Vision (2024) 132:1857–1879
https://doi.org/10.1007/s11263-023-01959-8

A Universal Event-Based Plug-In Module for Visual Object Tracking in Degraded Conditions

Jiqing Zhang[1,2] · Bo Dong[3] · Yingkai Fu[1] · Yuanchen Wang[1] · Xiaopeng Wei[1] · Baocai Yin[1] · Xin Yang[1]

Received: 17 May 2023 / Accepted: 13 November 2023 / Published online: 18 December 2023

Abstract

Most existing trackers based on RGB/grayscale frames may collapse due to the unreliability of conventional sensors in some challenging scenarios (e.g., motion blur and high dynamic range). Event-based cameras as bioinspired sensors encode brightness changes with high temporal resolution and high dynamic range, thereby providing considerable potential for tracking under degraded conditions. Nevertheless, events lack the fine-grained texture cues provided by RGB/grayscale frames. This complementarity encourages us to fuse visual cues from the frame and event domains for robust object tracking under various challenging conditions. In this paper, we propose a novel event feature extractor to capture spatiotemporal features with motion cues from event-based data by boosting interactions and distinguishing alterations between states at different moments. Furthermore, we develop an effective feature integrator to adaptively fuse the strengths of both domains by balancing their contributions. Our proposed module as the plug-in can be easily applied to off-the-shelf frame-based trackers. We extensively validate the effectiveness of eight trackers extended by our approach on three datasets: EED, VisEvent, and our collected frame-event-based dataset FE141. Experimental results also show that event-based data is a powerful cue for tracking.

Keywords Event-based camera · Visual object tracking · Multimodal fusion · Plug-in module

1 Introduction

Visual object tracking is a fundamental yet challenging topic in computer vision, which aims to predict the target state in each frame of a video sequence. Based on convolutional neural networks, object tracking using conventional RGB cameras has made remarkable progress in recent years (Bertinetto et al., 2016; Bhat et al., 2020; Chen et al., 2020; Guo et al., 2020; Li et al., 2019; Zhang et al., 2021; Fan et al., 2021; Zhang et al., 2018; Zhang and Peng, 2019; Zhao et al., 2022; Gao et al., 2020; Zhou et al., 2021; Cui et al., 2022; Cai et al., 2023). However, due to the frame rate and dynamic range limits of conventional sensors, current frame-based trackers are often overwhelmed in some degraded scenarios, such as fast motion and high dynamic range. By contrast, event-based cameras are bio-inspired vision sensors whose functioning principle fundamentally differs from conventional cameras. Instead of capturing frames at a fixed rate, an event camera measures brightness changes asynchronously and outputs a stream of events representing the location, timestamp, and sign of the brightness change. Compared to conventional cameras, event-based cameras

Communicated by Boxin Shi.

✉ Xin Yang
xinyang@dlut.edu.cn

Jiqing Zhang
jqz@dlmu.edu.com

Bo Dong
bo.dong@princeton.edu

Yingkai Fu
yingkaifu@mail.dlut.edu.cn

Yuanchen Wang
wangyc0604@mail.dlut.edu.cn

Xiaopeng Wei
xpwei@dlut.edu.cn

Baocai Yin
ybc@dlut.edu.cn

[1] Dalian University of Technology, Dalian, China

[2] Dalian Maritime University, Dalian, China

[3] Princeton University, Princeton, USA

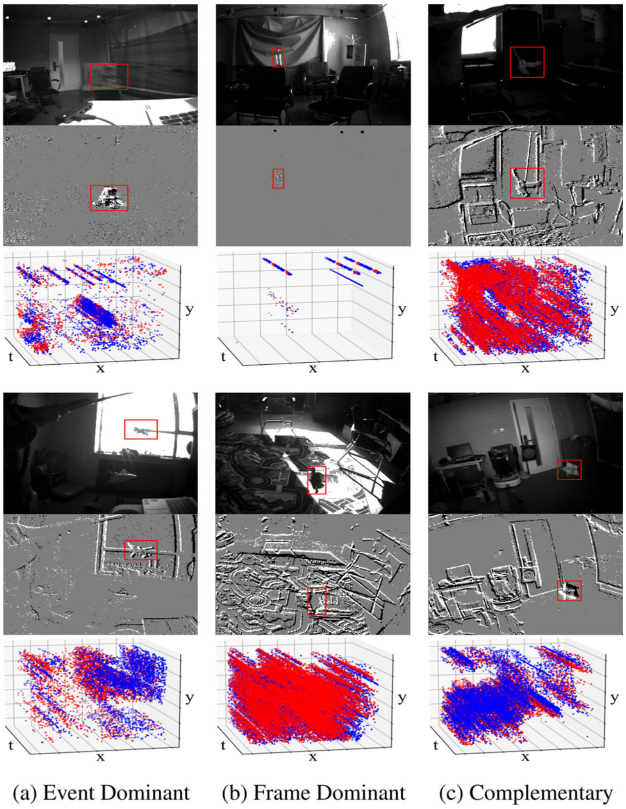


Fig. 1 Conventional frame-based and bio-inspired event-based cameras are complementary. **a** Event-based cameras can provide meaningful information under motion blur and HDR conditions, while frame-based cameras cannot; **b** conventional cameras can offer rich texture information, while event-based cameras suffer from sparse and texture-less cues; **c** both frame and event domains can offer valuable information. Red and blue points in histograms denote "On" and "Off" events, respectively

provide attractive advantages: high temporal resolution (in the order of μs), high dynamic range (140 dB vs. 60 dB), and low power consumption (Gallego et al., 2019). Therefore, encoded events can offer rich temporal cues for object tracking in degraded conditions. However, event-based cameras cannot measure fine-grained texture information like conventional cameras, which is crucial for distinguishing targets from backgrounds. Therefore, both sensors are complementary, Fig. 1 provides examples. The unique complementarity triggers us to introduce event-based information into existing frame-based trackers to leverage the advantages of both the frame and event domains for improving the tracking performance in degraded conditions.

To achieve our purpose, two challenges require to be addressed: (i) Extracting the spatial and temporal cues from event streams is a challenge. Since the asynchronous format of events is quite different from conventional frames, recent works (Gehrig et al., 2019; Messikommer et al., 2020; Wang et al., 2021; Zhu et al., 2019a; Zhang et al., 2021, 2023) aggregate events into frames and then utilize CNN-based methods to digest them. However, these approaches typically ignore the correlations between events occurring at distinct times, which is essential for spatiotemporal prediction. (ii) Effectively fusing the event and conventional frame domains regardless of the diversity of scenes is another challenge. Although multi-modal trackers (i.e., RGB-Thermal and RGB-Depth) (Xiao et al., 2017; Yan et al., 2021b; Zhu et al., 2019b; Zhang et al., 2022) have shown promising potential, unique properties of events prevent the direct application of these methods from providing an effective solution.

In this paper, we propose an event feature extractor and a multi-modal integrator to address the above two challenges, respectively. Specifically, (i) we employ a simple yet effective event accumulation approach to discretize the time domain of asynchronous events. Each discretized time slice can be accumulated into an intensity frame. Based on these aggregated event frames, we further design a novel spatiotemporal feature extractor, termed GM-LSTM, to fully extract the global spatiotemporal features with motion cues of events. The proposed GM-LSTM includes a self-attention scheme to capture temporal features with long-range spatial dependencies and a motion-aware module to enhance the representation of events at different moments, thereby boosting trackers' confidence. (ii) We design a cross-domain modulation and selection module (CDMS) to combine the benefits from the event domain and the frame domain in an efficient and adaptive manner. The effectiveness is enforced by a carefully designed feature enhancement module, which estimates attention from one domain that contributes to the feature expression of another domain. The adaptiveness is maintained by a specially designed weighting scheme to balance the contributions of the two domains, thereby determining which cue is reliable for the target location.

Lack of training data is also a major bottleneck for tracking using event and frame domains. We thus construct a large-scale multi-modality single-object tracking dataset, FE141, which contains 141 sequences with a total length of 2.0 h. FE141 provides ground truth annotations on both the frame and event domains. The annotation frequency is up to 240 Hz. To ensure diversity, we capture videos from diverse real-world scenes that differ significantly in object classes, location, shape, motion, and lighting conditions.

To demonstrate the effectiveness of our proposed method, we extend eight state-of-the-art frame-based trackers: ATOM (Danelljan et al., 2019), DiMP (Bhat et al., 2019), PrDiMP (Danelljan et al., 2020), STARK-S (Yan et al., 2021a), TransT (Chen et al., 2021), TrDiMP (Wang et al., 2021b), SparseTT (Fu et al., 2022) and ToMP (Mayer et al., 2022), to multi-model trackers. Take Fig. 2 as an example, experimental results on our FE141 dataset show that our proposed modules improve the performance of the existing frame-based trackers significantly. The main contributions of this work are four-fold.

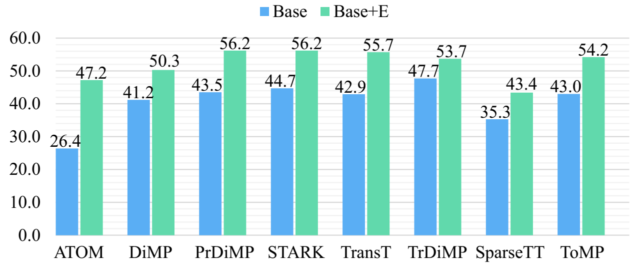


Fig. 2 RSR improvement of extended trackers with our proposed module on the FE141 dataset. 'Base': the original frame-bases trackers; 'Base+E': the extended trackers with our proposed module. The results demonstrate that all frame-based trackers are significantly improved by our proposed approach

- We introduce a novel event-based extractor to capture spatiotemporal features with motion cues and a well-designed cross-domain feature integrator to effectively and adaptively fuse the visual cues from both the frame and event domains.
- Our proposed approach can be readily extended to other frame-based trackers as a plug-in module, significantly boosting their performance.
- We contribute a large-scale frame-event-based dataset for single object tracking. The dataset provides a wide diversity in classes, location, shape, and degraded conditions.
- Experimental results on different datasets demonstrate the effectiveness of our approach.

A preliminary version of this work was presented at ICCV 2021 (Zhang et al., 2021), termed FENet. Compared to the preliminary version, we make several extensions in this work. (i) We propose a novel event-based extractor termed GM-LSTM, which can effectively extract the global spatial and rich temporal features with motion cues from event-based data; (ii) We conduct extensive experiments and validate that our proposed approach can be readily extended to other frame-based trackers as a plug-in module and significantly boost their performance; (iii) We collect 33 additional sequences containing scenes that are especially challenging for the event domain, such as severe camera motion, strobe light, static objects, and so on; (iv) We perform additional experiments and more analysis, including the comparison of state-of-the-art trackers on an additional benchmark VisEvent (Wang et al., 2021c), the comparison of different fusion strategies, computational costs, limitations, and future research directions.

2 Related Work

2.1 Single-domain Object Tracking

2.1.1 Frame-Based Object Tracking

Most of the current trackers leverage conventional frame-based sensors, among which the Siamese-based network has gained significant popularity. As one of the pioneering works, SiamFC Bertinetto et al. (2016) demonstrated that Siamese fully-convolutional deep networks have the ability to use the available data more efficiently for tracking task. Recently, several improvements have been made to enhance each part of the tracking pipeline, such as using deeper and wider backbone networks (Li et al., 2019; Zhang and Peng, 2019), introducing attention (Wang et al., 2018; Gao et al., 2022) and transformer mechanisms (Lin et al., 2022; Fu et al., 2022; Yan et al., 2021a), exploring unsupervised training (Wang et al., 2019, 2021a; Shen et al., 2022), exploiting model update mechanisms (Gao et al., 2019; Zhang et al., 2019), presenting online update schemes (Bhat et al., 2019; Danelljan et al., 2020), and so on.

2.1.2 Event-Based Object Tracking

Compared with the frame-based object tracking methods, only a few attempts have been made to track objects using event-based cameras. Event-based tracking can be generally classified into cluster-based and learning-based trackers. Piatkowska et al. (2012) extended a cluster based on the Gaussian Mixture Model to locate multiple persons in the occurrence of high occlusions. Camuñas-Mesa et al. (2017) proposed a cluster tracking algorithm based on a distance criterion between incoming events and a dynamic list of clusters. Barranco et al. (2018) redefined the well-known mean-shift clustering algorithm using asynchronous events for multi-object tracking. However, these methods rely on strong assumptions or assume restricted conditions. Based on the powerful representation ability of deep learning (Qiao et al., 2022; Ding et al., 2022; Liu et al., 2022; Wang et al., 2022), Chen et al. (2019, 2020) enhanced event representation method Time-Surface (Lagorce et al., 2016) and proposed tracking-by-detection networks for event-based single object tracking. Zhang et al. (2022) introduced spiking neural networks (Wu et al., 2018; Ding et al., 2022; Zhang et al., 2023) to extract the temporal features from asynchronous events for improving event-based tracking performance. Zhu et al. (2022) constructed an end-to-end learning-based paradigm that directly consumes event clouds.

In this paper, we focus on exploiting the complementary information between frame and event domains to improve the robustness of object tracking under degraded conditions.

2.2 Multi-domain Object Tracking

Exploiting the advantages of multiple sensors for robust tracking in challenging scenarios is an intuitive strategy. Thermal images are insensitive to illumination variations, they are thus introduced as a complementary domain to improve performance in extreme conditions like rainy and foggy (Lan et al., 2018; Li et al., 2018; Long Li et al., 2019; Wang et al., 2020; Zhang et al., 2019; Zhu et al., 2019b; Zhang et al., 2022; Hui et al., 2023; Zhang et al., 2023). For example, based on the relevance of attention (Yang et al., 2019a; Qiao et al., 2020; Liu et al., 2021), Zhang et al. (2022) collected a large-scale benchmark and designed a hierarchical multi-modal fusion tracker for visible-thermal UAV tracking. Another line of work (An et al., 2016; Camplani et al., 2015; Kart et al., 2018; Song and Xiao, 2013; Xiao et al., 2017; Yan et al., 2021b; Zhao et al., 2023; Yang et al., 2023) leverages the depth and conventional visible sensors to help solve the occlusion problem in object tracking. Depth cues provide a better object-to-background separation than the conventional frame and simplify reasoning about occlusion. For instance, Lukezic et al. (2019) validated that the performance of baseline RGB trackers can be improved from the straightforward addition of the depth information.

The asynchronous outputs of event-based cameras make combining frame and event domains a distinct challenge compared to the above multi-domain trackers. For example, Yang et al. (2019b) utilized convolution neural networks and spiking neural networks to extract the features of frames and events, respectively. Huang et al. (2018) fused conventional frame sequences at a low framerate and their corresponding high-frequency events for tracking high-speed moving objects. Due to dataset constraints, these methods validate the effectiveness of fusing frames and events within limited scenarios. Therefore in this paper, we first construct a large-scale single-object tracking dataset that provides synchronized conventional frames and raw event streams. Based on this dataset, we verify the effectiveness and generality of our proposed event feature extractor and multi-domain fusion network.

2.3 Spatiotemporal Prediction with LSTM

Sutskever et al. (2014) first proposed an end-to-end multilayer LSTM encoder-decoder framework for machine translation whose input data is one dimension. After that, Srivastava et al. (2015) introduced the LSTM to the field of video representations. However, their approach can only learn the temporal coherence of video sequences. To capture the spatial and temporal features of video frames simultaneously, Shi et al. (2015) proposed the convolutional LSTM (ConvLSTM) network for precipitation nowcasting. Based on this work, some variants have been proposed for achieving impressive results on spatiotemporal prediction. For example, Wang et al. (2017) designed a spatiotemporal LSTM unit that can model spatial and temporal representations in a unified memory cell and convey the memory both vertically across layers and horizontally over states. TrajGRU Shi et al. (2017) leveraged a convolutional layer to learn the receptive area offsets for a special application of precipitation nowcasting. Wu et al. (2021) focused on modelling the within-motion variations to learn the explicit transient variation and remember the motion trend in a unified way. However, due to limited receptive fields, these methods lack the ability to capture long-range spatial dependencies. Moreover, the simple interaction between the previous output state and the present input state ignores the correlations between the two states that are crucial for spatiotemporal prediction. In this paper, we introduce the self-attention mechanism into ConvLSTM (Shi et al., 2015) to effectively extract the global spatial features and temporal features simultaneously. Furthermore, we strengthen the interaction between states at different moments and enhance corresponding features by exploiting the motion information.

3 Methodology

As a supplementary modality, event-based information can efficiently boost the conventional frame-based trackers' outputs and significantly improve the tracking performance. However, incorporating the benefits of events into existing frame-based trackers requires addressing two challenges: (i) An event-based camera reports asynchronous per-pixel brightness changes, simultaneously extracting spatial and temporal information is challenging; (ii) Naively combining event and frame domains ignores circumstance in which one of the domains does not provide meaningful information. In this work, we propose the Event Feature Extractor (EFE; Sect. 3.2) and the Cross-Domain Modulation and Selection module (CDMS; Sect. 3.3) to tackle the above two challenges, respectively.

As shown in Fig. 3, the overall architecture has two branches: the reference branch (top) and the test branch (bottom). The reference and test branches share weights in a siamese style. Each branch has three components, namely: Frame-Feature Extractor (FFE), EFE, and CDMS. In particular, FFE takes a conventional frame as input to extract texture features; EFE extracts the spatial and temporal information

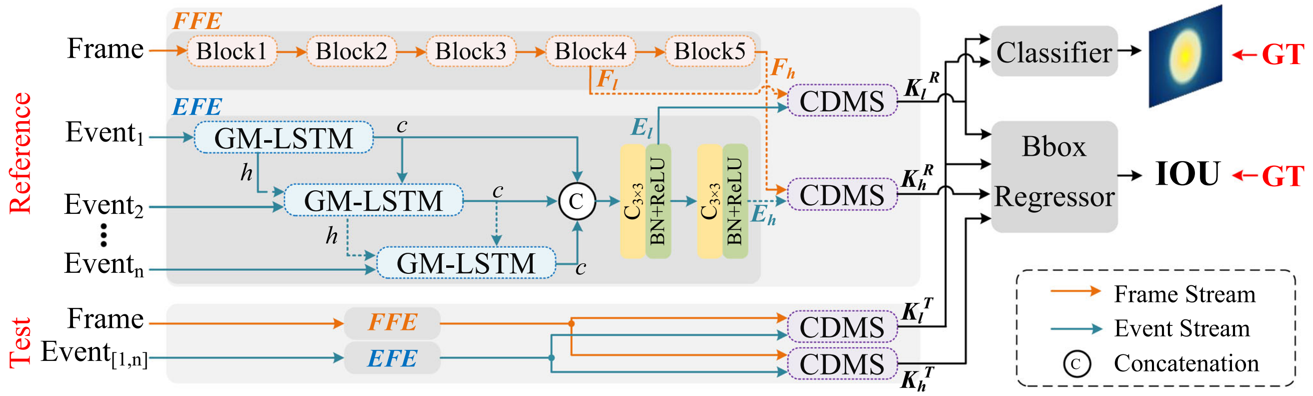


Fig. 3 Overview of the proposed architecture. It has a reference branch and a test branch, and they share the same architecture. Each branch contains a Frame Feature Extractor (FFE) to extract texture features from the conventional frame domain, an Event Feature Extractor (EFE) to extract global spatiotemporal information with motion cues from the event domain, and a Cross-Domain Modulation and Selection module (CDMS) to fuse two domains adaptively

from events captured between the sequential conventional frames. CDMS is responsible for integrating the advantages of the two domains and establishing fusion features (i.e., K^R and K^T). These fused features are then input to the classifier and regressor of base trackers to locate the target on the test frame. To facilitate comprehension, we next detail our proposed approach by using an example of extending the frame-based tracker PrDiMP (Danelljan et al., 2020) to a multi-modal tracker. In PrDiMP, the pretrained ResNet18 (He et al., 2016) is adopted as FFE. The features from the fourth and fifth blocks are employed as low-level and high-level frame features (*i.e.* F_l and F_h), respectively.

3.1 Event Representation

Event-based cameras asynchronously capture log intensity change for each pixel. An event will be triggered when:

$$L(x, y, t) - L(x, y, t - \Delta t) \geq pC, \tag{1}$$

where C denotes the contrast threshold; p is the polarity which means the sign of bright change, with $+1$ and -1 representing the positive and negative events, respectively. Δt is the time since the last event at location $(x, y)^\top$. In a given time interval, a set of events will be triggered:

$$\mathcal{E} = \{e_k\}_{k=1}^{N} = \{[x_k, y_k, t_k, p_k]\}_{k=1}^{N}. \tag{2}$$

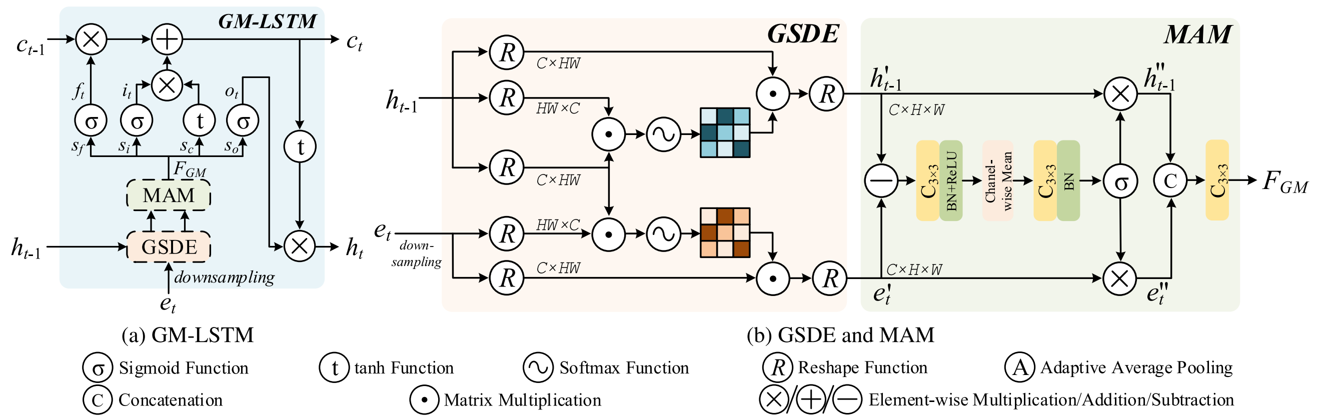
Due to the asynchronous event format differing significantly from the frames captured by conventional frame-based cameras, it is typical to convert the event set into a grid-like representation to use events with convolutional neural networks (Zhang et al., 2023; Mostafavi et al., 2021; Rebecq et al., 2018; Zhou et al., 2023; Zhu et al., 2019a). In this paper, we adopt a simple yet effective mapping between the event set and a grid-based tensor. Specifically, inspired by Zhu et al. (2019a), we first aggregate the events captured between two adjacent frames into an n-bin voxel grid to discretize the time dimension. Then, each 3D discretized slice is accumulated into a 2D frame, where a pixel of the frame records the polarity of the event with the latest timestamp at the pixel's location inside the current slice. Finally, the n generated frames are scaled by 255 for further processing. Given a set of events, $\mathcal{E}^i = \{e_k^i\}_{k=1}^{N_i}$, with the timestamps in the time range of i-th bin, the pixel located at (x, y) on the ith aggregated frame can be defined as follows:

$$\begin{aligned} g(x, y, i) &= \lfloor \frac{p_k^i \times \delta(t(x, y, i)_{max} - t_k^i) + 1}{2} \times 255 \rfloor, \\ t(x, y, i)_{max} &= max(t_k^i \times \delta(x - x_k^i, y - y_k^i)) \\ &\quad \forall t_k^i \in [T_j + (i-1)B, T_j + iB], \end{aligned} \tag{3}$$

where T_j is the timestamp of the jth frame in the frame domain; δ is the Dirac delta function; B is the bin size in the time domain, which is defined as: $B = (T_{j+1} - T_j)/n$. The proposed method leverages the latest timestamp to capture the latest motion cues inside each time slice. Our experimental results show that our used event processing method outperforms other commonly used approaches (see Table 5).

3.2 Event Feature Extractor (EFE)

The purpose of the EFE module is to extract the global spatial and temporal features with motion cues of the event-based data. As shown in Fig. 3, the key component of EFE is GM-LSTM which is a variant of ConvLSTM (Shi et al., 2015). Given a set of events, we first divide them into multiple bins according to Eq. 3. The GM-LSTM then processes each bin



(a) GM-LSTM

(b) GSDE and MAM

Fig. 4 **a** Detailed architectures of the proposed GM-LSTM and its two main building blocks: **b** Global Spatial Dependencies Extractor (GSDE) and Motion Aware Module (MAM)

while keeping the spatial dimension consistent. During the inference, each hidden state and cell state are propagated to the next GM-LSTM in a sequential manner. By fusing the cell state tensors and utilizing another two convolutional layers, we extract different-level event features, E_l and E_h, including both spatial and temporal information.

ConvLSTM (Shi et al., 2015) has achieved impressive results by replacing linear operations with the convolutional layer to capture temporal and spatial dependencies of conventional frames simultaneously. However, for accumulated event frames, it is difficult to extract local features due to sparseness and lack of texture information. Thus, establishing global relationships is essential for the use of event information. Due to the limited receptive field of convolutional layers, ConvLSTM (Shi et al., 2015) tends to suffer from a limited ability to capture long-range spatial dependencies. Besides, in ConvLSTM (Shi et al., 2015), the current input state and previous hidden output state interact through a concatenate operation. This simple interaction ignores the distinction between the two states, each of which contains different target motion, appearance, scene, and association cues across moments that are crucial for spatiotemporal prediction.

To address the above limitations, we replace the concatenation operation and convolutional layers in ConvLSTM with the proposed Global Spatial Dependencies Extractor (GSDE) and Motion Aware Module (MAM) to capture spatiotemporal information with motion cues for tracking, illustrated in Fig. 4a. Our GM-LSTM can be formulated as follows,

$$\begin{aligned} F_{GM} &= MAM(GSDE(\mathcal{D}(e_t), h_{t-1})), \\ s_f, s_i, s_c, s_o &= \mathcal{S}(F_{GM}), \\ f_t &= \sigma\left(s_f\right), i_t = \sigma\left(s_i\right), o_t = \sigma\left(s_o\right), \\ c_t &= f_t \otimes c_{t-1} + i_t \otimes \tanh\left(s_c\right), \\ h_t &= o_t \otimes \tanh\left(c_t\right), \end{aligned} \tag{4}$$

where $GSDE$ and MAM denote the GSDE and MAM modules, respectively; c_t, h_t and e_t represent the cell activation state, the hidden output state, and the input event state, respectively; $\mathcal{D}$ is the downsampling operation on e_t to reduce computation; F_{GM} is the aggregated feature of the current input state e_t and previous hidden output state h_{t-1}; $\mathcal{S}$ means splitting F_{GM} into equally sized chunks (i.e., s_f, s_i, s_c and s_o) in channel dimension; i_t, f_t, and o_t are the input, forget, and output gate at time t, respectively; σ is the Sigmoid function; $\otimes$ indicates element-wise multiplication. We next describe the architecture details of the two key components (i.e., GSDE and MAM) of the proposed GM-LSTM.

3.2.1 Global Spatial Dependencies Extractor (GSDE)

Given the excellent ability to model global dependencies, we introduce a self-attention mechanism into LSTM to extract spatial and temporal information for tracking, generating discriminative features to facilitate object localization. As shown in Fig. 4b, GSDE receives two inputs, the previous hidden output state h_{t-1} and the current input state e_t. Mathematically, for $h_{t-1} \in \mathbb{R}^{C\times H\times W}$,

$$h'_{t-1} = \mathcal{R}^{((C,H,W))}(\varphi(Q_h K_h)V_h), \tag{5}$$

where $\mathcal{R}^{((\cdot))}$ is a reshape function with a target shape $(\cdot)$; $V_h = Q_h = \mathcal{R}^{((C,HW))}(h_{t-1})$, $K_h = \mathcal{R}^{((HW,C))}(h_{t-1})$; φ denotes a softmax function. Similarly, for $e_t \in \mathbb{R}^{C\times H\times W}$,

$$e'_t = \mathcal{R}^{((C,H,W))}(\varphi(Q_e K_e)V_e), \tag{6}$$

where $V_e = \mathcal{R}^{((C,HW))}(e_t)$, $K_e = \mathcal{R}^{((HW,C))}(e_t)$; We believe that the prediction of the current time step can benefit from the past relevant features, we hence set $Q_e = Q_h$. In doing so, global spatial dependencies can be captured during the propagation of stacked LSTM cell layers.

3.2.2 Motion Aware Module (MAM)

Visual tracking relies on motion and temporal context, which motivates us to exploit the discriminating information between the previous output state and the current input state to improve the confidence of our tracker. As shown in Fig. 4b, we first perform a subtraction operation on h'_{t-1} and e'_t to acquire the most discriminative cues between different states, including target motion, appearance, scene. Then, we leverage the discriminative information to conduct a spatial attention scheme to guide h'_{t-1} and e'_t to focus on *where* is an informative part and predict the more distinctive states h''_{t-1} and e''_t. Finally, we concatenate h''_{t-1} and e''_t, and leverage a convolutional layer to generate the aggregated feature F_{GM}. The operation of first subtraction and then concatenation ensures the preservation of previous and present state information while focusing on discriminative cues. Formally, the MAM module is

$$
\begin{aligned}
F_{GM} &= \psi_{3\times3}([h''_{t-1}, e''_t]), &(7)\\
h''_{t-1} &= h'_{t-1} \otimes \kappa, &(8)\\
e''_t &= e'_t \otimes \kappa, &(9)\\
\kappa &= \sigma(\mathcal{B}(\psi_{3\times3}(\mathcal{M}(\gamma(\mathcal{B}(\psi_{3\times3}(h'_{t-1} - e'_t))))))), &(10)
\end{aligned}
$$

where $[\cdot]$ indicates channel-wise concatenation; $\psi_{k\times k}$ means a $k \times k$ convolution layer; $\mathcal{B}$ and γ represent the Batch Normalization (BN) and the ReLU activation function, respectively; σ is the Sigmoid function; κ is the discriminated attention map; $\mathcal{M}$ is channel-wise mean operation.

3.3 Cross-domain Modulation and Selection (CDMS)

Our CDMS is proposed to fuse valuable features from frame and event domains regardless of the diversity of scenes, as shown in Fig. 5. Simply fusing frame and event domains does not necessarily result in the desired improvement in performance. In typical cases, both the texture information provided by frames and the edge information provided by events offer meaningful cues for object tracking. However, in HDR scenes, for example, the frame domain cues may be weak or nonexistent, offering no valuable cues, while the event domain cues remain robust. Similarly, event-based cameras cannot successfully record object-related information in the absence of motion, whereas conventional frames can still provide rich texture features. Therefore, effectively and dynamically fusing between and within the multi-domain information is essential for robust object tracking.

As shown in Fig. 5, we first design a Cross-Attention Block (CAB) to complement one domain with information from another domain. Specifically, given the extracted frame features F and event features E, we define the following cross-domain attention scheme to generate an enhanced feature for E:

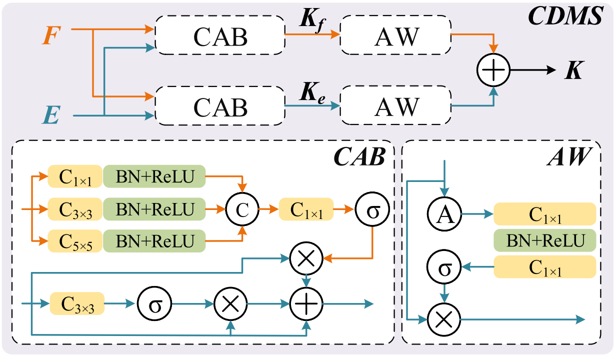


Fig. 5 Detailed architectures of the proposed cross-domain modulation and selection module (CDMS)

$$
\begin{aligned}
K_e &= K_e^1 + K_e^2 + E, &(11)\\
K_e^1 &= \sigma(\psi_{3\times3}(E)) \otimes E, &(12)\\
K_e^2 &= \sigma(\psi_{1\times1}[\gamma(\mathcal{B}(\psi_{1\times1}(F))),\\
&\quad \gamma(\mathcal{B}(\psi_{3\times3}(F))), \gamma(\mathcal{B}(\psi_{5\times5}(F)))]) \otimes E, &(13)
\end{aligned}
$$

where K_e^1 indicates a self-attention based on E; K_e^2 is a cross-domain attention scheme based on F to guide the feature of E; K_e denotes the enhanced event feature. Similarly, we can generate the enhanced frame feature K_f, directed by the event domain. Based on K_e and K_f, we further propose an adaptive weighted balance scheme (AW) to balance the contribution of the frame and event domains:

$$
\begin{aligned}
K &= w_f K_f + w_e K_e, &(14)\\
w_f &= \sigma(\psi_{1\times1}(\gamma(\mathcal{B}(\psi_{1\times1}(\mathcal{A}(K_f)))))), &(15)\\
w_e &= \sigma(\psi_{1\times1}(\gamma(\mathcal{B}(\psi_{1\times1}(\mathcal{A}(K_e)))))), &(16)
\end{aligned}
$$

where $\mathcal{A}$ is the adaptive average pooling.

3.4 Classifier and Bounding Box (BBox) Regressor

To enhance the generality, we do not modify the classifier, BBox regressor, and loss function of existing frame-based trackers. Taking PrDiMP (Danelljan et al., 2020) as an example, the BBox regressor contains an IoU modulation and an IoU predictor. The IoU modulation first maps K_l^R and K_h^R to different level modulation vectors v_l and v_h, respectively. Mathematically, the mapping is achieved as follows:

$$
\begin{aligned}
&v_l = \mathcal{F}(q), \quad v_h = \mathcal{F}(q),\\
&q = [\mathcal{F}(\mathcal{P}(\psi_{3\times3}(K_l^R), B^r)), \mathcal{P}(\psi_{3\times3}(K_h^R), B^r)],
\end{aligned} \qquad (17)
$$

where $\mathcal{F}$ is the fully connected layer; $\mathcal{P}$ denotes PrPool (Jiang et al., 2018); B^r is the target bounding box from the reference

Table 1 Analysis of existing intensity frame-based and event-based datasets for object tracking

	Classes	Frames	Events	Time	Frame(Hz)	Event(Hz)	Sensors
OTB-2013 (Wu et al., 2013)	10	29K	–	16.4m	30	–	Intensity
OTB-2015 (Wu et al., 2015)	16	59K	–	32.8m	30	–	Intensity
TC-128 (Liang et al., 2015)	27	55K	–	30.7m	30	–	Intensity
VOT-2014 (Kristan, 2014)	11	10K	–	5.7m	30	–	Intensity
VOT-2017 (Kristan, 2017)	24	21K	–	11.9m	30	–	Intensity
NUS-PRO (Li et al., 2015)	8	135K	–	75.2m	30	–	Intensity
UAV123 (Mueller et al., 2016)	9	113K	–	62.5m	30	–	Intensity
UAV20L (Mueller et al., 2016)	5	59K	–	32.6m	30	–	Intensity
GOT-10k (Huang et al., 2019)	563	1.5M	–	–	10	–	Intensity
LaSOT (Fan et al., 2021)	70	3.5M	–	32.5h	30	–	Intensity
EED (Mitrokhin et al., 2018)	2	0.2K	3.4M	10.2s	23	23	Intensity + Event
EV-IMO (Mitrokhin et al., 2019)	3	76K	–	32.0m	40	200	Intensity + Event
VisEvent (Wang et al., 2021c)	18	371K	–	–	25	25	Intensity + Event
Ours	21	251K	6.7G	119.8m	10/15/20/40	240	Intensity + Event

branch. Next, the IoU predictor predicts IoU based on the following equation:

$$IoU = \mathcal{F}([\mathcal{F}(\mathcal{P}(\psi_{3\times3}(\psi_{3\times3}(K_l^T)), B^t) \otimes v_l), \mathcal{F}(\mathcal{P}(\psi_{3\times3}(\psi_{3\times3}(K_h^T)), B^t) \otimes v_h)]), \tag{18}$$

The classifier predicts a target confidence score. It first maps K_l^R and B^r to an initial filter, which is then optimized by the optimizer. The optimizer uses the steepest descent methodology to obtain the final filter. The final filter is used as the convolutional layer's filter weight and applied to K_l^T to discriminate between the target object and background distractors robustly. The network is trained by minimizing the Kullback–Leibler divergence between the predicted and the label distribution. We refer to Danelljan et al. (2020) for details.

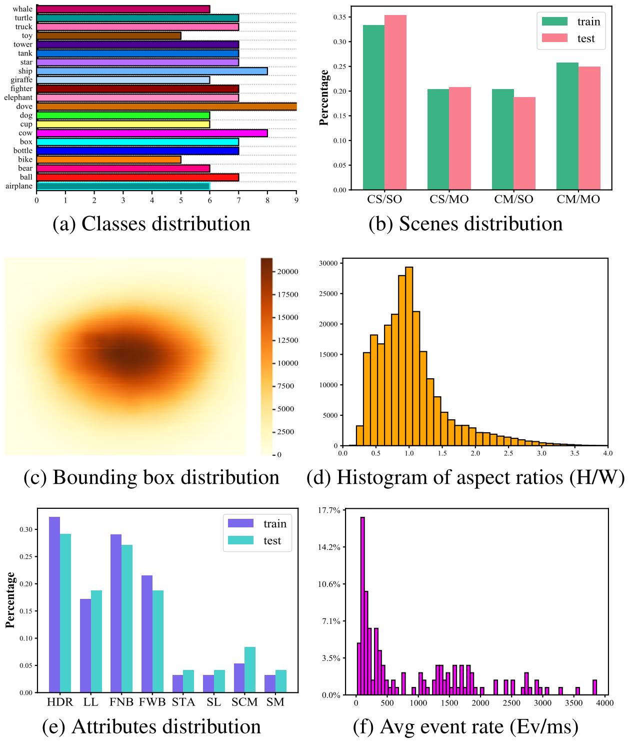
4 FE141 Dataset

4.1 Dataset Collection and Annotation

Our FE141 dataset is simultaneously recorded by a DAVIS346 camera and the Vicon motion capture system (https://www.vicon.com/). The DAVIS346 equips with a 346 × 260 pixels dynamic vision sensor (DVS) and a frame-based active pixel sensor (APS), and it can capture both events and grayscale frames simultaneously. Since event-based cameras only use power to process changing pixels, the power consumption is significantly lower than conventional cameras (i.e., ≤ 100 mW vs. ≥ 3W). The Vicon system can provide the 3D position and trajectory of targets with a high sample rate and sub-millimeter precision through 12 Vero motion capture infrared cameras. Since the Vicon system employs active sensing to track objects, the infrared light emitted from the system becomes noise in the events domain. To deal with it, we place an infrared filter in front of the DAVIS346 to filter out the light with a wavelength above 700 nm. We set the sampling rate of the DAVIS346 camera's APS to 10/15/20/40 Hz, and the sampling rate of the Vicon to 240 Hz. The data annotation is accomplished through 3D projection from Vicon to the DAVIS346 event camera, and we refer to Mitrokhin et al. (2019) for more details.

4.2 Dataset Statistics

Compared with RGB-based tracking benchmarks, a few attempts have been made at event-based tracking datasets. By recording monitors with neuromorphic vision sensors, Hu et al. (2016) converted existing RGB benchmark datasets to the DVS-based. However, this setting disregards the advantages of high temporal resolution and high dynamic range of event-based cameras, preventing the recording of motion information between consecutive frames and helpful cues in HDR scenes. Mitrokhin et al. (2018, 2019) presented two event-based tracking datasets in real scenes: EED (Mitrokhin et al., 2018) and EV-IMO (Mitrokhin et al., 2019). As shown in Table 1, the EED only has 234 frames (10.2 s) with two types of objects. EV-IMO offers a better package with motion masks and high-frequency events annotations, up to 200Hz. But, similar to EED, limited object types block it to be used practically. Recently, Wang et al. (2021c) proposed an RGB-Event tracking dataset, termed VisEvent, but the annotation frequency of events on this dataset is only 25Hz. In addition, partial data in VisEvent has a mismatched timestamp or lacks the raw events, resulting in the availability of just incomplete



(a) Classes distribution

(b) Scenes distribution

(c) Bounding box distribution

(d) Histogram of aspect ratios (H/W)

(e) Attributes distribution

(f) Avg event rate (Ev/ms)

Fig. 6 Statistics of FE141 dataset in terms of **a** classes, **b** scenes, **c** bounding box position, **d** aspect ratios (H/W), **e** attributes, and **f** average event rate (Ev/ms)

data. To address this lack of training data with the high annotated rate for multi-modal learning with events, we collect a large-scale dataset termed FE141, which has 141 sequences with a total length of 2.0 h. We also provide the distribution differences compared to existing intensity image-based object tracking datasets in Table 1, including OTB-2013 (Wu et al., 2013), OTB-2015 (Wu et al., 2015), TC-128 (Liang et al., 2015), VOT-2014 (Kristan, 2014), VOT-2017 (Kristan, 2017), NUS-PRO (Li et al., 2015), UAV123 (Mueller et al., 2016), UAV20L (Mueller et al., 2016), GOT-10k (Huang et al., 2019), and LaSOT (Fan et al., 2021). We further present statistics of FE141 from multiple perspectives to highlight its diversity.

4.2.1 Object and Scene Category

We aim to build a highly diverse dataset with sufficient objects and scene classes. As shown in Fig. 6a, our FE141 dataset includes 21 different object classes, covering most categories in real applications. These objects can be divided into three categories: animals, vehicles, and daily goods (*e.g.,* bottle, box). Due to the sensitivity of event-based cameras to movement, we construct our dataset under scenes with different motions. Specifically, according to the camera movement and the number of objects, as shown in Fig. 6b, FE141 has four types of scenes: static shots with a single object or multiple objects (CS/SO and CS/MO); dynamic shots with a single object or multiple objects (CM/SO and CM/MO).

4.2.2 Annotated Bounding Box Statistics

To investigate the location distribution of bounding boxes in FE141, we plot the distribution of all annotated bounding box locations. As shown in Fig. 6c, the overall bounding box distribution tends to be centered. We further present the distribution of the bounding box aspect ratios (i.e., height over width), see Fig. 6d. It demonstrates that the bounding boxes of the FE141 dataset have various shapes.

4.2.3 Attributes Definition

As shown in Fig. 6e, we define eight attributes in our FE141 dataset: high dynamic range (HDR), low-light (LL), fast motion without and with motion blur on APS frame (FWB and FNB), static object (STA), scenes illuminated with a strobe light (SL), severe camera motion (SCM), and objects similar to the object being tracked (SM). The first four are difficult for conventional frame-based tracking, whereas the latter four are challenging for event-based tracking.

4.2.4 Event Rate

To analyze the properties of the raw events stream, we compute the rate of event stream generated during the recordings as follows: for the event stream of each sequence, we first discretize its time dimension into $1ms$ intervals. We then count the number of events of each interval and calculate the average number of events of all intervals as the event rate. The distribution of the event rate in Fig. 6f evidences the motion diversity is pretty wide.

5 Experiments

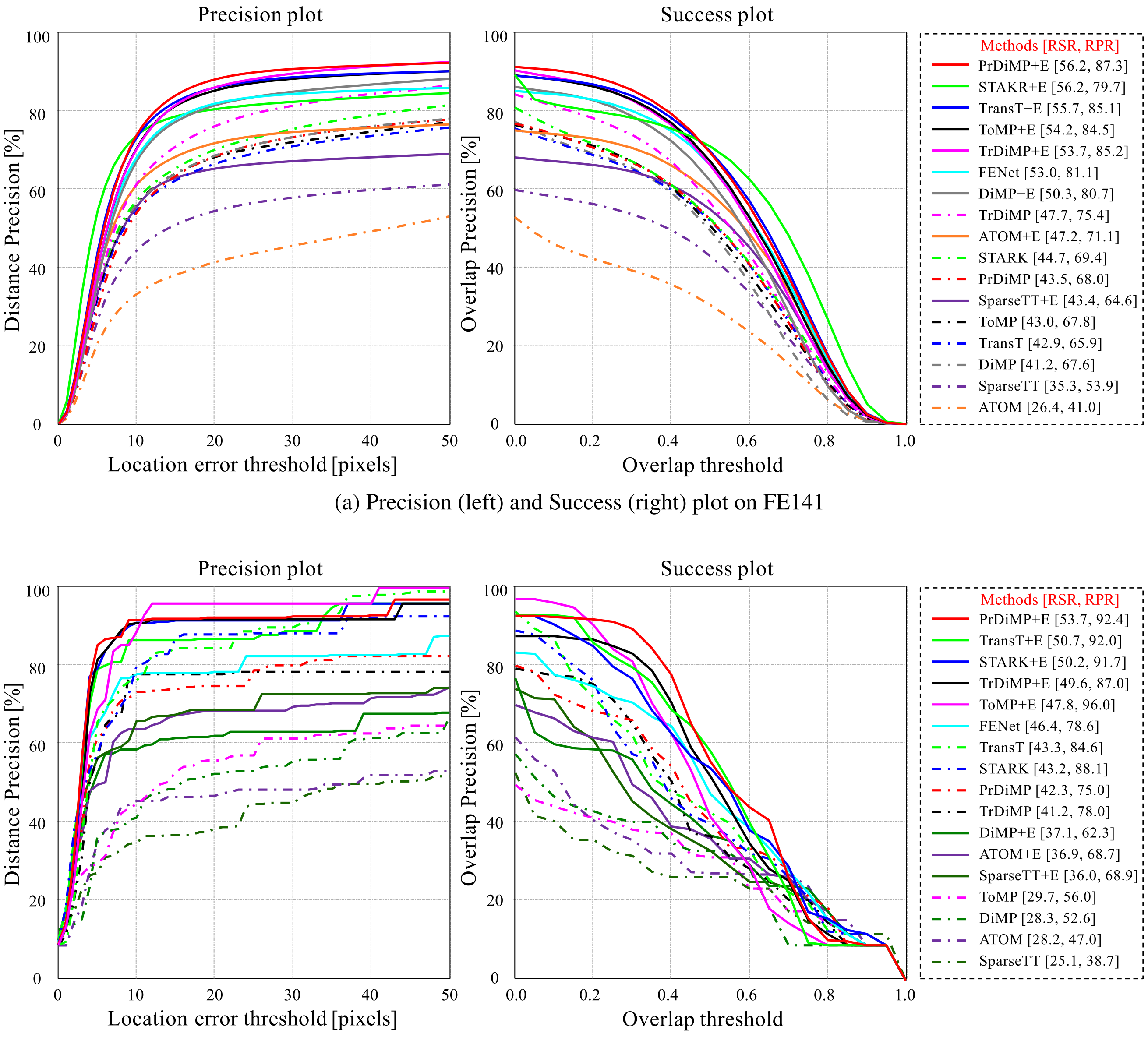
5.1 Experimental Settings

5.1.1 Implementation Details

We implement the proposed network in PyTorch (Paszke et al., 2019). For different extended multi-modal trackers, we adopt the same training strategies of corresponding original frame-based trackers, including the learning rate, the optimizer, etc. All approaches are trained on a 20-core i9-10900K 3.7 GHz CPU, 64 GB RAM, and an NVIDIA RTX3090 GPU.

5.1.2 Dataset

We evaluate our proposed approach on three event-frame-based datasets: FE141, EED (Mitrokhin et al., 2018), and



(a) Precision (left) and Success (right) plot on FE141

(b) Precision (left) and Success (right) plot on EED [53]

Fig. 7 Precision (left) and Success (right) plot on FE141 and EED datasets. In terms of both metrics, the extended trackers with our approach outperform corresponding original frame-based trackers by a large margin

VisEvent (Wang et al., 2021c). Our FE141 dataset contains 141 sequences, of which 93 are utilized for training and 48 are used for testing. The EED (Mitrokhin et al., 2018) is only used for evaluation, it provides five challenging sequences: fast drone, light variations, occlusions, what is background, and multiple objects. The first two sequences both record a fast-moving drone under low illumination. The third and the fourth sequences record a moving ball with another object and a net as foreground, respectively. The fifth sequence consists of multiple moving objects under normal lighting conditions. Compared with the FE141 and EED datasets, VisEvent (Wang et al., 2021c) provide RGB frames instead of grayscale frames. Following Zhang et al. (2022), we filter sequences that miss raw event streams or have misaligned timestamps, leaving 205 sequences for training and 172 for testing.

5.1.3 Evaluation Metric

To show the quantitative performance of each tracker, we utilize three widely used metrics: success rate (SR), precision rate (PR), and overlap precision (OP_T). These metrics represent the percentage of three particular types of frames. SR cares the frame of that overlap between ground truth and predicted bounding box is larger than a threshold; PR focuses on the frame of that the center distance between ground truth and predicted bounding box within a given threshold; OP_T represents SR with T as the threshold. For SR, we employ the area under curve (AUC) of an SR plot as representative SR (RSR). For PR, we use the PR score associated with a 20-pixel threshold as representative PR (RPR).

5.2 Plug-in Module

Our proposed GM-LSTM for extracting event features and CDMS for fusing two domains can be readily extended to the existing frame-based tracking approaches. To validate the effectiveness and generality of our proposed approach, we plug the GM-LSTM and CDMS modules into eight state-of-the-art frame-based trackers: ATOM (Danelljan et al., 2019), DiMP (Bhat et al., 2019), PrDiMP (Danelljan et al., 2020), STARK-S (Yan et al., 2021a), TransT (Chen et al., 2021), TrDiMP (Wang et al., 2021b), SparseTT (Fu et al., 2022) and ToMP (Mayer et al., 2022).

5.2.1 Comparisons on the FE141 Dataset

As shown in Fig. 7a, on the FE141 dataset, all our extended multi-modal approaches outperform corresponding original frame-based trackers by a large margin in terms of both precision and success rate. For example, the extended ATOM (Danelljan et al., 2019) with our approach outperforms the base ATOM by 20.8% and 30.1% in terms of RSR and RPR, respectively; The extended PrDiMP (Danelljan et al., 2020) with our method achieves 56.2% and 87.3% overall RSR and RPR, outperforming the original frame-based model by 12.7% and 19.3%, respectively. The experimental results demonstrate the effectiveness and generality of our approach. To provide a more comprehensive comparison, we compare our extended trackers to two event-only methods: (i) Only using events and EFE for tracking, termed Only-Event; (ii) STNet (Zhang et al., 2022), a state-of-the-art event-based tracker that combines Swin-Transformer (Liu et al., 2021) and Spiking Neural Networks to extract spatial and temporal features to improve tracking performance. We also compare our extended trackers to two frame-event multi-modality methods: (i) FENet (Zhang et al., 2021), our preliminary version of this paper; (ii) ViPT (Jiawen et al., 2023), a state-of-the-art multi-modal tracking method introducing the prompt-learning ideology.

To further validate the effectiveness of our multi-domain fusion, we also show the performances under four different challenging conditions, including high dynamic range (HDR), low light (LL), fast motion with blur (FWB), and fast motion without blur (FNB), which are extremely challenging for the frame domain. As shown in Table 2, our extended methods offer better results than the corresponding original methods under all four conditions, especially in HDR and LL conditions. For example, in HDR condition, the extended STARK-S (Yan et al., 2021a) with our approach outperforms the base STARK-S by 14.7% and 15.6% in terms of RSR and RPR, respectively; In the LL condition, the extended TransT (Chen et al., 2021) with our modules outperforms the original TransT by 27.0% and 38.8% in terms of RSR and RPR, respectively. The results demonstrate that our extended multi-modal trackers can effectively extract and leverage the information provided by the event domain. To estimate the effect of the frame domain in multimodality, we report the tracking performance on FE141 under another four degraded conditions that reduce the quality of the event data: (a) scenes with objects similar to the object being tracked (SM); (b) severe camera motion (SCM); (c) scenes illuminated with a strobe light (SL); and (d) static object (STA). As shown in Table 3, all extended trackers outperform STNet except extended ATOM (Danelljan et al., 2019) and SparseTT (Fu et al., 2022). For example, the extended PrDiMP (Danelljan et al., 2020) achieves a 56.2% overall RSR and 87.3% RPR, outperforming the STNet by 8.0% and 11.6%, respectively. It shows that the frame domain indeed improves the robustness of tracking.

The above results illustrate that our method can still locate targets effectively with another modality even if one modality lacks available information. Multiple visual examples under different degraded conditions are shown in Fig. 10, where we can see our extended approaches can effectively track the target under all conditions. We further provide additional qualitative comparisons of extended trackers with our modules compared to base trackers under different conditions in the supplemental video. The supplemental video is available at https://youtu.be/ul-8poOPgs8.

5.2.2 Comparisons on the EED Dataset

Although EED has very limited frames and corresponding events, the sequences it provides are still challenging for object tracking. As shown in Fig. 7b, we can see that extended trackers with our proposed modules improve the base trackers in terms of RSR significantly.

5.2.3 Comparisons on the VisEvent Dataset

To confirm that our method maintains effectiveness on RGB frames and associated events, we further train and validate our approach on the VisEvent dataset. As shown in Table 4, the extended trackers with our modules still offer better performance than base trackers in both accuracy. These results again indicate the effectiveness and generality of our proposed network on RGB-event-based dataset. We also provide more visual results in the supplemental video.

Table 2 State-of-the-art comparison on the FE141 dataset under four challenging conditions: high dynamic range (HDR), low light (LL), fast motion with blur (FWB), and fast motion without blur (FNB)

Methods	HDR				LL				FWB				FNB				ALL			
	RSR	$OP_{0.50}$	$OP_{0.75}$	RPR	RSR	$OP_{0.50}$	$OP_{0.75}$	RPR	RSR	$OP_{0.50}$	$OP_{0.75}$	RPR	RSR	$OP_{0.50}$	$OP_{0.75}$	RPR	RSR	$OP_{0.50}$	$OP_{0.75}$	RPR
ATOM (Danelljan et al., 2019)	13.4	12.8	3.6	20.0	37.2	38.3	3.9	60.6	34.1	44.5	16.0	50.7	46.8	58.2	24.7	70.5	26.4	30.2	10.6	41.0
DiMP (Bhat et al., 2019)	32.6	36.6	9.8	52.5	35.0	34.2	4.4	55.8	44.2	58.2	16.1	69.9	55.1	71.1	21.1	86.3	41.2	49.0	12.1	65.9
PrDiMP (Danelljan et al., 2020)	37.9	45.2	16.0	55.3	45.9	48.2	11.7	65.6	46.1	58.5	18.7	73.7	56.0	71.2	30.3	84.2	43.5	51.9	17.3	68.0
STARK-S (Yan et al., 2021a)	49.4	57.8	28.3	71.2	41.6	46.1	10.6	61.3	46.8	59.6	22.7	71.5	59.0	73.9	34.2	90.6	44.7	50.0	20.1	69.4
TransT (Chen et al., 2021)	47.0	56.2	22.6	68.1	40.8	40.9	8.4	58.8	43.2	56.9	21.0	65.6	58.3	72.8	30.1	88.0	42.9	51.6	18.2	65.9
TrDiMP (Wang et al., 2021b)	50.1	60.0	24.8	74.7	45.5	49.6	8.9	71.3	46.5	59.5	20.9	72.2	59.8	72.6	30.2	92.4	47.7	56.1	19.1	75.4
SparseTT (Fu et al., 2022)	43.7	50.5	21.9	62.9	29.0	36.0	6.1	43.9	38.8	49.7	18.1	60.4	56.2	70.2	31.0	84.1	35.3	42.6	15.8	53.9
ToMP (Mayer et al., 2022)	49.2	59.2	24.3	71.6	46.4	51.2	8.0	71.7	44.3	54.9	18.6	72.3	58.9	72.9	30.5	90.0	43.0	50.3	17.1	67.8
Only Event	47.4	56.4	16.5	75.0	63.1	78.7	33.3	95.5	46.2	55.5	15.4	79.1	44.6	48.4	17.1	72.8	47.3	59.1	19.9	74.5
STNet (Zhang et al., 2022)	54.8	67.0	27.7	82.3	62.6	81.1	25.4	94.1	47.0	58.2	30.6	70.2	59.7	73.6	37.7	87.9	48.2	57.9	22.8	75.7
FENet (Zhang et al., 2021)	59.9	75.1	33.8	85.8	65.9	89.2	36.1	95.4	55.7	67.6	29.2	87.3	62.4	78.2	37.7	**93.4**	53.0	65.7	26.0	81.1
ViPT (Jiawen et al., 2023)	58.8	74.2	32.2	85.1	**70.5**	94.3	44.5	99.0	51.6	66.7	26.5	77.8	**65.7**	**84.8**	43.9	92.4	56.1	**72.3**	31.2	82.0
ATOM + E	49.9	63.9	28.7	70.3	69.0	89.5	44.7	98.1	56.6	68.7	29.7	89.3	55.1	70.7	36.2	78.6	47.2	58.6	24.7	71.1
DiMP + E	54.5	68.3	19.1	82.8	67.5	92.2	31.0	97.6	54.0	62.7	16.9	94.4	60.1	78.0	24.2	92.7	50.3	61.7	18.0	80.7
PrDiMP + E	60.9	76.4	34.1	87.1	69.3	92.6	41.8	**99.2**	**57.8**	68.2	29.2	**95.0**	62.8	78.3	38.4	92.6	**56.2**	68.8	27.2	**87.3**
STARK-S + E	**64.1**	**79.8**	**47.0**	86.8	68.7	86.8	**59.0**	89.4	57.6	**73.2**	**37.4**	84.9	63.8	80.4	**44.9**	89.5	**56.2**	70.5	**37.8**	79.7
TransT + E	61.7	77.2	36.0	**87.9**	67.8	89.6	38.1	97.6	55.5	68.7	28.1	87.9	62.5	79.3	38.7	91.1	55.7	69.1	28.0	85.1
TrDiMP + E	54.5	66.8	26.5	81.0	65.6	88.2	28.7	98.2	54.7	65.3	21.6	91.5	61.4	75.7	35.8	92.4	53.7	65.6	22.5	85.2
SparseTT + E	53.0	65.1	27.5	75.7	69.0	**94.7**	35.6	98.2	44.0	54.8	24.4	67.5	61.8	77.0	37.3	91.1	43.4	54.4	22.6	64.6
ToMP + E	58.2	72.5	30.8	83.6	65.7	89.1	30.9	96.5	55.9	66.2	25.8	92.1	61.7	78.2	35.3	91.2	54.2	66.6	24.3	84.5

Bold values indicate the best result

Table 3 State-of-the-art comparison on the FE141 dataset under four degraded conditions: scenes with objects similar to the object being tracked (SM), severe camera motion (SCM), scenes illuminated with a strobe light (SL), and static object (STA)

Methods	SM				SCM				SL				STA				ALL			
	RSR	$OP_{0.50}$	$OP_{0.75}$	RPR	RSR	$OP_{0.50}$	$OP_{0.75}$	RPR	RSR	$OP_{0.50}$	$OP_{0.75}$	RPR	RSR	$OP_{0.50}$	$OP_{0.75}$	RPR	RSR	$OP_{0.50}$	$OP_{0.75}$	RPR
ATOM (Danelljan et al., 2019)	15.8	15.7	6.8	24.1	27.2	29.7	10.8	42.4	12.6	9.6	2.5	20.9	6.0	4.3	0.6	29.9	26.4	30.2	10.6	41.0
DiMP (Bhat et al., 2019)	54.1	69.0	17.4	88.7	43.1	46.5	8.6	74.9	34.7	33.0	4.6	67.0	28.3	34.3	7.4	49.4	41.2	49.0	12.1	65.9
PrDiMP (Danelljan et al., 2020)	42.0	53.5	23.7	62.4	46.9	52.3	16.2	**77.4**	23.5	19.8	5.0	49.9	27.7	30.4	4.8	50.2	43.5	51.9	17.3	68.0
STARK-S (Yan et al., 2021a)	38.9	46.3	17.5	61.8	41.4	43.8	13.0	70.2	19.3	14.1	2.0	38.4	21.7	16.7	1.2	48.0	44.7	50.0	20.1	69.4
TransT (Chen et al., 2021)	38.7	44.9	19.0	60.9	35.7	43.0	12.0	56.3	19.5	14.9	2.4	43.8	28.6	34.8	6.3	49.7	42.9	51.6	18.2	65.9
TrDiMP (Wang et al., 2021b)	52.8	66.3	21.5	85.7	44.4	47.0	12.7	72.5	20.6	15.4	3.7	38.7	31.0	25.2	3.3	65.2	47.7	56.1	19.1	75.4
SparseTT (Fu et al., 2022)	34.4	42.3	14.3	55.9	30.5	36.8	10.3	49.9	10.5	7.1	1.4	19.2	3.9	2.5	0.1	9.0	35.3	42.6	15.8	53.9
ToMP (Mayer et al., 2022)	32.4	37.3	11.6	54.7	39.1	41.0	11.8	63.0	15.1	8.2	1.1	30.5	21.3	21.5	4.1	41.5	43.0	50.3	17.1	67.8
Only-Event	17.2	21.3	6.9	26.1	32.4	28.1	9.4	48.7	17.7	6.8	2.8	24.1	27.1	30.2	7.1	45.1	47.3	59.1	19.9	74.5
STNet (Zhang et al., 2022)	21.9	29.9	6.9	41.1	38.3	30.8	11.5	52.8	16.3	18.2	3.6	33.2	27.5	32.3	7.3	48.6	48.2	57.9	22.8	75.7
FENet (Zhang et al., 2021)	35.8	39.5	11.6	65.1	40.5	41.7	13.5	66.4	22.2	26.7	5.2	39.2	39.7	48.3	8.3	70.1	53.0	65.7	26.0	81.1
ViPT (Jiawen et al., 2023)	54.4	70.7	35.7	77.7	**48.3**	**62.2**	**20.5**	73.8	20.9	25.4	8.0	33.8	56.0	**74.8**	21.0	86.5	56.1	**72.3**	31.2	82.0
ATOM + E	21.1	24.1	11.3	35.1	31.6	33.4	12.6	47.3	17.0	18.0	6.2	28.3	55.9	71.1	15.4	94.8	47.2	58.6	24.7	71.1
DiMP + E	27.5	34.0	6.5	46.8	33.3	38.3	8.3	54.8	24.4	14.5	2.8	53.3	27.4	29.7	4.8	51.1	50.3	61.7	18.0	80.7
PrDiMP + E	57.3	70.9	28.3	**88.8**	42.3	44.7	15.3	69.0	33.5	36.0	6.1	63.4	55.4	65.0	16.9	94.8	**56.2**	68.8	27.2	**87.3**
STARK-S + E	**62.1**	**76.5**	**53.9**	81.6	43.5	53.4	20.2	65.2	24.9	28.6	**9.5**	42.5	41.6	54.7	**25.0**	59.2	**56.2**	70.5	**37.8**	79.7
TransT + E	54.5	67.5	26.0	84.3	43.9	49.4	16.9	70.3	33.5	**37.0**	8.2	61.6	52.7	65.0	17.4	87.5	55.7	69.1	28.0	85.1
TrDiMP + E	49.0	61.2	22.9	77.3	38.8	41.8	14.9	62.1	**35.9**	35.2	6.6	**69.2**	**57.7**	72.0	14.3	**98.3**	53.7	65.6	22.5	85.2
SparseTT + E	17.2	20.4	7.3	27.8	22.3	26.5	10.4	34.3	19.0	22.1	4.0	35.1	29.1	34.0	8.0	52.2	43.4	54.4	22.6	64.6
ToMP + E	43.4	54.9	17.4	69.6	42.5	46.0	14.4	67.7	31.8	30.5	5.3	62.6	51.4	62.6	14.6	88.0	54.2	66.6	24.3	84.5

Bold values indicate the best result

5.3 Ablation Study

In concert with Sect. 3, our ablation experiments are based on the extended multi-modal tracker PrDiMP (Danelljan et al., 2020).

5.3.1 Impact of Multi-modal Input

To demonstrate the efforts of multi-modal input on object tracking, we conduct the following two experiments: (i) Frame only: only using frames and **FFE**; (ii) Event only: only using events and **EFE**. As shown in the rows *A* and *B* of Table 5, when using only frame or event as input, the RSR/RPR scores are 43.5%/68.0% and 39.7%/64.5%, respectively. These results are significantly worse than tracking results with multi-modal inputs (the row *R* of Table 5), which validates the effectiveness of multi-modal fusion for tracking in degraded conditions.

5.3.2 Effectiveness of GM-LSTM

The proposed GSDE and MAM are the two key components in GM-LSTM. To verify their effectiveness, we modify the original model by dropping each of the components and retrain the modified models. Correspondingly, we obtain four retrained models: (i) without GM-LSTM; Inside GM-LSTM, (ii) without GSDE; (iii) without MAM; (iv) replacing GSDE and MAM with a convolutional layer (i.e., ConvLSTM (Shi et al., 2015)). The corresponding experimental results are shown in the rows from *C* to *F* of Table 5. We can see that when the GM-LSTM is removed, the RSR and RPR drop significantly by 5.1% and 9.1%, respectively. This illustrates that the temporal information plays a key role in our proposed tracker. The performance of RSR drops by 2.0% when the GSDE is removed, as shown in the row *D*, which demonstrates capturing long-range spatial dependencies can boost the feature expression ability of LSTM. When the MAM is removed, the RSR and RPR drop by 1.4% and 2.4% as shown in the row *E*. It suggests discriminating information between the different states can improve the confidence of our tracker. The effectiveness of the proposed GSDE and MAM is again illustrated by comparing rows *D* to *F*, where the performance decreases further when both the GSDE and MAM are removed.

In our setting, the query vectors Q_e and Q_h in GSDE are the same, and both are estimated from h_{t-1}. To make our design more convincing, we conduct the following two experiments: (i) Q_e and Q_h are estimated from e_t and h_{t-1}, respectively; (ii) Q_e and Q_h are the same, and both are estimated from e_t. When Q_e and Q_h are not shared between e_t and h_{t-1}, the RSR and RPR drop by 2.1% and 4.1% as shown in the row *G* of Table 5. It suggests that the interaction between the current input state and the previous output state strengthens the confidence of our tracker. When Q_e and Q_h are both estimated from e_t, as shown in the second row, $\mathrm{OP}_{0.75}$ is higher than ours. However, our setting outperforms this setting in the other three metrics, 1.1% in RSR, 0.2% in $\mathrm{OP}_{0.50}$, and 3.3% in RPR. We argue that this is due to the fact that h_{t-1} is generated by multiple iterations in GM-LSTM, which stores richer temporal information than e_t.

Table 4 Comparisons on the VisEvent (Wang et al., 2021c) dataset

Methods		RSR	RPR	Methods		RSR	RPR
ATOM	Base	38.4	52.4	TransT	Base	44.9	65.0
	+ E	45.3	61.3		+ E	51.8	69.4
DiMP	Base	41.4	56.3	TrDiMP	Base	42.9	63.3
	+ E	47.1	63.4		+ E	50.5	67.6
PrDiMP	Base	42.1	55.0	SparseT	Base	40.8	53.9
	+ E	48.5	62.0		+ E	43.0	63.5
STARK-S	Base	40.8	55.9	ToMP	Base	43.7	57.3
	+ E	46.7	61.6		+ E	49.6	67.0

5.3.3 Effectiveness of CDMS

We investigate the impact of the proposed CDMS module by removing it and its components from the extended tracker. There are three key components in CDMS: self-attention (Eq. 12), cross-attention (Eq. 13), and adaptive weighting (Eq. 14). Correspondingly, we conduct four retrained comparative models: (i) without CDMS; Inside CDMS, (ii) without self-attention (CDMS w/o SA); (iii) without cross-attention (CDMS w/o CA); (iv) without adaptive weighting (CDMS w/o AW). The results of the four modified models are shown in the rows *I* to *L* of Table 5, respectively. Compared to the original model, removing CDMS had the greatest influence on performance, with RSR and RPR scores dropping by 3.9% and 4.9%, respectively. When removing distinct components of CDMS, the performance degrades to differing degrees. These results demonstrate that the proposed CDMS and its components all contribute to the tracking performance.

To further assess our proposed adaptive weighting scheme, we report the estimated two weights (i.e., w_f for the frame domain; w_e for the event domain) of extended PrDiMP tracker in Fig. 10. In Fig. 10a–d, the frame domain cannot provide reliable visual cues. Correspondingly, we can see the w_e in these examples are significantly higher than w_f. In Fig. 10e–h, when the event data appears deceptive, we can see that w_f tends to increase and w_e decreases accordingly. In Fig. 10e, the tracked object moves from the overexposed region to the normal region, and the frame domain's weight w_e shifts from low to high correspondingly. In the overexposed scene, our method tends to give higher weight to the

Table 5 Ablation study results on the FE141 datset. $A \Leftarrow B$ denotes A is estimated from B

	Models	RSR ↑	$\text{OP}_{0.50}$ ↑	$\text{OP}_{0.75}$ ↑	RPR ↑
A.	Frame only	43.5	51.9	17.3	68.0
B.	Event only	47.3	59.1	19.9	74.5
C.	w/o GM-LSTM	51.1	63.2	25.0	78.2
D.	GM-LSTM w/o GSDE	54.2	67.1	26.4	83.5
E.	GM-LSTM w/o MAM	54.8	67.4	26.1	84.9
F.	ConvLSTM (Shi et al., 2015)	53.7	66.1	26.1	82.8
G.	$Q_e \Leftarrow e_t, Q_h \Leftarrow h_{t-1}$	54.1	67.8	27.2	83.2
H.	$Q_e, Q_h \Leftarrow e_t$	55.1	68.6	**32.0**	84.0
I.	w/o CDMS	52.3	63.3	24.1	82.4
J.	CDMS w/o SA	55.4	67.6	26.9	86.0
K.	CDMS w/o CA	55.1	67.9	27.2	84.9
L.	CDMS w/o AW	54.1	66.1	26.5	84.0
M.	TSLTD (Chen et al., 2020)	54.7	67.6	27.4	83.8
N.	Event count (Maqueda et al., 2018)	53.6	66.5	24.7	82.7
O.	Event frame (Rebecq et al., 2017)	53.2	65.6	22.9	83.3
P.	Zhu et al. (2019a)	55.4	68.0	26.3	85.7
Q.	Time surface (Sironi et al., 2018)	53.6	66.6	24.2	83.2
R.	Ours	**56.2**	**68.8**	27.2	**87.3**

Bold values indicate the best result

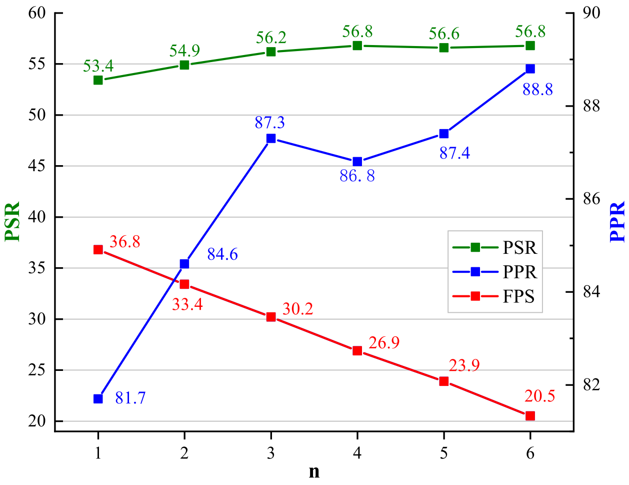


Fig. 8 Trade-off between accuracy and efficiency introduced by the number of slices of event aggregation (i.e., n)

event domain to fully exploit its advantages; In normal illumination, the weight of the frame domain is higher than that of the event domain because it provides texture information to distinguish similar objects. In Fig. 10f, g, we think the reason w_e is larger than w_f is that the model is trained to focus on texture cues in the frame domain, but no texture cues can be extracted in these cases. Fig. 10h depicts a target transitioning from motion to rest. *#12* shows that during motion, the event domain provides obvious edge information compared to the frame, so w_e is greater than w_f; While the target is stationary in *#16*, the event domain cannot provide valuable information, causing w_e to drop sharply. It is worthwhile to mention that the extended trackers with our modules can successfully track the target in all examples. It demonstrates that our proposed approach dynamically balances the contributions of the frame and event domains. Even if one domain is limited, the proposed CDMS can still provide valuable information for object localization based on another domain.

5.3.4 Impact of Event Representation

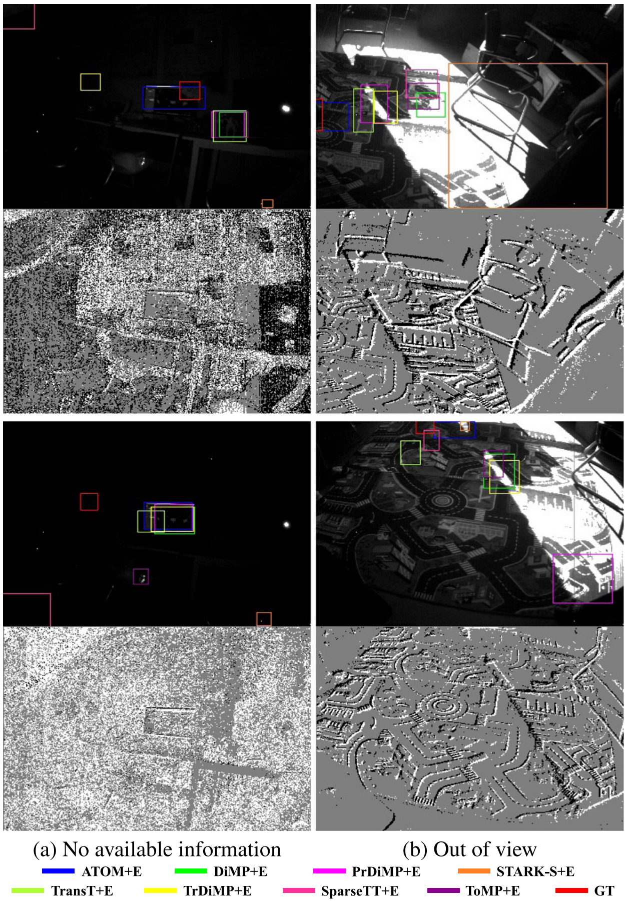
In this work, two primary factors about event representation impact tracking performance: (i) The way of accumulating raw events. Our proposed accumulated method retains the latest timestamps to record the most recent tracking-critical motion cues. To confirm its effectiveness, we conduct experiments with five commonly used event aggregation methods (Chen et al., 2020; Maqueda et al., 2018; Rebecq et al., 2017; Zhu et al., 2019a; Sironi et al., 2018). The results are shown in the rows *M*-*Q* of Table 5, which shows our method outperforms other compared representation approaches; (ii) The hyperparameter n. For events recorded between two consecutive frames, we slice them into n blocks in the temporal domain and then accumulate them as input to EFE. As shown in Fig. 8, both RSR and RPR scores show an increasing trend with increasing n value. However, with a larger n value, it slows down the inference time. We can see $n = 3$ offers the best trade-off between accuracy and efficiency. These results suggest that converting event streams to spatiotemporal voxel grids can improve tracking robustness.

5.4 Comparison of Fusion Strategies

To further demonstrate the necessity that the event feature extractor GM-LSTM and the fusion module CDMS need to be carefully designed, we first combine conventional frame and event aggregated frame by concatenation manner to train and test the top three frame-based trackers (i.e., PrDiMP (Danelljan et al., 2020), TransT (Chen et al., 2021), and STARK-S (Yan et al., 2021a)). Here, we adopt the following two kinds of fusion strategies: (a) Early Fusion (EF), we first concatenate corresponding frame and event data as one unified data, and then feed the fused data into the tracking model; (b) Middle Fusion (MF), we first use the backbone of frame-based trackers to extract frame and event features separately, then concatenate the extracted features and feed them to the regressor. In addition, we extend our preliminary version FENet (Zhang et al., 2021) to the top-3 trackers, turning them into multi-modal methods. As shown in Table 6, the extended multi-modal trackers with our approach still outperform all others by a considerable margin. It reflects the effectiveness of our specially designed event features extractor and cross-domain feature integrator. We also witness that the performance of the three chosen approaches can be improved significantly only by naively combining the frame and event domains. It means event information definitely plays an important role in dealing with degraded conditions. Figure 11 provides more visual examples. An interesting observation is that applying the proposed method significantly boosts the performance for TransT in RPR of SCM scene, but leads to performance drops for PrDiMP and STARK. SCM is one of the most challenging conditions for the event domain. When an event camera moves drastically, almost all edges in a scene trigger events and make tracking challenging. Thus, the RPR for extended PrDiMP and STARK have decreased. TransT designs elaborate self-attention and cross-attention schemes to adaptively focus on abundant context and semantic information in frame and event domains. We believe this is the primary cause for the increase in RPR on the extended TransT.

5.5 Computational Cost

As shown in Table 7, we evaluate the trade-off between the performance and the number of network complexity of top-3 trackers from two aspects: (i) Comparison between the base tracker and the extended tracker with our proposed modules. Extending base trackers to multi-modal trackers involves event feature extraction and multi-modal fusion, which inevitably introduces additional computational consumption. The extended methods significantly improve the robustness of tracking in real degraded conditions, and part of them can still run in real-time at over 30.0 fps. (ii) Comparison between the extended tracker with ConvLSTM (Shi et al., 2015) and with our proposed GM-LSTM. The results show that our GM-LSTM achieves significant performance gains at an extremely low computational cost compared to ConvLSTM. For example, on the extended tracker PrDiMP (Danelljan et al., 2020), our method adds only 0.6G MACs and 0.4M Params compared to ConvLSTM-based, which improves the RSR and RPR scores by 2.5% and 4.5%, respectively. This shows that our approach can exploit spatiotemporal features with motion cues from events in a nearly cost-free fashion. Our GM-LSTM remains lightweight for two reasons: we only add three extra convolutional layers in GM-LSTM, slightly increasing the number of parameters; and we downsample the input to GM-LSTM to achieve low multiply-add operations.

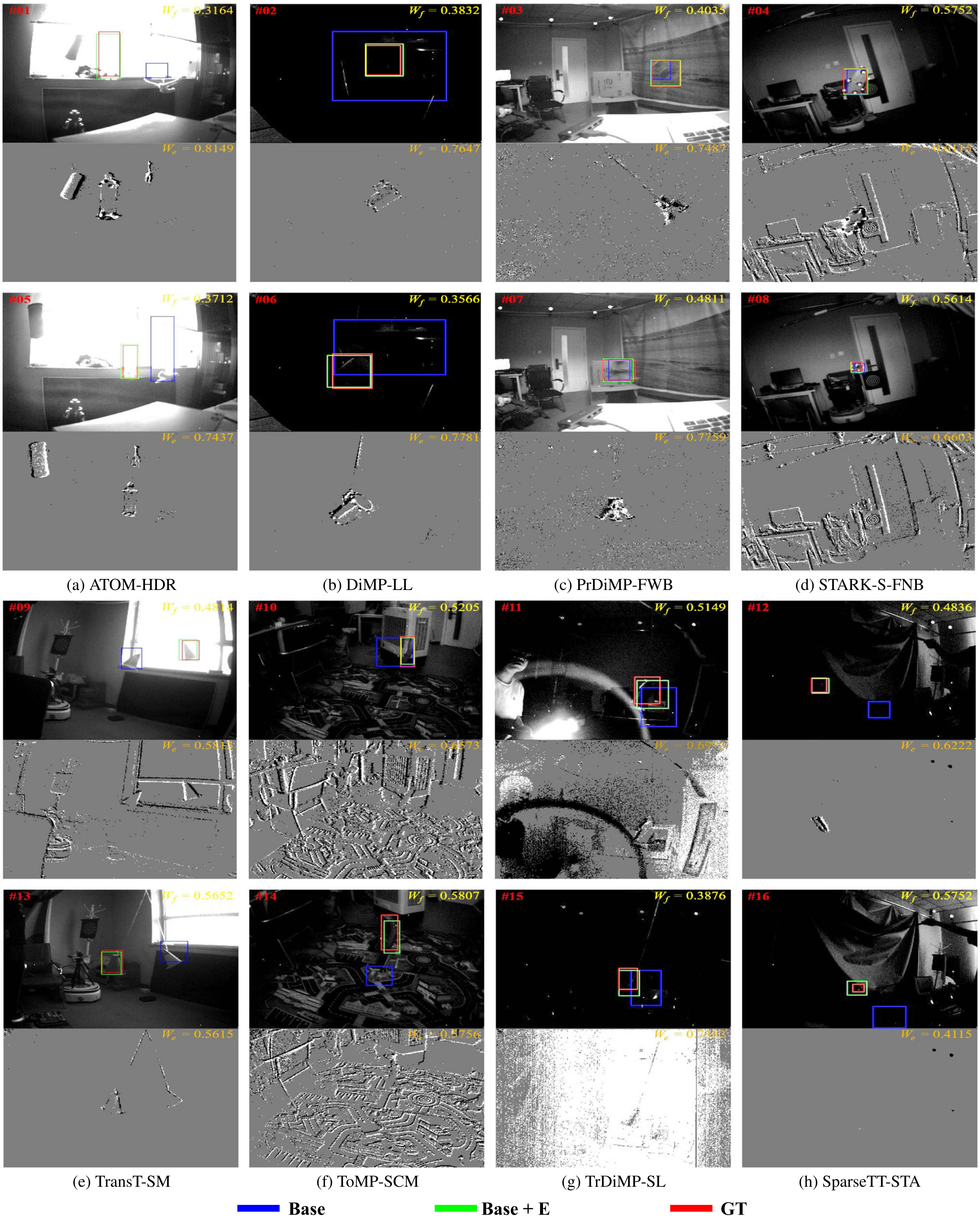

(a) No available information

(b) Out of view

Fig. 9 Failure cases. Extended trackers may fail in cases where the frame and event images can not provide sufficient information (**a**), or some portion of the target leaves the view (**b**)

6 Discussion and Conclusion

In this paper, we introduce the event domain into frame-based tracking approaches for boosting tracking performance under different challenging conditions. Our proposed event



(a) ATOM-HDR (b) DiMP-LL (c) PrDiMP-FWB (d) STARK-S-FNB

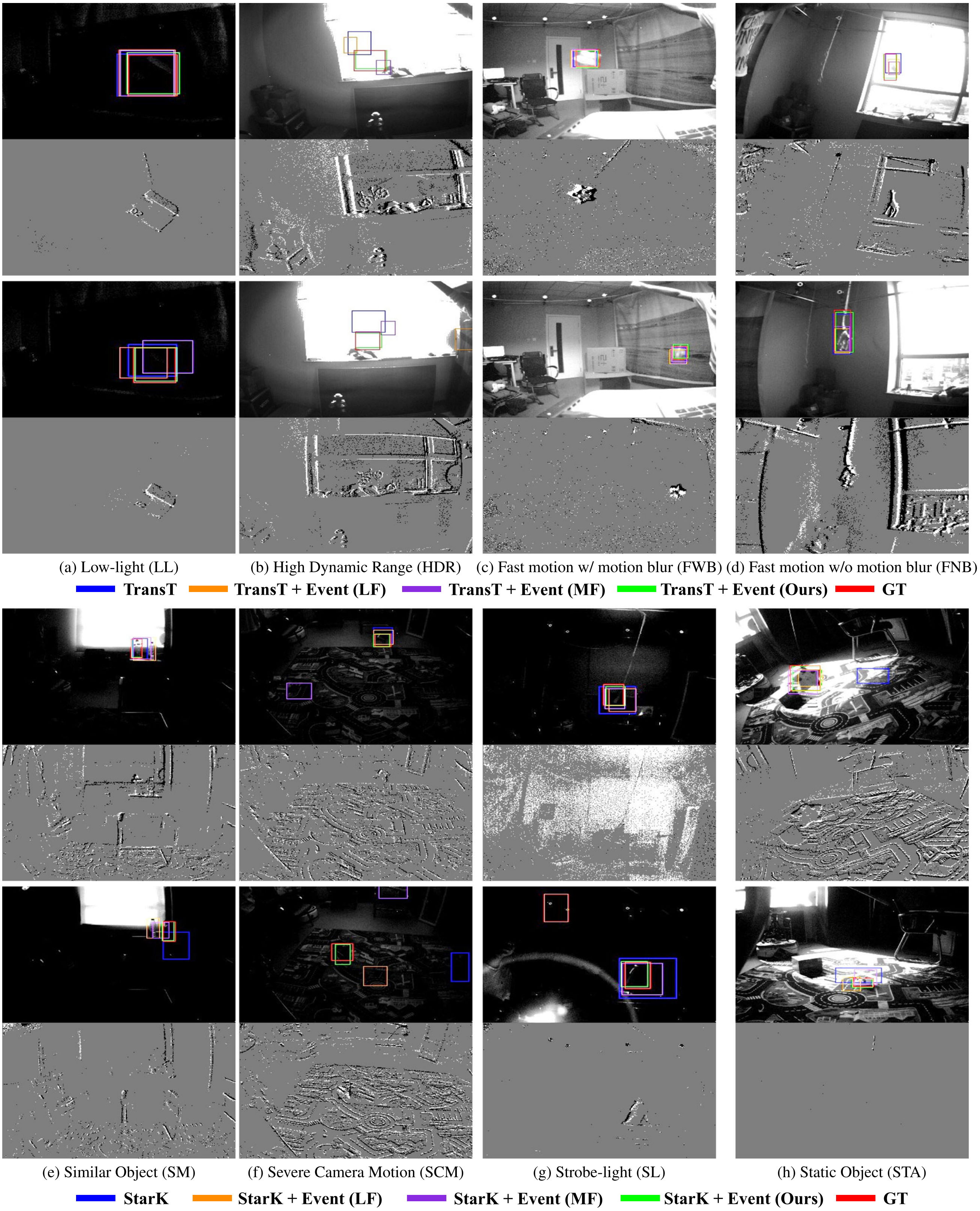
(e) TransT-SM (f) ToMP-SCM (g) TrDiMP-SL (h) SparseTT-STA

Fig. 10 Visual outputs of state-of-the-art algorithms on FE141 dataset. 'A–B' means approach A under challenging condition B. 'Base': the original frame-bases trackers; 'Base+E': the extended trackers with our proposed modules. To better visualize the scene, we only visualize the first of the n event frames and apply a gamma correction to it

Table 6 Comparison of different fusion strategies based on PrDiMP (Danelljan et al., 2020), TransT (Chen et al., 2021) and Stark-S (Yan et al., 2021a) on the FE141 dataset under different challenge scenes

Methods	HDR		LL		FWB		FNB		SM		SCM		SL		STA		ALL	
	RSR	RPR	RSR	RPR	RSR	RPR	RSR	RPR	RSR	RPR	RSR	RPR	RSR	RPR	RSR	RPR	RSR	RPR
PrDiMP (Danelljan et al., 2020)	37.9	55.3	45.9	65.6	46.1	73.7	56.0	84.2	42.0	62.4	**46.9**	**77.4**	23.5	49.9	27.7	50.2	43.5	68.0
PrDiMP + E (EF)	50.4	72.9	49.0	72.8	44.8	70.9	59.7	91.9	45.8	76.1	41.9	68.3	22.0	41.8	32.0	67.8	47.1	73.9
PrDiMP + E (MF)	51.9	77.3	63.5	95.8	49.0	77.0	58.0	86.3	51.6	83.6	33.6	53.7	16.6	30.0	23.2	39.0	47.6	73.7
PrDiMP + E (FENet (Zhang et al., 2021))	59.9	85.8	65.9	95.4	55.7	87.3	62.4	**93.4**	35.8	65.1	40.5	66.4	22.2	39.2	39.7	70.1	53.0	81.1
PrDiMP + E (Ours)	**60.9**	**87.1**	**69.3**	**99.2**	**57.8**	**95.0**	**68.8**	92.6	**57.3**	**88.8**	42.3	69.0	**33.5**	**63.4**	**55.4**	**94.8**	**56.2**	**87.3**
TransT (Chen et al., 2021)	47.0	68.1	40.8	58.8	43.2	65.6	58.3	88.0	38.7	60.9	35.7	56.3	19.5	43.8	28.6	49.7	42.9	65.9
TransT + E (EF)	49.6	72.0	47.3	70.1	46.6	73.4	59.7	**91.8**	46.7	77.8	42.7	69.3	22.0	44.7	29.3	58.5	46.6	73.2
TransT + E (MF)	55.2	80.0	52.2	77.4	46.2	71.7	59.8	91.7	48.9	74.6	43.5	**71.6**	19.5	42.0	52.2	**96.5**	48.3	75.3
TransT + E (FENet (Zhang et al., 2021))	58.9	85.5	63.2	95.5	52.3	82.4	61.2	91.4	45.3	**85.0**	41.5	67.6	21.4	31.1	52.4	93.2	52.2	82.2
TransT + E (Ours)	**61.7**	**87.9**	**67.8**	**97.6**	**55.5**	**87.9**	**62.5**	91.1	**54.5**	84.3	**43.9**	70.3	**33.5**	**61.6**	**52.7**	87.5	**55.7**	**85.1**
STARK-S (Yan et al., 2021a)	49.4	71.2	41.6	61.3	46.8	71.5	59.0	90.6	38.9	61.8	41.4	**70.2**	19.3	38.4	21.7	48.0	44.7	69.4
STARK-S + E (EF)	50.1	72.8	64.5	80.9	52.7	**89.4**	58.9	**91.2**	47.2	74.0	40.4	67.2	20.3	43.8	26.1	51.3	48.6	77.5
STARK-S + E (MF)	51.3	75.8	51.2	78.0	51.3	81.9	57.9	89.9	43.9	71.2	40.5	65.4	24.8	**50.6**	26.8	53.8	49.8	78.7
STARK-S + E (FENet (Zhang et al., 2021))	61.6	84.8	68.3	87.3	47.1	68.8	62.5	89.6	60.7	81.2	**45.1**	70.1	22.0	40.1	41.0	57.0	54.1	78.2
STARK-S + E (Ours)	**64.1**	**86.8**	**68.7**	**89.4**	**57.6**	84.9	**63.8**	89.5	**62.1**	**81.6**	43.5	67.2	**24.9**	42.5	**41.6**	**59.2**	**56.2**	**79.7**

Bold values indicate the best result



(a) Low-light (LL) (b) High Dynamic Range (HDR) (c) Fast motion w/ motion blur (FWB) (d) Fast motion w/o motion blur (FNB)

(e) Similar Object (SM) (f) Severe Camera Motion (SCM) (g) Strobe-light (SL) (h) Static Object (STA)

Fig. 11 Visual results of different fusion strategies based on TransT (Chen et al., 2021) and StarK (Yan et al., 2021a) on FE141 under different challenge scenes. To better visualize the scene, we only visualize the first of the n event frames and apply a gamma correction to it

Table 7 Comparison of the computational efficiency on the FE141 dataset

Networks	MACs (G)	Params (M)	fps	RSR	RPR
PrDiMP (Danelljan et al., 2020)	6.0	11.7	43.1	43.5	68.0
PrDiMP + E (ConvLSTM)	13.4	23.5	32.0	53.7	82.8
PrDiMP + E (ours)	14.0	23.9	30.2	56.2	87.3
TransT (Chen et al., 2021)	11.9	18.4	43.9	42.9	65.9
TransT + E (ConvLSTM)	22.0	27.8	32.9	53.7	82.8
TransT + E (ours)	22.4	28.2	31.4	55.7	85.1
STARK-S (Yan et al., 2021a)	9.7	21.4	41.5	44.7	69.4
STARK-S + E (ConvLSTM)	19.5	30.8	29.1	53.4	78.5
STARK-S + E (ours)	19.8	31.2	28.7	56.2	79.7

Note that the listed MACs and parameters do not include the classifier and regressor

feature extractor can effectively extract spatial and temporal information with motion cues from event-based data. Our novel designed attention schemes effectively and adaptively fuse the information obtained from both the frame and event domains. We also introduce a large-scale frame-event-based object tracking dataset to train our networks and stimulate further research in this area. Multiple extended trackers with our approach outperform corresponding original trackers, which indicates leveraging the complementarity of events and frames boosts the robustness of object tracking in degraded conditions.

6.1 Limitation

Although extended trackers with our proposed method have achieved state-of-the-art tracking performance on the FE141, EED and VisEvent testing set, they do have their limitations. In particular, extended multi-modal methods may fail in cases where the scene is very complex and data of both domains provides insufficient information, as shown in Fig. 9a. When some portion of the target leaves the view, these trackers do not successfully match the target, as shown in Fig. 9b. We believe that a deeper research is needed to study these extreme cases.

6.2 Future Work

At this moment, we mainly focus on developing a cross-domain fusion scheme that can enhance visual tracking robustness, especially in degraded conditions. Although the ablation studies indeed verified the effectiveness of our proposed GM-LSTM and CDMS. Simplifying the architecture while keeping or even improving the performance is part of our future work. Furthermore, we have not leveraged the high measurement rate of event-based cameras to achieve high temporal resolution tracking in this paper. Leveraging the high event measurement rate to achieve higher tracking speed is attractive to many real-world applications. One possible way to achieve it is generating latent frames based on one frame and the events captured afterwards. However, the computational latency introduced by latent frame estimations may be a barrier to achieving desired tracking speed. Another possible solution is leveraging temporal alignment between events and frames. We think graph-based algorithms should be a promising direction. Moreover, using advanced high-quality APS and high-resolution DVS for tracking provides significant promise, yet it also presents certain challenges. For example, multi-modality data with differing viewpoints may need alignment and synchronization to ensure temporal and spatial consistency; one modality may lose critical information, thereby increasing the complexity of the fusion. Our further work will also focus on expanding the FE141 dataset by collecting more sequences, with a specific focus on enhancing data quality and incorporating common non-rigid objects.

Acknowledgements This work was supported in part by National Key Research and Development Program of China (2022ZD0210500), the National Natural Science Foundation of China under Grants 62332019/61972067, and the Distinguished Young Scholars Funding of Dalian (No. 2022RJ01).

References

An, N., Zhao, X. G., & Hou, Z. G. (2016). Online RGB-D tracking via detection-learning-segmentation. In *ICPR* (pp. 1231–1236).

Barranco, F., Fermuller, C., & Ros, E. (2018). Real-time clustering and multi-target tracking using event-based sensors. In *IROS* (pp. 5764–5769).

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., & Torr, P. H. (2016). Fully-convolutional Siamese networks for object tracking. In *ECCV* (pp. 850–865).

Bhat, G., Danelljan, M., Gool, L. V., & Timofte, R. (2019). Learning discriminative model prediction for tracking. In *ICCV* (pp. 6182–6191).

Bhat, G., Danelljan, M., Van Gool, L., & Timofte, R. (2020). Know your surroundings: Exploiting scene information for object tracking. In *ECCV* (pp. 205–221).

Cai, L., McGuire, N. E., Hanlon, R., Mooney, T. A., & Girdhar, Y. (2023). Semi-supervised visual tracking of marine animals using autonomous underwater vehicles. *International Journal of Computer Vision, 131*(6), 1406–1427.

Camplani, M., Hannuna, S. L., Mirmehdi, M., Damen, D., Paiement, A., Tao, L., & Burghardt, T. (2015). Real-time RGB-D tracking with depth scaling kernelised correlation filters and occlusion handling. In *BMVC* (Vol. 4, p. 5).

Camuñas-Mesa, L. A., Serrano-Gotarredona, T., Ieng, S. H., Benosman, R., & Linares-Barranco, B. (2017). Event-driven stereo visual tracking algorithm to solve object occlusion. *IEEE Transactions on Neural Networks and Learning Systems, 29*(9), 4223–4237.

Chen, H., Suter, D., Wu, Q., & Wang, H. (2020). End-to-end learning of object motion estimation from retinal events for event-based object tracking. In *AAAI* (Vol. 34, pp. 10,534–10,541).

Chen, H., Wu, Q., Liang, Y., Gao, X., & Wang, H. (2019). Asynchronous tracking-by-detection on adaptive time surfaces for event-based object tracking. In *ACM MM* (pp. 473–481).

Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., & Lu, H. (2021). Transformer tracking. In *CVPR* (pp. 8126–8135).

Chen, Z., Zhong, B., Li, G., Zhang, S., & Ji, R. (2020). Siamese box adaptive network for visual tracking. In *CVPR* (pp. 6668–6677).

Cui, Y., Guo, D., Shao, Y., Wang, Z., Shen, C., Zhang, L., & Chen, S. (2022). Joint classification and regression for visual tracking with fully convolutional Siamese networks. *International Journal of Computer Vision, 130*(2), 550–566.

Danelljan, M., Bhat, G., Khan, F. S., & Felsberg, M. (2019). ATOM: Accurate tracking by overlap maximization. In *CVPR* (pp. 4660–4669).

Danelljan, M., Gool, L. V., & Timofte, R. (2020). Probabilistic regression for visual tracking. In *CVPR* (pp. 7183–7192).

Ding, J., Dong, B., Heide, F., Ding, Y., Zhou, Y., Yin, B., & Yang, X. (2022). Biologically inspired dynamic thresholds for spiking neural networks. In *NeurIPS* (Vol. 35, pp. 6090–6103).

Ding, J., Gao, L., Liu, W., Piao, H., Pan, J., Du, Z., Yang, X., & Yin, B. (2022). Monocular camera-based complex obstacle avoidance via efficient deep reinforcement learning. *IEEE Transactions on Circuits and Systems for Video Technology, 33*(2), 756–770.

Fan, H., Bai, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Huang, M., Liu, J., Xu, Y., et al. (2021). LaSOT: A high-quality large-scale single object tracking benchmark. *International Journal of Computer Vision, 129*, 439–461.

Fu, Z., Fu, Z., Liu, Q., Cai, W., & Wang, Y. (2022). Sparsett: Visual tracking with sparse transformers. In *IJCAI* (pp. 905–912).

Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., Leutenegger, S., Davison, A. J., Conradt, J., Daniilidis, K., & Scaramuzza, D. (2019). Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(1), 154–180.

Gao, J., Hu, W., & Lu, Y. (2020). Recursive least-squares estimator-aided online learning for visual tracking. In *CVPR* (pp. 7386–7395).

Gao, J., Zhang, T., & Xu, C. (2019). Graph convolutional tracking. In *CVPR* (pp. 4649–4659).

Gao, S., Zhou, C., Ma, C., Wang, X., & Yuan, J. (2022). AiATrack: Attention in attention for transformer visual tracking. In *ECCV* (pp. 146–164).

Gehrig, D., Loquercio, A., Derpanis, K. G., & Scaramuzza, D. (2019). End-to-end learning of representations for asynchronous event-based data. In *ICCV* (pp. 5633–5643).

Guo, D., Wang, J., Cui, Y., Wang, Z., & Chen, S. (2020). SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In *CVPR* (pp. 6269–6277).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR* (pp. 770–778).

Huang, J., Wang, S., Guo, M., & Chen, S. (2018). Event-guided structured output tracking of fast-moving objects using a Celex sensor. *IEEE Transactions on Circuits and Systems for Video Technology, 28*(9), 2413–2417.

Huang, L., Zhao, X., & Huang, K. (2019). GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(5), 1562–1577.

Hui, T., Xun, Z., Peng, F., Huang, J., Wei, X., Wei, X., Dai, J., Han, J., & Liu, S. (2023). Bridging search region interaction with template for RGB-T tracking. In *CVPR* (pp. 9516–9526).

Hu, Y., Liu, H., Pfeiffer, M., & Delbruck, T. (2016). DVS benchmark datasets for object tracking, action recognition, and object recognition. *Frontiers in Neuroscience, 10*, 405.

Jiang, B., Luo, R., Mao, J., Xiao, T., & Jiang, Y. (2018). Acquisition of localization confidence for accurate object detection. In *ECCV* (pp. 784–799).

Jiawen, Z., Simiao, l., Xin, C., Wang, D., & Lu, H. (2023). Visual prompt multi-modal tracking. In *CVPR*.

Kart, U., Kämäräinen, J. K., Matas, J., & Matas, J. (2018). How to make an RGBD tracker? In *ECCVw* (pp. 148–161).

Kristan, M. E. A. (2014). The visual object tracking VOT2014 challenge results. In *ECCVW* (pp. 191–217).

Kristan, M. E. A. (2017). The visual object tracking VOT2017 challenge results. In *ICCVW* (pp. 1949–1972).

Lagorce, X., Orchard, G., Galluppi, F., Shi, B. E., & Benosman, R. B. (2016). HOTS: A hierarchy of event-based time-surfaces for pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(7), 1346–1359.

Lan, X., Ye, M., Zhang, S., & Yuen, P. C. (2018). Robust collaborative discriminative learning for RGB-infrared tracking. In *AAAI* (Vol. 32, pp. 7008–7015).

Li, P., Chen, B., Ouyang, W., Wang, D., Yang, X., & Lu, H. (2019). GradNet: Gradient-guided network for visual object tracking. In *ICCV* (pp. 6162–6171).

Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., & Yan, J. (2019). SiamRPN++: Evolution of Siamese visual tracking with very deep networks. In *CVPR* (pp. 4282–4291).

Li, C., Zhu, C., Huang, Y., Tang, J., & Wang, L. (2018). Cross-modal ranking with soft consistency and noisy labels for robust RGB-T tracking. In *ECCV* (pp. 808–823).

Liang, P., Blasch, E., & Ling, H. (2015). Encoding color information for visual tracking: Algorithms and benchmark. *IEEE Transactions on Image Processing, 24*(12), 5630–5644.

Li, A., Lin, M., Wu, Y., Yang, M. H., & Yan, S. (2015). NUS-PRO: A new visual tracking challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(2), 335–349.

Lin, L., Fan, H., Zhang, Z., Xu, Y., & Ling, H. (2022). SwinTrack: A simple and strong baseline for transformer tracking. In *NeurIPS* (Vol. 35, pp. 16,743–16,754).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV* (pp. 10,012–10,022).

Liu, Y., Xie, J., Shi, X., Qiao, Y., Huang, Y., Tang, Y., & Yang, X. (2021). Tripartite information mining and integration for image matting. In *ICCV* (pp. 7555–7564).

Liu, Y., Long, C., Zhang, Z., Liu, B., Zhang, Q., Yin, B., & Yang, X. (2022). Explore contextual information for 3D scene graph generation. *IEEE Transactions on Visualization and Computer Graphics, 29*(12), 5556–5568.

Long Li, C., Lu, A., Hua Zheng, A., Tu, Z., & Tang, J. (2019). Multi-adapter RGBT tracking. In *ICCVW* (pp. 2262–2270).

Lukezic, A., Kart, U., Kapyla, J., Durmush, A., Kamarainen, J.K., Matas, J., & Kristan, M. (2019). CDTB: A color and depth visual object tracking dataset and benchmark. In *ICCV* (pp. 10,013–10,022).

Maqueda, A. I., Loquercio, A., Gallego, G., García, N., & Scaramuzza, D. (2018). Event-based vision meets deep learning on steering prediction for self-driving cars. In *CVPR* (pp. 5419–5427).

Mayer, C., Danelljan, M., Bhat, G., Paul, M., Paudel, D. P., Yu, F., & Van Gool, L. (2022). Transforming model prediction for tracking. In *CVPR* (pp. 8731–8740).

Messikommer, N., Gehrig, D., Loquercio, A., & Scaramuzza, D. (2020). Event-based asynchronous sparse convolutional networks. In *ECCV* (pp. 415–431).

Mitrokhin, A., Fermüller, C., Parameshwara, C., & Aloimonos, Y. (2018). Event-based moving object detection and tracking. In *IROS* (pp. 1–9).

Mitrokhin, A., Ye, C., Fermuller, C., Aloimonos, Y., & Delbruck, T. (2019). EV-IMO: Motion segmentation dataset and learning pipeline for event cameras. In *IROS* (pp. 6105–6112).

Mostafavi, M., Wang, L., & Yoon, K. J. (2021). Learning to reconstruct HDR images from events, with applications to depth and flow prediction. *International Journal of Computer Vision, 129*, 900–920.

Mueller, M., Smith, N., & Ghanem, B. (2016). A benchmark and simulator for UAV tracking. In *ECCV* (pp. 445–461).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS* (Vol. 32).

Piatkowska, E., Belbachir, A. N., Schraml, S., & Gelautz, M. (2012). Spatiotemporal multiple persons tracking using dynamic vision sensor. In *CVPRW* (pp. 35–40).

Qiao, Y., Liu, Y., Yang, X., Zhou, D., Xu, M., Zhang, Q., & Wei, X. (2020). Attention-guided hierarchical structure aggregation for image matting. In *CVPR* (pp. 13,676–13,685).

Qiao, Y., Zhu, J., Long, C., Zhang, Z., Wang, Y., Du, Z., & Yang, X. (2022). CPRAL: Collaborative panoptic-regional active learning for semantic segmentation. In *AAAI* (Vol. 36, pp. 2108–2116).

Rebecq, H., Gallego, G., Mueggler, E., & Scaramuzza, D. (2018). EMVS: Event-based multi-view stereo-3D reconstruction with an event camera in real-time. *International Journal of Computer Vision*.

Rebecq, H., Horstschaefer, T., & Scaramuzza, D. (2017). Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization. In *BMVC* (pp. 16–1).

Shen, Q., Qiao, L., Guo, J., Li, P., Li, X., Li, B., Feng, W., Gan, W., Wu, W., & Ouyang, W. (2022). Unsupervised learning of accurate Siamese tracking. In *CVPR* (pp. 8101–8110).

Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. c. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NeurIPS* (Vol. 28).

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2017). Deep learning for precipitation nowcasting: A benchmark and a new model. In *NeurIPS* (Vol. 30).

Sironi, A., Brambilla, M., Bourdis, N., Lagorce, X., & Benosman, R. (2018). HATS: Histograms of averaged time surfaces for robust event-based object classification. In *CVPR* (pp. 1731–1740).

Song, S., & Xiao, J. (2013). Tracking revisited using RGBD camera: Unified benchmark and baselines. In *ICCV* (pp. 233–240).

Srivastava, N., Mansimov, E., & Salakhudinov, R. (2015). Unsupervised learning of video representations using LSTMs. In *ICLR* (pp. 843–852).

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *NeurIPS* (Vol. 27).

Vicon motion capture. https://www.vicon.com/

Wang, X., Li, J., Zhu, L., Zhang, Z., Chen, Z., Li, X., Wang, Y., Tian, Y., & Wu, F. (2021). VisEvent: Reliable object tracking via collaboration of frame and event flows. arXiv:2108.05015

Wang, Y., Long, M., Wang, J., Gao, Z., & Yu, P. S. (2017). PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. In *NeurIPS* (vol. 30).

Wang, N., Song, Y., Ma, C., Zhou, W., Liu, W., & Li, H. (2019). Unsupervised deep tracking. In *CVPR* (pp. 1308–1317).

Wang, Q., Teng, Z., Xing, J., Gao, J., Hu, W., & Maybank, S. (2018). Learning attentions: residual attentional siamese network for high performance online visual tracking. In *CVPR* (pp. 4854–4863).

Wang, C., Xu, C., Cui, Z., Zhou, L., Zhang, T., Zhang, X., & Yang, J. (2020). Cross-modal pattern-propagation for RGB-T tracking. In *CVPR* (pp. 7064–7073).

Wang, N., Zhou, W., Wang, J., & Li, H. (2021). Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *CVPR* (pp. 1571–1580).

Wang, T., Li, J., Wu, H. N., Li, C., Snoussi, H., & Wu, Y. (2022). ResLNet: Deep residual LSTM network with longer input for action recognition. *Frontiers of Computer Science, 16*, 166,334.

Wang, Y., Zhang, X., Shen, Y., Du, B., Zhao, G., Lizhen, L. C. C., & Wen, H. (2021). Event-stream representation for human gaits identification using deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(7), 3436–3449.

Wang, N., Zhou, W., Song, Y., Ma, C., Liu, W., & Li, H. (2021). Unsupervised deep representation learning for real-time tracking. *International Journal of Computer Vision, 129*, 400–418.

Wu, Y., Lim, J., & Yang, M. H. (2013). Online object tracking: A benchmark. In *CVPR* (pp. 2411–2418).

Wu, Y., Lim, J., & Yang, M. H. (2015). Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(9), 1834–1848.

Wu, H., Yao, Z., Wang, J., & Long, M. (2021). MotionRNN: A flexible model for video prediction with spacetime-varying motions. In *CVPR* (pp. 15,435–15,444).

Wu, Y., Deng, L., Li, G., Zhu, J., & Shi, L. (2018). Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in Neuroscience, 12*, 331.

Xiao, J., Stolkin, R., Gao, Y., & Leonardis, A. (2017). Robust fusion of color and depth data for RGB-D target tracking using adaptive range-invariant depth models and spatio-temporal consistency constraints. *IEEE Transactions on Cybernetics, 48*(8), 2485–2499.

Yan, B., Peng, H., Fu, J., Wang, D., & Lu, H. (2021). Learning spatio-temporal transformer for visual tracking. In *ICCV* (pp. 10,448–10,457).

Yan, S., Yang, J., Käpylä, J., Zheng, F., Leonardis, A., & Kämäräinen, J. K. (2021). DepthTrack: Unveiling the power of RGBD tracking. In *ICCV* (pp. 10,725–10,733).

Yang, J., Gao, S., Li, Z., Zheng, F., & Leonardis, A. (2023). Resource-efficient RGBD aerial tracking. In *CVPR* (pp. 13,374–13,383).

Yang, X., Mei, H., Xu, K., Wei, X., Yin, B., & Lau, R. W. (2019). Where is my mirror? In *ICCV* (pp. 8809–8818).

Yang, Z., Wu, Y., Wang, G., Yang, Y., Li, G., Deng, L., Zhu, J., & Shi, L. (2019). DashNet: A hybrid artificial and spiking neural network for high-speed object tracking. arXiv:1909.12942

Zhang, Z., & Peng, H. (2019). Deeper and wider siamese networks for real-time visual tracking. In *CVPR* (pp. 4591–4600).

Zhang, L., Danelljan, M., Gonzalez-Garcia, A., van de Weijer, J., & Shahbaz Khan, F. (2019). Multi-modal fusion for end-to-end RGB-T tracking. In *ICCVW* (pp. 2252–2261).

Zhang, J., Dong, B., Zhang, H., Ding, J., Heide, F., Yin, B., & Yang, X. (2022). Spiking transformers for event-based single object tracking. In *CVPR* (pp. 8801–8810).

Zhang, L., Gonzalez-Garcia, A., Weijer, J. V. D., Danelljan, M., & Khan, F. S. (2019). Learning the model update for siamese trackers. In *ICCV* (pp. 4010–4019).

Zhang, T., Guo, H., Jiao, Q., Zhang, Q., & Han, J. (2023). Efficient RGB-T tracking via cross-modality distillation. In *CVPR* (pp. 5404–5413).

Zhang, J., Wang, Y., Liu, W., Li, M., Bai, J., Yin, B., & Yang, X. (2023). Frame-event alignment and fusion network for high frame rate tracking. In *CVPR* (pp. 9781–9790).

Zhang, J., Yang, X., Fu, Y., Wei, X., Yin, B., & Dong, B. (2021). Object tracking by jointly exploiting frame and event domain. In *ICCV* (pp. 13,043–13,052).

Zhang, H., Zhang, J., Dong, B., Peers, P., Wu, W., Wei, X., Heide, F., & Yang, X. (2023). In the blink of an eye: Event-based emotion recognition. In *SIGGRAPH* (pp. 1–11).

Zhang, P., Zhao, J., Wang, D., Lu, H., & Ruan, X. (2022). Visible-thermal UAV tracking: A large-scale benchmark and new baseline. In *CVPR* (pp. 8886–8895).

Zhang, T., Xu, C., & Yang, M. H. (2018). Learning multi-task correlation particle filters for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(2), 365–378.

Zhang, H., Zhang, L., Dai, Y., Li, H., & Koniusz, P. (2023). Event-guided multi-patch network with self-supervision for non-uniform motion deblurring. *International Journal of Computer Vision, 131*(2), 453–470.

Zhang, J., Zhao, K., Dong, B., Fu, Y., Wang, Y., Yang, X., & Yin, B. (2021). Multi-domain collaborative feature representation for robust visual object tracking. *The Visual Computer, 37*(9–11), 2671–2683.

Zhao, H., Chen, J., Wang, L., & Lu, H. (2023). ArkitTrack: A new diverse dataset for tracking using mobile RGB-D data. In *CVPR* (pp. 5126–5135).

Zhao, H., Yan, B., Wang, D., Qian, X., Yang, X., & Lu, H. (2022). Effective local and global search for fast long-term tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(1), 460–474.

Zhou, C., Teng, M., Han, J., Liang, J., Xu, C., Cao, G., & Shi, B. (2023). Deblurring low-light images with events. *International Journal of Computer Vision, 126*(12), 1394–1414.

Zhou, Q., Wang, R., Li, J., Tian, N., & Zhang, W. (2021). Siamese single object tracking algorithm with natural language prior. *Frontiers of Computer Science, 15*, 1–2.

Zhu, Z., Hou, J., & Lyu, X. (2022). Learning graph-embedded key-event back-tracing for object tracking in event clouds. In *NeurIPS* (Vol. 35, pp. 7462–7476).

Zhu, Y., Li, C., Luo, B., Tang, J., & Wang, X. (2019). Dense feature aggregation and pruning for RGBT tracking. In *ACM MM* (pp. 465–472).

Zhu, A. Z., Yuan, L., Chaney, K., & Daniilidis, K. (2019). Unsupervised event-based learning of optical flow, depth, and egomotion. In *CVPR* (pp. 989–997).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.