Object Tracking by Jointly Exploiting Frame and Event Domain

Jiqing Zhang[1,*], Xin Yang[1,*], Yingkai Fu[1], Xiaopeng Wei[1], Baocai Yin[1,†], Bo Dong[2,†]
[1]Dalian University of Technology, [2] SRI International

Abstract

*Inspired by the complementarity between conventional frame-based and bio-inspired event-based cameras, we propose a multi-modal based approach to fuse visual cues from the frame- and event-domain to enhance the single object tracking performance, especially in degraded conditions (*e.g.*, scenes with high dynamic range, low light, and fast-motion objects). The proposed approach can effectively and adaptively combine meaningful information from both domains. Our approach's effectiveness is enforced by a novel designed cross-domain attention schemes, which can effectively enhance features based on self- and cross-domain attention schemes; The adaptiveness is guarded by a specially designed weighting scheme, which can adaptively balance the contribution of the two domains. To exploit event-based visual cues in single-object tracking, we construct a large-scale frame-event-based dataset, which we subsequently employ to train a novel frame-event fusion based model. Extensive experiments show that the proposed approach outperforms state-of-the-art frame-based tracking methods by at least 10.4% and 11.9% in terms of representative success rate and precision rate, respectively. Besides, the effectiveness of each key component of our approach is evidenced by our thorough ablation study.*

1. Introduction

Recently, convolutional neural networks (CNNs) based approaches show promising performance in object tracking tasks [4, 6, 10, 12, 14, 18, 20, 30, 37, 46, 48, 49]. These approaches mainly use conventional frame-based cameras as sensing devices since they can effectively measure absolute light intensity and provide a rich representation of a scene. However, conventional frame-based sensors have limited frame rates (*i.e.*, $\leq$ 120 FPS) and dynamic range (*i.e.*, $\leq$ 60 dB). Thus, they do not work robustly in degraded conditions. Figure 1 (a) and (b) show two examples of degraded conditions, high dynamic range, and fast-moving object, respectively. Under both conditions, we hardly see the moving objects. Thus, obtaining meaningful visual cues of the objects is challenging. By contrast, an event-based camera, a bio-inspired sensor, offers high temporal resolution (up to 1MHz), high dynamic range (up to 140 dB), and low energy consumption [7]. Nevertheless, it cannot measure absolute light intensity and thus texture cues (as shown in Figure 1 (d)). Both sensors are, therefore, complementary. The unique complementarity triggers us to propose a multi-modal sensor fusion-based approach to improve the tracking performance in degraded conditions, which leverages the advantages of both the frame- and event-domain.

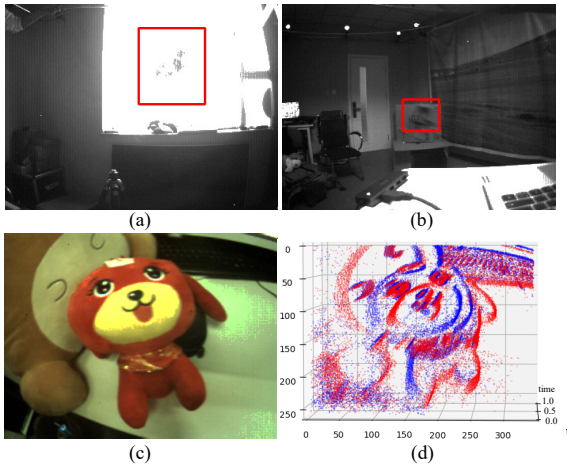


Figure 1. **Limitations of conventional frame-based and bio-inspired event-based cameras.** (a) and (b) show the limitation of a frame-based camera under HDR and fast-moving object, respectively. (d) shows an event-based camera's asynchronous output of the scene shown in (c), sparse and no texture information.

Yet, event-based cameras measure light intensity changes and output events asynchronously. It differs significantly from conventional frame-based cameras, which represent scenes with synchronous frames. Besides, CNNs-based approaches are not designed to digest asynchronous inputs. Therefore, combining asynchronous events and synchronous images remains challenging. To address the challenge, we propose a simple yet effective event aggregation approach to discretize the time domain of asynchronous events. Each of the discretized time slices can be accumu-

* Joint first authors. †Baocai Yin (ybc@dlut.edu.cn) and Bo Dong (bo.dong@sri.com) are the corresponding authors.

lated to a conventional frame, thus can be easily processed by a CNNs-based model. Our experimental results show the proposed aggregation method outperforms other commonly used event accumulation approaches [9, 26, 32, 40, 50]. Another critical challenge, similar to other multi-modal fusion-based approaches [2, 8, 27, 29, 33, 41], is grasping meaningful cues from both domains effectively regardless the diversity of scenes. In doing so, we introduce a novel cross-domain feature integrator, which leverages self- and cross-domain attention schemes to fuse visual cues from both the event- and frame-domain effectively and adaptively. The effectiveness is enforced by a novel designed feature enhancement module, which enhances its own domain's feature based on both domains' attentions. Our approach's adaptivity is held by a specially designed weighting scheme to balance the contributions of the two domains. Based on the two domains' reliabilities, the weighting scheme adaptively regulates the two domains' contributions. We extensively validate our multi-modal fusion-based method and demonstrate that our model outperforms state-of-the-art frame-based methods by a significant margin, at least 10.4% and 11.9% in terms of representative success rate and precision rate, respectively.

To exploit event-based visual cues in single object tracking and enable more future research on multi-modal learning with asynchronous events, we construct a large-scale single-object tracking dataset, FE108, which contains 108 sequences with a total length of 1.5 hours. FE108 provides ground truth annotations on both the frame- and event-domain. The annotation frequency is up to 40Hz and 240Hz for the frame and event domains, respectively. To the best of our knowledge, FE108 is the largest event-frame-based dataset for single object tracking, which also offers the highest annotation frequency in the event domain.

To sum up, our contributions are as follows:

- We introduce a novel cross-domain feature integrator, which can effectively and adaptively fuse the visual cues provided from both the frame and event domains.
- We construct a large-scale frame-event-based dataset for single object tracking. The dataset covers wide challenging scenes and degraded conditions.
- Our extensively experimental results show our approach outperforms other state-of-the-art methods by a significant margin. Our ablation study evidences the effectiveness of the novel designed attention-based schemes.

2. Related Work

Single-Domain Object Tracking. Recently, deep-learning-based methods have dominated the frame-based object tracking field. Most of the methods [4, 12, 14, 30, 37, 48, 49] leverage conventional frame-based sensors. Only a few attempts have been made to track objects using event-based cameras. Piatkowska *et al.* [39] used an event-based camera for multiple persons tracking in the occurrence of high occlusions, which is enabled by the Gaussian Mixture Model based events clustering algorithm. Barranco *et al.* [3] proposed a real-time mean-shift clustering algorithm using events for multi-object tracking. Mitrokhin *et al.* [35] proposed a novel events representation, time-image, to utilize temporal information of the event stream. With it, they achieve an event-only feature-less motion compensation pipeline. Chen *et al.* [9] pushed the event representation further and proposed a synchronous Time-Surface with Linear Time Decay representation to effectively encode the Spatio-temporal information. Although these approaches reported promising performance in object tracking tasks, they did not consider leveraging frame domains. By contrast, our approach focuses on leveraging complementarity between frame and event domains.

Multi-Domain Object Tracking. Multi-modal-based tracking approaches have been getting more attention. Most of the works leverage RGB-D (RGB + Depth) [2, 8, 24, 41, 44] and RGB-T (RGB + Thermal) [27, 29, 31, 42, 47, 51] as multi-modal inputs to improve tracking performance. Depth is an important cue to help solve the occlusion problem in tracking. When a target object is hidden partially by another object with a similar appearance, the difference in their depth levels will be distinctive and help detect the occlusion. Illumination variations and shadows do not influence images from the thermal infrared sensors. They thus can be combined with RGB to improve performance in degraded conditions (*e.g.*, rain and smog). Unlike these multi-domain approaches, fusing frame and event domains brings a unique challenge caused by the asynchronous outputs of event-based cameras. Our approach aims to solve the problem, which effectively leverage events for improving robustness, especially under degraded conditions.

3. Methodology

3.1. Background: Event-based Camera

An event-based camera is a bio-inspired sensor. It asynchronously measures light intensity changes in scene-illumination at a pixel level. Hence, it provides a very high measurement rate, up to 1MHz [7]. Since light intensity changes are measured in log-scale, an event-based camera also offers a very high dynamic range, 140 dB vs. 60 dB of a conventional camera [17]. When the change of pixel intensity in the log-scale is greater than a threshold, an event is triggered. The polarity of an event reflects the direction of the changes. Mathematically, a set of events can be defined as:

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

where e_k is the k-th event; (x_k, y_k) is the pixel location of event e_k; t_k is the timestamp; $p_k \in \{-1, 1\}$ is the polarity of an event. In a stable lighting condition, events are

triggered by moving edges (*e.g.*, object contour and texture boundaries), making an event-based camera a natural edge extractor.

3.2. Event Aggregation

Since the asynchronous event format differs significantly from the frames generated by conventional frame-based cameras, vision algorithms designed for frame-based cameras cannot be directly applied. To deal with it, events are typically aggregated into a frame or grid-based representation first [19, 26, 32, 34, 40, 43, 50].

We propose a simple yet effective pre-processing method to map events into a grid-based presentation. Specifically, inspired by [50], 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 to 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 i-th aggregated frame can be defined as follows:

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

where T_j is the timestamp of the j-th 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 event processing method outperforms other commonly used approaches (see Table 4).

3.3. Network Architecture

The overall architecture of the proposed approach is illustrated in Figure 2, which has two branches: reference branch (top) and test branch (bottom). The reference and test branches share weights in a siamese style. The core of our approach is the Cross-Domain Feature Integrator (CDFI), designed to leverage both domains' advantages. Specifically, the frame domain provides rich texture information, whereas the event domain is robust to challenging scenes and provides edge information. As shown in Figure 2, the inputs of CDFI are a frame and events captured between the frame and its previous one. We preprocess the events based on Eq. 2. The outputs of CDFI are one low-level (*i.e.*, K_l) and one high-level (*i.e.*, K_h) fused features. The classifier uses the extracted low-level fused features

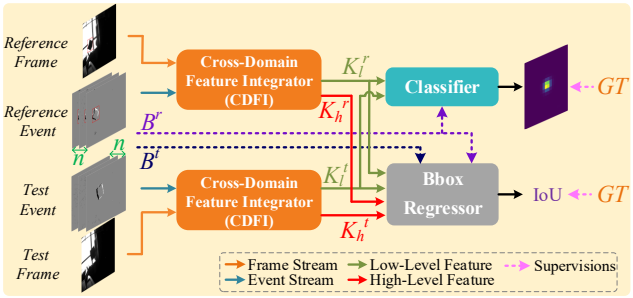


Figure 2. Overview of the proposed architecture.

from both reference and test branches to estimate a confidence map. Finally, the bbox-regressor reports IoU between the ground truth bounding box and estimated bounding box to help locate a target on the test frame.

3.3.1 Cross-Domain Feature Integrator

The overall structure of the proposed CDFI is shown in Figure 3 (a). It has three components, namely: Frame-Feature Extractor (FFE), Event-Feature Extractor (EFE), and Cross-Domain Modulation and Selection Block (CDMS).

FFE is for extracting features from the frame domain. We adopt ResNet18 [21] as our frame feature extractor. The 4th and 5th blocks' features are used as the low-level and high-level frame features (*i.e.*, F_l and F_h), respectively.

EFE generates features to represent the encoded information in the event domain. Similar to FFE, EFE extracts low-/high-level features from the event domain (*i.e.*, E_l and E_h). Since each aggregated event frame conveys different temporal information, each of them is processed by a dedicated sub-branch. Like other feature extractors, each sub-branch of EFE leverages stacked convolutional layers to increase receptive field at higher levels. We also introduce a self-attention scheme to each sub-branch to focus on more critical features. It is achieved by a specially designed Edge Attention Block (EAB), illustrated in Figure 3 (b). As shown in Figure 3 (a), two EABs are added behind the third and fourth convolutional layers. Then, the low-level (*i.e.*, e_i^l) and high-level (*i.e.*, e_i^h) features on the ith sub-branch are generated by the first and second EABs, respectively. Finally, all generated e_i^l and e_i^h are fused in a weighted sum manner to obtain the E_l and E_h. Mathematically, EFE is defined as (here we ignore l and h to bring a general form):

$$E = w_1 e_1 \oplus ... \oplus w_n e_n, \quad (3)$$
$$e_i = \sigma(\psi_{1\times1}(\mathcal{C}(\kappa_i^m))) \otimes \kappa_i, \quad (4)$$
$$\kappa_i^m = \sigma(\mathcal{A}(\kappa_i)) \otimes \kappa_i, \quad (5)$$

where w_i is a learned weight; $\psi_{1\times1}$ means a 1×1 convolution layer; σ is the Sigmoid function; κ_i, e_i, $\mathcal{C}$, and $\mathcal{A}$ are the input, output features of the EAB on the ith sub-branch, channel-wise addition, and adaptive average

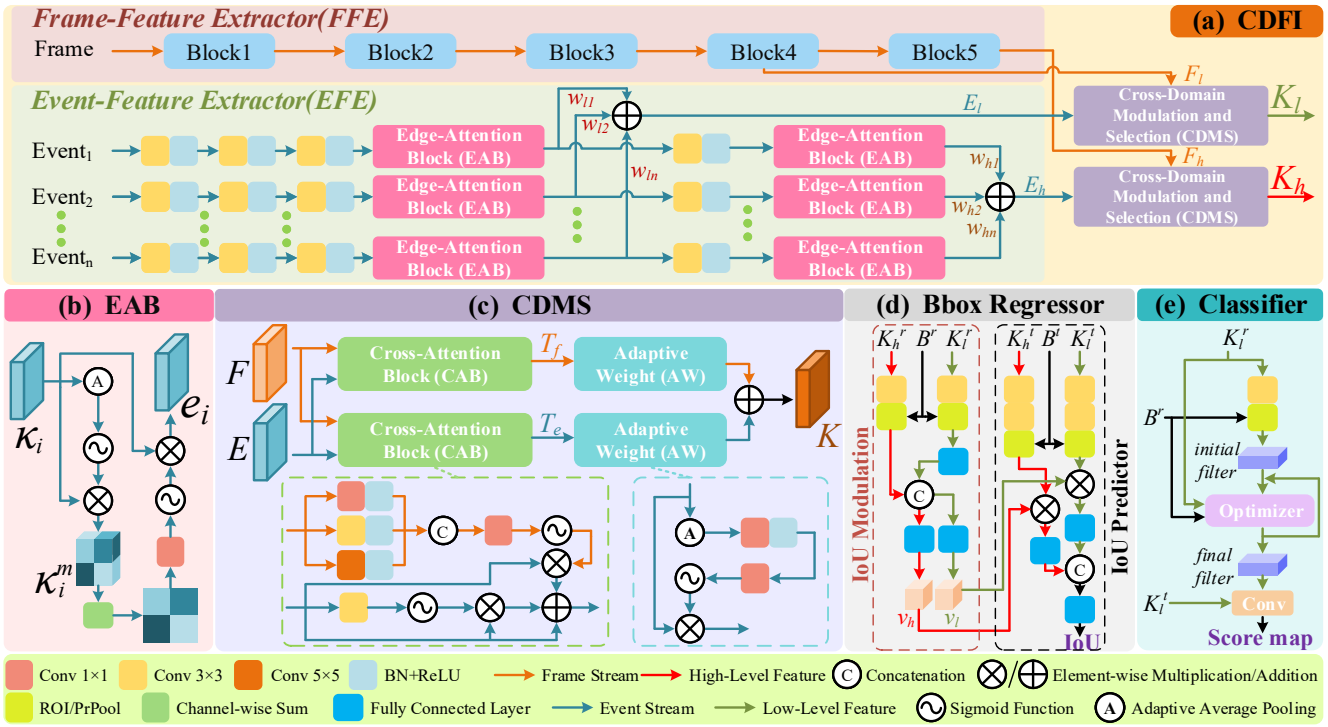


Figure 3. Detailed architectures of the proposed components. (a) Overview of Cross-Domain Feature Integrator (CDFI), (b) Edge-Attention Block (EAB), (c) Cross-Domain Modulation and Selection block (CDMS), (d) Bbox Regressor, and (e) Classifier.

pooling, respectively; $\oplus$/$\otimes$ indicates element-wise summation/multiplication;

CDMS is designed to fuse the extracted frame and event features, shown in Figure 3 (c). The key to the proposed CDMS is a cross-domain attention scheme designed based on the following observations: (i) Rich textural and semantic cues can easily be captured by a conventional frame-based sensor, whereas an event-based camera can easily capture edge information. (ii) The cues provided by a conventional frame-based sensor become less effective in challenging scenarios. By contrast, an event-based camera does not suffer from these scenarios. (iii) In the case of multiple moving objects crossing each other, it is hard to separate them trivially based on edges. However, the problem can be addressed well with texture information.

To address the first observation, we design a Cross-Attention Block (CAB) to fuse features of the two domains based on cross-domain attentions. Specifically, given two features from two different domains, D_1 and D_2, we define the following cross-domain attention scheme to generate an enhanced feature for D_1:

$$T_{D_1} = T_{D_1}^{1\to1} \oplus T_{D_1}^{2\to1} \oplus D_1 \tag{6}$$

$$T_{D_1}^{1\to1} = \sigma(\psi_{3\times3}(D_1)) \otimes D_1, \tag{7}$$

$$T_{D_1}^{2\to1} = \sigma(\psi_{1\times1}[\xi(\psi_{1\times1}(D_2)), \xi(\psi_{3\times3}(D_2)), \xi(\psi_{5\times5}(D_2))]) \otimes D_1, \tag{8}$$

where $[\cdot]$ indicates channel-wise concatenation; ξ is the Batch Normalization (BN) followed by a ReLU activation function; $T_{D_1}^{1\to1}$ indicates a self-attention based on D_1. $T_{D_1}^{2\to1}$ is a cross-domain attention scheme based on D_2 to enhance the feature of D_1. When D_1 and D_2 represent the event- and frame-domain, the enhanced feature of the event-domain, T_e, is obtained. Inversely, the enhanced feature of the frame-domain, T_f, can be generated.

To address the second and third observations, we propose an adaptive weighted balance scheme to balance the contribution of the frame- and event- domains:

$$K = W_f T_f \oplus W_e T_e, \tag{9}$$

$$W_{D_i} = \sigma(\psi_{1\times1}(\xi(\psi_{1\times1}(\mathcal{A}(T_{D_i}))))). \tag{10}$$

3.3.2 Bounding Box (BBox) Regressor and Classifier

For the BBox regressor and classifier, we adopt the target estimation network of ATOM[13] and the classifier of DiMP [5], respectively. The architecture of BBox regressor is shown in Figure 3 (d). The IoU modulation 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:

$$v_l = \mathcal{F}(q), \qquad v_h = \mathcal{F}(q),$$
$$q = [\mathcal{F}(\mathcal{P}(\psi_{3\times3}(K_l^r), B^r)), \mathcal{P}(\psi_{3\times3}(K_h^r), B^r)] \tag{11}$$

where $\mathcal{F}$ is fully connected layer; $\mathcal{P}$ denotes PrPool [23]; B^r is the target bounding box from reference branch. The

IoU predictor predicts IoU based on the following equation:

$$IoU = \mathcal{F}([\mathcal{F}(\mathcal{P}(\psi_{3\times3}(\psi_{3\times3}(K_l^t)), B^t) \otimes v_l), \\ \mathcal{F}(\mathcal{P}(\psi_{3\times3}(\psi_{3\times3}(K_h^t)), B^t) \otimes v_h)]) \quad (12)$$

For the classifier, following [5], we use it to predict a target confidence score. As shown in Figure 3 (e), the classifier 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 robustly discriminate between the target object and background distractors.

3.4. Loss Function

We adopt the loss function of [5], which is defined as:

$$L_{\text{tot}} = \beta L_{\text{cls}} + L_{\text{b}}, \quad (13)$$

$$L_{cls} = \frac{1}{N}\sum_{i=1}^{N}(\ell\left(s_i, z_c\right))^2, \quad (14)$$

$$\ell(s_i, z_c) = \begin{cases} s_i - z_c, & z_c > 0.05 \\ \max(0, s_i), & z_c \leq 0.05, \end{cases} \quad (15)$$

$$L_b = \frac{1}{N}\sum_{i=1}^{N}(IoU_i - IoU_{gt})^2, \quad (16)$$

where s_i is the i-th classification score predicted by the classifier, and z_c is obtained by setting to a Gaussian function centered as the target c. The loss function has two components: classification loss L_{cls}, and bounding box regressor loss L_b. The L_{cls} estimates Mean Squared Error (MSE) between s_i and z_c. The idea behind Eq. 15 is to alleviate the impact of unbalanced negative samples (*i.e.*, background). A hinge function is applied to clip the scores at zero in the background region so that the model can equally focus on both positive and negative samples. The L_b estimates MSE between the predicted IoU overlap IoU_i obtained from test branch and the ground truth IoU_{gt}.

4. Dataset

Currently, Hu *et al.* [22] collected a dataset by placing an event-based camera in front of a monitor and recorded large-scale annotated RGB/grayscale videos (*e.g.*, VOT2015 [25]). However, the dataset based on RGB tracking benchmarks cannot faithfully represent events captured in real scenes since the events between adjacent frames are missing. Mitrokhin *et al.* [35, 36] collected two event-based tracking datasets: EED [35] and EV-IMO [36]. As shown in Table 1, the EED only has 179 frames (7.8 seconds) 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. To enable further research on multi-modal learning with events, we collect a large-scale dataset termed FE108, which has 108 sequences with a total length of 1.5 hours. The dataset contains 21 different types of objects and covers four challenging scenarios. The annotation frequency is up to 20/40 Hz for the frame domain (20 out of 108 sequences are 20Hz) and 240 Hz for the event domain.

	Classes	Frames	Events	Time	Frame(Hz)	Event(Hz)
EED [35]	2	179	3.4M	7.8s	23	23
EV-IMO [36]	3	76,800	–	32.0m	40	200
Ours	**21**	**208,672**	**5.6G**	**96.9m**	**20/40**	**240**

Table 1. **Analysis of existing event-based datasets.** Our FE108 offers the best in terms of all listed metrics.

4.1. Dataset Collection and Annotation

The FE108 dataset is captured by a DAVIS346 event-based camera [7], which equips a 346×260 pixels dynamic vision sensor (DVS) and an active pixel sensor (APS). It can simultaneously provide events and aligned grayscale images of a scene. The ground truth bounding boxes of a moving target are provided by the Vicon motion capture system [1], which captures motion with a high sampling rate (up to 330Hz) and sub-millimeter precision. During the capturing process, we fix APS's frame rate to 20/40 FPS and Vicon's sampling rate to 240Hz, which are also the annotation frequency of the captured APS frame and accumulated events (*i.e.*, accumulated every $1/240$ second), respectively.

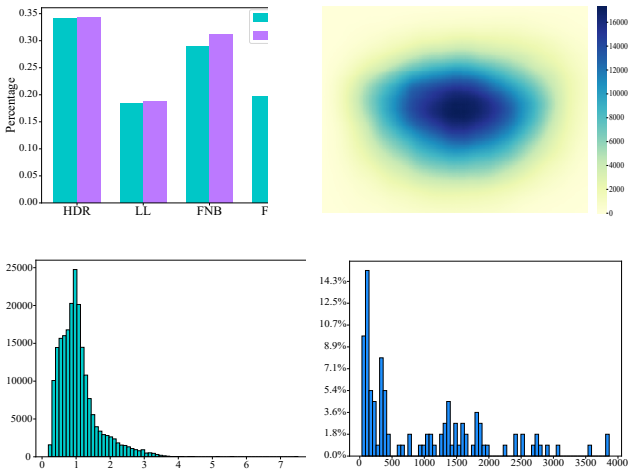
4.2. Dataset Facts

We introduce critical aspects of the constructed FE108. More details about the FE108 are described in the *supplementary material*.

Categorical Analysis. The FE108 dataset can be categorized differently from different perspectives. The first perspective is the number of object classes. There are 21 different object classes, which can be divided into three categories: animals, vehicles, and daily goods (*e.g.,* bottle, box). Second, as shown in Figure 4 (a), the FE108 contains four types of challenging scenes: low-light (LL), high dynamic range (HDR), fast motion with and without motion blur on APS frame (FWB and FNB). Third, based on the camera movement and number of objects, FE108 has four types of scenes: static shots with a single object or multiple objects; dynamic shots with a single object or multiple objects.

Annotated Bounding Box Statistics. In Figure 4 (b), we plot out the distribution of all annotated bounding box locations, which shows most annotations are close to frames' centers. In Figure 4 (c), we also show the distribution of the bounding box aspect ratios (H/W) .

Event Rate. The FE108 dataset is collected in a constant lighting condition. It means all events are triggered by mo-



(c) Histogram of aspect ratios (d) Avg event rate (Ev/ms)

Figure 4. Statistics of FE108 dataset in terms of (a) attributes, (b) bounding box center position, (c) aspect ratios, and (d) event rate.

tions (*e.g.*, moving objects, camera motion). Therefore, the distribution of the event rate can represent the motion distribution of FE108. As shown in Figure 4 (d), the distribution of the event rate is diverse. It indicates the captured 108 scenes offer wide motion diversity.

5. Experiments

We implement the proposed network in PyTorch [38]. In the training phase, random initialization is used for all components except the FFE (which is a ResNet18 pre-trained on ImageNet). The initial learning rate for the classifier, the bbox regressor, and the CDFI are set to 1e-3, 1e-3, and 1e-4, respectively. The learning rate is adjusted by a decay scheduler, which is scaled by 0.2 for every 15 epochs. We use Adam optimizer to train the network for 50 epochs. The batch size is set to 26. It takes about 20 hours on a 20-core i9-10900K 3.7 GHz CPU, 64 GB RAM, and an NVIDIA RTX3090 GPU.

5.1. Comparison with State-of-the-art Trackers

To validate the effectiveness of our method, we compare the proposed approach with the following eight state-of-the-art frame-based trackers: SiamRPN [28], ATOM [13], DiMP [5], SiamFC++ [45], SiamBAN [11], KYS [6], CLNet [16], and PrDiMP [15]. 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

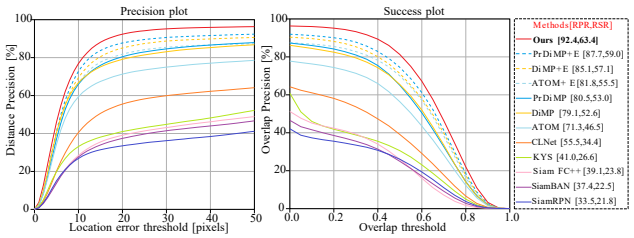


Figure 5. Precision (left) and Success (right) plot on FE108. In terms of both metric, our approach outperforms the state-of-the-art by a large margin.

(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).

Illustrated as the solid curves in Figure 5, on FE108 dataset, our method outperforms other compared approaches by a large margin in terms of both precision and success rate. Specifically, the proposed approach achieves a 92.4% overall RPR and 63.4% RSR, and it outperforms the runner-up by 11.9% and 10.4%, respectively. To get more insights into the effectiveness of the proposed approach, we also show the performances under four different challenging conditions provided by FE108. As shown in Table 2, our method offers the best results under all four conditions, especially in LL and HDR conditions. Eight visual examples under different degraded conditions are shown in Figure 7, where we can see our approach can accurately track the target under all conditions.

Even though EED [35] has very limited frames and associated events, it provides five challenging sequences: fast drone (FD), light variations (LV), occlusions (Occ), what is background (WiB), and multiple objects (MO). 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. Therefore, we also compare our approach against other methods on EED [35]. The experimental results are shown in Figure 6 and Table 3. Our method significantly outperforms other approaches in all conditions except WiB. But with limited frames, the experimental result is less convincing and meaningful compared to the ones obtained from FE108.

One question in our mind is whether combining the

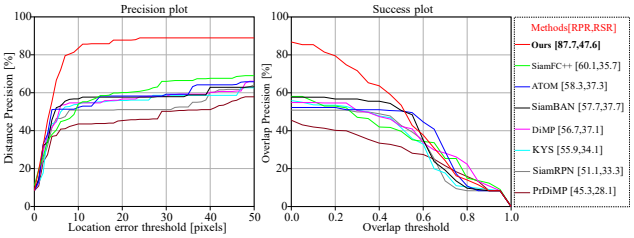


Figure 6. Precision (left) and Success (right) plot on EED [35].

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
SiamRPN [28]	15.3	16.9	6.1	21.6	10.1	8.3	1.4	14.5	26.2	32.1	6.1	44.1	33.2	42.9	11.5	51.9	21.8	26.1	7.0	33.5
ATOM [13]	36.6	41.8	14.4	56.0	28.6	29.1	5.8	45.0	66.8	89.6	32.6	96.7	57.1	71.0	28.0	88.6	46.5	56.4	20.1	71.3
DiMP [5]	41.8	50.0	17.9	62.7	45.6	52.8	11.2	69.5	69.4	**94.7**	37.1	99.7	60.5	75.6	29.3	93.2	52.6	65.4	23.4	79.1
SiamFC++ [45]	15.3	15.0	1.3	25.2	13.4	8.7	0.8	15.3	28.6	36.3	6.0	48.2	36.8	42.7	7.4	63.1	23.8	26.0	3.9	39.1
SiamBAN [11]	16.3	16.4	3.9	26.6	15.5	14.8	2.3	26.5	25.2	26.3	5.8	46.7	32.0	39.6	9.1	51.4	22.5	25.0	5.6	37.4
KYS [6]	15.7	14.5	5.2	23.0	12.0	8.0	1.1	18.0	47.0	63.9	14.8	73.3	36.9	44.5	15.2	57.9	26.6	30.6	9.2	41.0
CLNet [16]	30.0	33.5	9.6	48.3	13.7	6.0	0.9	23.6	52.9	71.2	23.3	80.3	40.8	46.3	14.2	67.7	34.4	39.1	11.8	55.5
PrDiMP [15]	44.3	52.8	19.6	66.3	44.6	48.2	8.9	69.5	67.0	89.9	33.6	99.7	60.6	75.8	29.7	93.3	53.0	65.0	23.3	80.5
ATOM [13] + Event	49.0	59.2	21.0	68.8	50.8	67.8	27.7	72.6	68.5	90.4	42.0	97.2	57.4	71.1	28.3	90.2	55.5	70.0	27.4	81.8
DiMP [5] + Event	50.1	60.2	23.7	74.8	57.0	70.4	28.2	82.8	70.1	94.2	44.2	99.9	60.8	75.9	29.1	93.6	57.1	71.2	28.6	85.1
PrDiMP [15] + Event	53.1	65.3	24.9	79.1	60.3	79.6	29.8	90.5	70.0	93.8	44.8	99.8	61.8	76.3	29.4	93.6	59.0	74.4	29.8	87.7
Ours	**59.9**	**74.4**	**33.0**	**86.0**	**65.6**	**86.0**	**30.8**	**95.7**	**71.2**	**94.7**	**45.9**	**100.0**	**62.8**	**80.5**	**32.0**	**94.5**	**63.4**	**81.3**	**34.4**	**92.4**

Table 2. State-of-the-art comparison on FE108 in terms of representative success rate (RSR), representative precision rate (RPR), and overlap precision (OP).

Methods	FD		LV		Occ		WiB		MO		ALL	
	RSR	RPR	RSR	RPR	RSR	RPR	RSR	RPR	RSR	RPR	RSR	RPR
SiamRPN [28]	23	43	11	10	38	40	43	63	53	100	33	51
ATOM [13]	12	19	7	12	47	60	74	100	47	100	37	58
DiMP [5]	9	19	2	4	48	60	**79**	100	50	100	37	57
SiamFC++ [45]	17	52	10	26	45	60	58	63	50	100	36	60
SiamBAN [11]	22	43	8	6	36	40	69	100	54	100	38	58
KYS [6]	19	38	6	19	46	60	46	63	54	100	34	56
CLNet [16]	10	19	2	6	19	20	13	25	4	13	9	17
PrDiMP [15]	9	14	4	22	19	20	78	100	31	70	28	45
Ours	**32**	**81**	**35**	**98**	**48**	**60**	69	**100**	**55**	**100**	**48**	**88**

Table 3. State-of-the-art comparison on EED [35] in terms of RSR and RPR.

frame and event information can make other frame-based approaches outperform our approach. To answer this question, we combine APS and event aggregated frame by concatenation manner to train and test the top three frame-based performers (*i.e.*, PrDiMP [15], DiMP [5], and ATOM [13]). We report their RSR and RPR in Table 2 and show the corresponding results as the dashed curve in Figure 5. As we can see, our approach still outperforms all others by a considerable margin. It reflects the effectiveness of our specially designed 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.

5.2. Ablation Study

Multi-modal Input. We design the following experiments to show the effectiveness of multi-modal input. 1. Frame only: only using frames and **FFE**; 2. Event only: only using events and **EFE**; 3. Event to Frame: combining frames and events by concatenation as input to **FFE**; 4. Frame to Event: the same as 3, but input to **EFE**. For each setup, we train a dedicated model and test with it. As shown in the row *A*-*D* of Table 4, the models with multi-modal inputs perform better than the ones with unimodal input. It shows the effectiveness of multi-modal fusion and our CDFI.

Effectiveness of the proposed key components. There are

	Models	RSR ↑	$OP_{0.50}$ ↑	$OP_{0.75}$ ↑	RPR ↑
A.	Frame Only	45.6	54.6	21.0	73.1
B.	Event Only	52.0	63.2	20.3	82.0
C.	Event to Frame	55.5	70.0	27.4	82.8
D.	Frame to Event	53.6	66.5	25.9	80.4
E.	w/o EAB	60.7	77.9	31.7	88.6
F.	w/o CDMS	59.8	75.8	31.0	88.1
G.	CDMS w/o SA	62.6	79.8	33.8	91.5
H.	CDMS w/o CA	61.9	78.8	33.0	90.7
I.	CDMS w/o AW	60.9	77.2	32.0	89.9
J.	TSLTD [9]	60.4	77.0	31.2	89.2
K.	Time Surfaces [26]	61.4	78.5	32.9	90.1
L.	Event Count [32]	59.6	76.4	27.4	88.6
M.	Event Frame [40]	59.0	74.5	29.9	87.7
N.	Zhu *et al.* [50]	61.9	79.2	32.3	91.2
O.	All $w = 1$	61.3	78.1	31.6	90.1
P.	Ours	**63.4**	**81.3**	**34.4**	**92.4**

Table 4. Ablation study results.

two key components in our approach: EAB and CDMS. Inside the CDMS, there are three primary schemes: self-attention (Eq. 7), cross-attention (Eq. 8), and adaptive weighting (Eq. 9). To verify their effectiveness, we modify the original model by dropping each of the components and retrain the modified models. Correspondingly, we obtain five retrained models: (i) without EAB; (ii) without CDMS; Inside CDMS, (iii) without self-attention (CDMS w/o SA); (iv) without cross-attention (CDMS w/o CA); (v) without adaptive weighting (CDMS w/o AW). The results of the five modified models are shown in the row *E*-*I* of Table 4, respectively. Compared to the original model, removing CDMS has the most considerable impact on the performance, whereas removing the self-attention influences the least. It confirms the proposed CDMS is the key to our outstanding performance. Moreover, removing EAB also influences performance significantly. It shows that the EAB indeed enhances the extracted edge features.

Inside CDMS, removing adaptive weighting scheme degrades performance the most. To get more insights into it, we report the estimated two weights (*i.e.*, w_f for the frame domain; w_e for the event domain) of all eight visual exam-

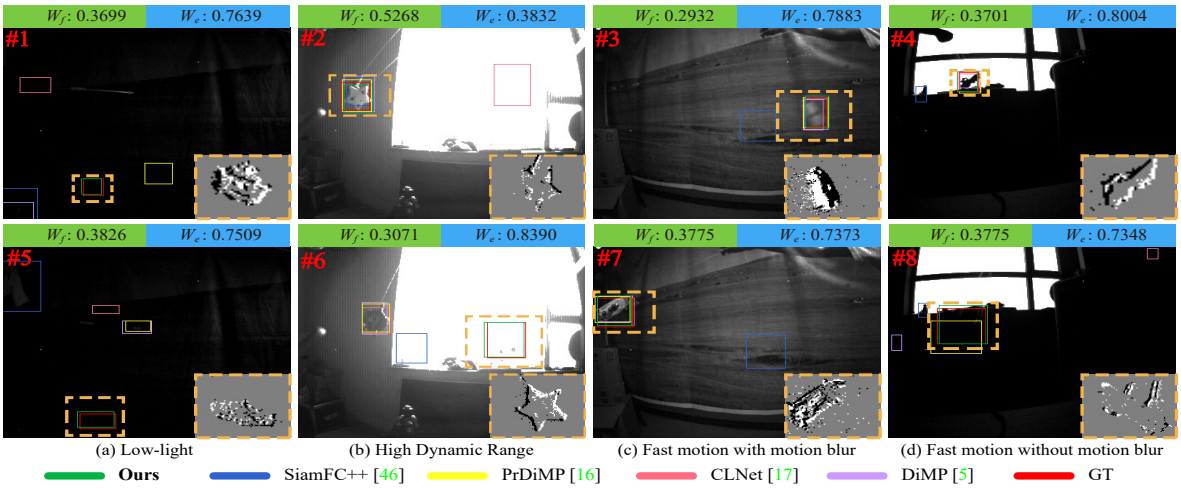


Figure 7. Visual outputs of state-of-the-art algorithms on FE108 dataset. The lower-right dashed boxes show accumulated event frame of the dashed boxes inside the frames.

ples in Figure 7. Except for the second one, the frame domain cannot provide reliable visual cues. Correspondingly, we can see the w_e in these seven examples are significantly higher than w_f, whereas w_e is much lower than w_f in the second scene. The fourth one provides an interesting observation. We can see the object clearly in the frame domain, but w_e is still higher than w_f. We think it is because the model is trained to focus on texture cues in the frame domain, but no texture cues can be extracted in this case. It is worthwhile to mention that only our method can successfully track the target in all examples.

Event Aggregation. For the events captured between two adjacent frames, we slice them into n chunks in the time domain and then aggregate them as EFE's inputs. Here, we study the impacts of hyperparameter n. As shown in Table 5, both RSR and RPR scores increase with a larger 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. The way of aggregating events is another factor that has an impact on the performance. We conducted experiments with five commonly used event aggregation methods [9, 26, 32, 40, 50]. The results are shown in the row *J-N* of Table 4, and our method still delivers the best performance. It suggests that discretizing the time dimension and leveraging the recent timpstamp information are effective for tracking. Another component associated with event aggregation is the weights in Eq. 3, which are learned during the training process. We manually set the weights to 1 with $n = 3$. The result is shown in row *O* of Table 4, and we can see the corresponding performance is worse than the original model.

n	1	2	3	4	5	6
RPR ↑	89.3	90.1	92.4	92.6	92.6	92.7
RSR ↑	60.2	61.7	63.4	63.8	63.4	63.9
FPS ↓	35.1	32.7	30.1	27.9	25.2	22.7

Table 5. Trade-off between accuracy and efficiency introduced by the number of slices of event aggregation (*i.e.*, n).

6. Discussion and Conclusion

In this paper, we introduce a frame-event fusion-based approach for single object tracking. Our novel designed attention schemes effectively fuse the information obtained from both the frame and event domains. Besides, the novel developed weighting scheme is able to balance the contributions of the two domains adaptively. To enable further research on multi-modal learning and object tracking with events, we construct a large-scale dataset, FE108, comprising events, frames, and high-frequency annotations. Our approach outperforms frame-based state-of-the-art methods, which indicates leveraging the complementarity of events and frames boosts the robustness of object tracking in degraded conditions. Our current focus is on developing a cross-domain fusion scheme that can enhance visual tracking robustness, especially in degraded conditions. However, we have not leveraged the high measurement rate of event-based cameras to achieve low-latency tracking and the frame rate in the frame-domain bounds the tracking frequency of the proposed approach. One limitation of our frame-event-based dataset, FE108, is that no sequence contains the scenario of no events. Our further work will focus on these two aspects: 1) We will investigate the feasibility of increasing tracking frequency by leveraging the high measurement rate of event-based cameras; 2) We will expand the FE108 by collecting more challenging sequences, especially with no events and more realistic scenes.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 61632006, Grant 61972067, and the Innovation Technology Funding of Dalian (Project No. 2018J11CY010, 2020JJ26GX036).

References

[1] Vicon motion capture. https://www.vicon.com/. 5

[2] Ning An, Xiao-Guang Zhao, and Zeng-Guang Hou. Online rgb-d tracking via detection-learning-segmentation. In *ICPR*, 2016. 2

[3] Francisco Barranco, Cornelia Fermuller, and Eduardo Ros. Real-time clustering and multi-target tracking using event-based sensors. In *IROS*, 2018. 2

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

[5] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *ICCV*, 2019. 4, 5, 6, 7

[6] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Know your surroundings: Exploiting scene information for object tracking. In *ECCV*, 2020. 1, 6, 7

[7] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 2014. 1, 2, 5

[8] Massimo Camplani, Sion L Hannuna, Majid Mirmehdi, Dima Damen, Adeline Paiement, Lili Tao, and Tilo Burghardt. Real-time rgb-d tracking with depth scaling kernelised correlation filters and occlusion handling. In *BMVC*, 2015. 2

[9] 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 *AAAI*, 2020. 2, 7, 8

[10] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *CVPR*, 2020. 1

[11] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *CVPR*, 2020. 6, 7

[12] Kenan Dai, Dong Wang, Huchuan Lu, Chong Sun, and Jianhua Li. Visual tracking via adaptive spatially-regularized correlation filters. In *CVPR*, 2019. 1, 2

[13] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *CVPR*, 2019. 4, 6, 7

[14] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *CVPR*, 2017. 1, 2

[15] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *CVPR*, 2020. 6, 7

[16] Xingping Dong, Jianbing Shen, Ling Shao, and Fatih Porikli. Clnet: A compact latent network for fast adjusting siamese trackers. In *ECCV*, 2020. 6, 7

[17] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[18] Jin Gao, Weiming Hu, and Yan Lu. Recursive least-squares estimator-aided online learning for visual tracking. In *CVPR*, 2020. 1

[19] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *ICCV*, 2019. 3

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

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[22] Yuhuang Hu, Hongjie Liu, Michael Pfeiffer, and Tobi Delbruck. Dvs benchmark datasets for object tracking, action recognition, and object recognition. *Frontiers in neuroscience*, 2016. 5

[23] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *ECCV*, 2018. 4

[24] Uğur Kart, Joni-Kristian Kämäräinen, Jiří Matas, and Jiri Matas. How to make an rgbd tracker? In *ECCV*, 2018. 2

[25] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Tomas Vojir, Gustav Hager, Georg Nebehay, and Roman Pflugfelder. The visual object tracking vot2015 challenge results. In *ICCVW*, 2015. 5

[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 transactions on pattern analysis and machine intelligence*, 2016. 2, 3, 7, 8

[27] Xiangyuan Lan, Mang Ye, Shengping Zhang, and Pong C Yuen. Robust collaborative discriminative learning for rgb-infrared tracking. In *AAAI*, 2018. 2

[28] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, 2018. 6, 7

[29] Chenglong Li, Chengli Zhu, Yan Huang, Jin Tang, and Liang Wang. Cross-modal ranking with soft consistency and noisy labels for robust rgb-t tracking. In *ECCV*, 2018. 2

[30] Peixia Li, Boyu Chen, Wanli Ouyang, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Gradnet: Gradient-guided network for visual object tracking. In *ICCV*, 2019. 1, 2

[31] Cheng Long Li, Andong Lu, Ai Hua Zheng, Zhengzheng Tu, and Jin Tang. Multi-adapter rgbt tracking. In *ICCVW*, 2019. 2

[32] Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *CVPR*, 2018. 2, 3, 7, 8

[33] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. In *CVPR*, 2021. 2

[34] Nico Messikommer, Daniel Gehrig, Antonio Loquercio, and Davide Scaramuzza. Event-based asynchronous sparse convolutional networks. In *ECCV*, 2020. 3

[35] Anton Mitrokhin, Cornelia Fermüller, Chethan Parameshwara, and Yiannis Aloimonos. Event-based moving object detection and tracking. In *IROS*, 2018. 2, 5, 6, 7

[36] Anton Mitrokhin, Chengxi Ye, Cornelia Fermuller, Yiannis Aloimonos, and Tobi Delbruck. Ev-imo: Motion segmentation dataset and learning pipeline for event cameras. In *IROS*, 2019. 5

[37] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR*, 2016. 1, 2

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*, 2019. 6

[39] Ewa Piatkowska, Ahmed Nabil Belbachir, Stephan Schraml, and Margrit Gelautz. Spatiotemporal multiple persons tracking using dynamic vision sensor. In *CVPRW*, 2012. 2

[40] Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Real-time visual-inertial odometry for event cameras using keyframe-based nonlinear optimization. In *BMVC*, 2017. 2, 3, 7, 8

[41] Shuran Song and Jianxiong Xiao. Tracking revisited using rgbd camera: Unified benchmark and baselines. In *ICCV*, 2013. 2

[42] Chaoqun Wang, Chunyan Xu, Zhen Cui, Ling Zhou, Tong Zhang, Xiaoya Zhang, and Jian Yang. Cross-modal pattern-propagation for rgb-t tracking. In *CVPR*, 2020. 2

[43] Yanxiang Wang, Xian Zhang, Yiran Shen, Bowen Du, Guangrong Zhao, Lizhen Cui Cui Lizhen, and Hongkai Wen. Event-stream representation for human gaits identification using deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[44] Jingjing Xiao, Rustam Stolkin, Yuqing Gao, and Aleš Leonardis. 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*, 2017. 2

[45] Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *AAAI*, 2020. 6, 7

[46] Jiqing Zhang, Kai Zhao, Bo Dong, Yingkai Fu, Yuxin Wang, Xin Yang, and Baocai Yin. Multi-domain collaborative feature representation for robust visual object tracking. *The Visual Computer*, 2021. 1

[47] Lichao Zhang, Martin Danelljan, Abel Gonzalez-Garcia, Joost van de Weijer, and Fahad Shahbaz Khan. Multi-modal fusion for end-to-end rgb-t tracking. In *ICCVW*, 2019. 2

[48] Tianzhu Zhang, Changsheng Xu, and Ming-Hsuan Yang. Learning multi-task correlation particle filters for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 1, 2

[49] Zhipeng Zhang and Houwen Peng. Deeper and wider siamese networks for real-time visual tracking. In *CVPR*, 2019. 1, 2

[50] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *CVPR*, 2019. 2, 3, 7, 8

[51] Yabin Zhu, Chenglong Li, Bin Luo, Jin Tang, and Xiao Wang. Dense feature aggregation and pruning for rgbt tracking. In *ACM MM*, 2019. 2