Frame-Event Alignment and Fusion Network for High Frame Rate Tracking

Jiqing Zhang[1], Yuanchen Wang[1], Wenxi Liu[2], Meng Li[3], Jinpeng Bai[1], Baocai Yin[1], Xin Yang[1,⋆]
[1]Dalian University of Technology, [2]Fuzhou University, [3]HiSilicon(Shanghai) Technologies Co.,Ltd

Abstract

Most existing RGB-based trackers target low frame rate benchmarks of around 30 frames per second. This setting restricts the tracker's functionality in the real world, especially for fast motion. Event-based cameras as bioinspired sensors provide considerable potential for high frame rate tracking due to their high temporal resolution. However, event-based cameras cannot offer fine-grained texture information like conventional cameras. This unique complementarity motivates us to combine conventional frames and events for high frame rate object tracking under various challenging conditions. In this paper, we propose an end-to-end network consisting of multi-modality alignment and fusion modules to effectively combine meaningful information from both modalities at different measurement rates. The alignment module is responsible for cross-style and cross-frame-rate alignment between frame and event modalities under the guidance of the moving cues furnished by events. While the fusion module is accountable for emphasizing valuable features and suppressing noise information by the mutual complement between the two modalities. Extensive experiments show that the proposed approach outperforms state-of-the-art trackers by a significant margin in high frame rate tracking. With the FE240hz dataset, our approach achieves high frame rate tracking up to 240Hz.

1. Introduction

Visual object tracking is a fundamental task in computer vision, and deep learning-based methods [7,9,10,15,35,56] have dominated this field. Limited by the conventional sensor, most existing approaches are designed and evaluated on benchmarks [13,24,38,53] with a low frame rate of approximately 30 frames per second (FPS). However, the value of a higher frame rate tracking in the real world has been proved [16,21–23]. For example, the shuttlecock can reach speeds of up to $493km/h$, and analyzing its position is essential for athletes to learn how to improve their skills [46]. Utilizing professional high-speed cameras is one strategy for high frame rate tracking, but these cameras are inaccessible to casual users. Consumer devices with cameras, such as smartphones, have made attempts to integrate sensors with similar functionalities into their systems. However, these sensors still suffer from large memory requirements and high power consumption [49].

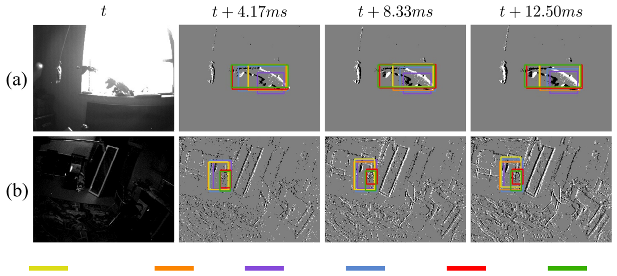


Figure 1. A comparison of our AFNet with SOTA trackers. All competing trackers locate the target at time $t + \Delta t$ with conventional frames at time t and aggregated events at time $t + \Delta t$ as inputs. Our method achieves high frame rate tracking up to 240Hz on the FE240hz dataset. The two examples also show the complementary benefits of both modalities. (a) The event modality does not suffer from HDR, but the frame does; (b) The frame modality provides rich texture information, while the events are sparse.

As bio-inspired sensors, event-based cameras measure light intensity changes and output asynchronous events to represent visual information. Compared with conventional frame-based sensors, event-based cameras offer a high measurement rate (up to 1MHz), high dynamic range (140 dB vs. 60 dB), low power consumption, and high pixel bandwidth (on the order of kHz) [14]. These unique properties offer great potential for higher frame rate tracking in challenging conditions. Nevertheless, event-based cameras cannot measure fine-grained texture information like conventional cameras, thus inhibiting tracking performance. Therefore, in this paper, we exploit to integrate the valuable information from event-based modality with that of frame-based modality for high frame rate single object tracking under various challenging conditions.

To attain our objective, two challenges require to be addressed: (i) The measurement rate of event-based cameras is much higher than that of conventional cameras. Hence

[⋆] Xin Yang (xinyang@dlut.edu.cn) is the corresponding author.

for high frame rate tracking, low-frequency frames must be aligned with high-frequency events to disambiguate target locations. Although recent works [34, 45, 48, 50] have proposed various alignment strategies across multiple frames for video-related tasks, they are specifically designed for conventional frames of the same modality at different moments. Thus, applying these approaches directly to our cross-modality alignment does not offer an effective solution. (ii) Effectively fusing complementary information between modalities and preventing interference from noise is another challenge. Recently, Zhang *et al.* [61] proposed a cross-domain attention scheme to fuse visual cues from frame and event modalities for improving the single object tracking performance under different degraded conditions. However, the tracking frequency is bounded by the conventional frame rate since they ignore the rich temporal information recorded in the event modality.

To tackle the above challenges, we propose a novel end-to-end framework to effectively combine complementary information from two modalities at different measurement rates for high frame rate tracking, dubbed AFNet, which consists of two key components for alignment and fusion, respectively. Specifically, (i) we first propose an event-guided cross-modality alignment (ECA) module to simultaneously accomplish cross-style alignment and cross-frame-rate alignment. Cross-style alignment is enforced by matching feature statistics between conventional frame modality and events augmented by a well-designed attention scheme; Cross-frame-rate alignment is based on deformable convolution [8] to facilitate alignment without explicit motion estimation or image warping operation by implicitly focusing on motion cues. (ii) A cross-correlation fusion (CF) module is further presented to combine complementary information by learning a dynamic filter from one modality that contributes to the feature expression of another modality, thereby emphasizing valuable information and suppressing interference. Extensive experiments on different event-based tracking datasets validate the effectiveness of the proposed approach (see Figure 1 as an example).

In summary, we make the following contributions:

- Our AFNet is, to our knowledge, the first to combine the rich textural clues of frames with the high temporal resolution offered by events for high frame rate object tracking.

- We design a novel event-guided alignment framework that performs cross-modality and cross-frame-rate alignment simultaneously, as well as a cross-correlation fusion architecture that complements the two modalities.

- Through extensive experiments, we show that the proposed approach outperforms state-of-the-art trackers in various challenging conditions.

2. Related Work

2.1. Visual Object Tracking

Visual object tracking based on the conventional frame has undergone astonishing progress in recent years, which can be generally divided into two categories, *i.e.*, correlation filter (CF) trackers [1, 4, 18, 33], and deep trackers [2, 26, 39, 55, 64, 65]. CF trackers learn a filter corresponding to the object of interest in the first frame, and this filter is used to locate the target in subsequent frames. While mainstream deep trackers estimate a general similarity map by cross-correlation between template and search images. However, limited by sensors and benchmarks, those methods are mainly applied to low frame rate (30FPS) tracking.

The high temporal resolution of event cameras allows tracking targets at a higher frame rate. Compared with conventional frame-based tracking, a few attempts have been made at event-based tracking, which can be generally classified into cluster-based and learning-based methods. Litzenberger *et al.* [28] assigned each new event to a cluster based on distance criteria, which is continuously updated for locating the target. Linares *et al.* [27] used software to initialize the size and location of clusters, then proposed an FPGA-based framework for tracking. Piatkowska *et al.* [41] extended the clustering method by a stochastic prediction of the objects' states to locate multiple persons. However, these methods involve handcrafted strategies and only apply in simple situations. Based on the powerful representation ability of deep learning [30, 52], Chen *et al.* [5, 6] designed two different event representation algorithms based on Time Surface [25] for target location regression. Zhang *et al.* [59] combined Swin-Transformer [31] and spiking neural network [12, 29, 58] to extract spatial and temporal features for improving event-based tracking performance. However, these event-based trackers often fail to locate targets accurately when events are too sparse or insufficient.

To combine benefits from frame and event modalities, [61] employed attention schemes [36, 37, 42, 43, 60] to balance the contributions of the two modalities. This work is most closely related to ours, but it does not exploit the high measurement rate of event-based cameras to accomplish a higher frame rate tracking, thus the tracking frequency is constrained by the frame rate in the frame modality. In contrast, our approach attains high frame rate tracking under various challenging conditions by aligning and fusing frame and event modalities with different measurement rates.

2.2. Alignment between Multiple Frames

Alignment across multiple frames in the same sequence is essential to exploit the temporal information for video-related tasks, such as video super-resolution [48, 50] and compressed video quality enhancement [11, 63]. A line of works [44, 47, 54] performs alignment by estimating the optical flow field between the reference and its neighbouring frames. In another line of works [11,45,48], implicit motion compensation is accomplished by deformable convolution. Deformable convolution was first proposed in [8], which improves the ability of convolutional layers to model geometric transformations by learning additional offsets. Although the deformable convolution has shown superiority in alignment on the conventional frame domain, aligning on the frame and event modalities brings unique challenges caused by the different styles. In this paper, we propose a novel alignment strategy to simultaneously achieve cross-modality and cross-frame-rate alignment.

3. Methodology

3.1. Events 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 a certain 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 $(x, y)^\top$.

Suppose two sequential conventional frames F_i and F_{i+1} are captured at times i and $i+1$, respectively. $E_{i \to i+1} = \{[x_k, y_k, t_k, p_k]\}_{k=0}^{N-1}$ contains N events triggered during the interval $[i, i+1]$. Our goal is to achieve high frame rate tracking by aligning and fusing conventional frame F_i and $E_{i \to t}$ at any time $t \in [i, i+1]$. The apart in time between dual-modality inputs depends on their frame rates. Specifically, $t - i = \frac{n}{\gamma_e}$, where n is an integer in $[1, \frac{\gamma_e}{\gamma_f}]$; γ_e and γ_f denote the frame rates of event and frame modalities, respectively. Following [61], we represent events $E_{i \to t}$ as:

$$g(x, y) = \lfloor \frac{p_k \times \delta(x - x_k, y - y_k, t - t_k) + 1}{2} \times 255 \rfloor, \tag{2}$$

where $g(x, y)$ denotes the pixel value of aggregated events at $(x, y)^\top$; δ is the Dirac delta function. In this way, the asynchronous event stream $E_{i \to t}$ is accumulated to a 2D event frame, denoted E_t.

3.2. Network Overview

Following DiMP [3], as illustrated in Figure 2 (a), the overall architecture of our proposed AFNet contains three components: the feature extractor (*i.e.*, backbone, ECA, and CF), the target classifier, and the bbox regressor. The feature extractors of the template branch and the search branch share the same architecture. Each branch receives an RGB image F_i and aggregated events E_t at different times as inputs, and corresponding features F_f and F_e can be extracted by the backbone network. ECA and CF are two key components of our method. The goal of ECA is to address the misalignment between the conventional frames and aggregated event frames at different moments. While CF aims to combine the strengths of both modalities by complementing one modality with information from another. Both target classifier and bbox regressor receive the fused features from feature extractors. Given a template set of fused features and corresponding target boxes, the model predictor generates the weights of the target classifier. Applying these weights to the features collected from search branch predicts the target confidence scores. The bbox regressor estimates the IoU of the groundtruth and the predicted bounding box.

3.3. Event-guided Cross-modality Alignment

The ECA module is proposed to align conventional frames to the reference aggregated events at the feature level. The key to ECA is designed based on the following challenges: (i) Cross-style alignment is a challenge. Frames and events are recorded by different sensors and thus have different styles, making alignment challenging. (ii) Cross-frame-rate alignment is another challenge. The frame rate of aggregated event frames is far higher than that of conventional images, resulting in target location ambiguity that confuses the tracker's predictions. As shown in Figure 2 (b), ECA contains three modules: Motion Aware (MA), Style Transformer (ST), and Deformable Alignment (DA).

MA. Since event-based cameras respond to changes in light intensity, they provide natural motion cues that can effectively facilitate multi-modality alignment. We thus first enhance the valuable motion information of event modality by visual attention mechanisms. As shown in Figure 2 (b), given event modality features $F_e \in \mathbb{R}^{C \times H \times W}$, we design spatial and channel attention schemes to emphasize the meaningful moving cues while suppressing noise,

$$F_e^c = \sigma(\psi_1(\psi_1(\mathcal{R}^{(C,1,1)}(F_e^s))))F_e, \tag{3}$$

$$F_e^s = \mathcal{R}^{(1,C,HW)}(F_e) \times \mathcal{R}^{(1,HW,1)}(\mathcal{S}(\psi_1(F_e))), \tag{4}$$

where F_e^s and F_e^c are event features enhanced in the spatial and channel dimensions, respectively. ψ_k denotes the convolution operation where kernel size is $k \times k$; $\mathcal{S}$ and σ denote the softmax and the sigmoid function, respectively; $\mathcal{R}(\cdot)$ is a reshape function with a target shape $(\cdot)$.

ST. ST is responsible for combining the content of conventional frames and the style of events to meet the first challenge. Specifically, F_e^c is employed to guide the frame features F_f to focus on the motion cues that aid in alignment,

$$F_f^m = \sigma(F_e^c)F_f + F_f, \tag{5}$$

where F_f^m denotes frame features fused with moving information provided by events. Then, we adopt the adaptive instance normalization (AdaIN) [19] to adjust the mean and

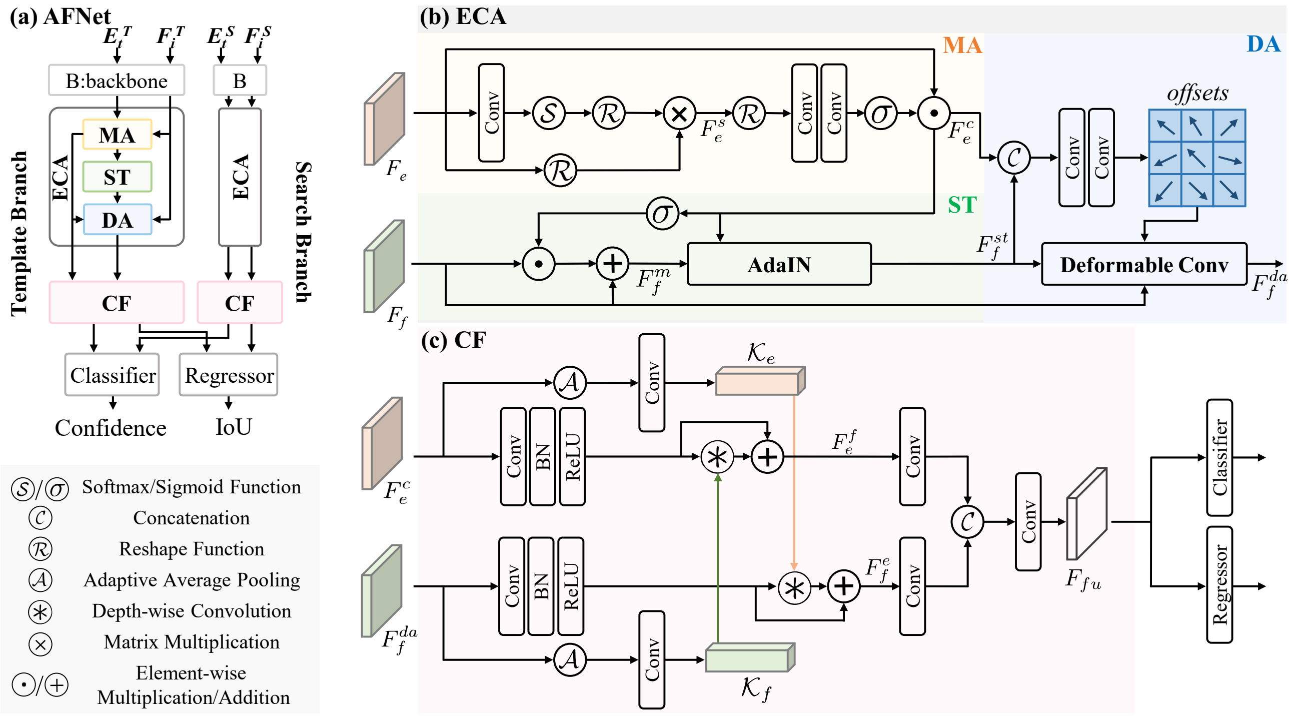


Figure 2. (a) Overview of our AFNet; (b) two key components in the event-guided cross-modality alignment (ECA) module: style transformer (ST) and deformable alignment (DA); and (c) the cross-correlation fusion (CF) module.

variance of the content input (*i.e.*, frame features) to match those of the style input (*i.e.*, event features). Formally,

$$\begin{aligned} F_f^{st} &= \mathrm{AdaIN}(F_f^m, F_e^c) \\ &= \sigma(F_e^c)\left(\frac{F_f^m - \mu(F_f^m)}{\sigma(F_f^m)}\right) + \mu(F_e^c), \end{aligned} \tag{6}$$

where F_f^{st} denotes the output of our ST module, which combines the content of frame modality and the style of event modality. μ and σ are the mean and standard deviation, computed independently across batch size and spatial dimensions for each feature channel.

DA. To address the second challenge, inspired by [50], we propose the DA module to adaptively align the conventional frames and aggregated events at different frame rates without explicit motion estimation and image warping operations. As shown in Figure 2 (b), DA first predict the offsets $\mathcal{O}$ of the convolution kernels according to F_e^c and F_f^{st},

$$\mathcal{O} = \psi_3(\psi_1([F_e^c, F_f^{st}])), \tag{7}$$

where $[\cdot]$ denotes channel-wise concatenation. The learnable offsets will implicitly focus on motion cues and explore similar features across modalities for alignment. With $\mathcal{O}$ and F_f, the aligned feature F_f^{da} of the conventional frame can be computed by the deformable convolution $\mathcal{D}$ [8],

$$F_f^{da} = \mathcal{D}(F_f, \mathcal{O}). \tag{8}$$

3.4. Cross-correlation Fusion

Our CF is proposed to robustly fuse frame and event correlations by adaptively learning a dynamic filter from one modality that contributes to the feature expression of another modality. Simply fusing frame and event modalities ignores circumstance in which one of the modalities does not provide meaningful information. In an HDR scene, for instance, the frame modality will provide no useful information, yet the event modality still exhibits strong cues. Conversely, in the absence of motion, event-based cameras cannot successfully record target-related information, while conventional frames can still deliver rich texture features. Therefore, we propose a cross-correlation scheme to complement one domain with information from another domain as shown in Figure 2 (c). Specifically, given the aligned frame feature F_f^{da} and enhanced event feature F_e^c, the proposed CF first adaptively estimates a dynamic filter of high-level contextual information from one modality. Then, this dynamic filter serves to enhance the features of another modality. Formally,

$$\begin{aligned} F_f^e &= \mathcal{F} \circledast \mathcal{K}_e + \mathcal{F}, \\ \mathcal{F} &= \vartheta(\psi_3(F_f^{da})), \\ \mathcal{K}_e &= \psi_3(\mathcal{A}(F_e^c)), \end{aligned} \tag{9}$$

where F_f^e denotes the enhanced feature of the frame modality based on the dynamic filter $\mathcal{K}_e$ from event modality; $\circledast$

is the depthwise convolution; $\mathcal{A}$ denotes the adaptive average pooling; ϑ is the Batch Normalization (BN) followed by a ReLU activation function. Similarly, we can extract the complementary feature F_e^f of the event modality based on the dynamic filter $\mathcal{K}_f$ from frame modality. Finally, F_f^e and F_e^f are concatenated to build the fused feature F_{fu},

$$F_{fu} = \psi_1([\psi_1(F_f^e), \psi_1(F_e^f)]), \tag{10}$$

F_{fu} will be fed into the classifier and regressor to locate the target. The classifier adopts an effective model initializer and a steepest descent based optimizer to predict the score map. The regressor employs the overlap maximization strategy for the task of accurate bounding box estimation. We refer to [3] for details.

3.5. Implementation Details

We adopt the pretrained ResNet18 [17] as the backbone to extract frame and event features. Following [3, 61], the loss function is defined as:

$$L = \beta L_{cls} + L_{bb}, \tag{11}$$

where L_{cls} is the target classification loss which includes a hinge function to equally focus on both positive and negative samples. L_{bb} is the bounding box regressor loss which estimates MSE between the predicted IoU and the groundtruth. β is set to 100.

We implemented our approach in PyTorch [40] and trained our network for 100 epochs with a batch size of 32 using Adam optimizer with the default parameters. We set the initial learning rate of the feature extraction network, the classifier, and the regressor to 2e-4, 1e-3, 1e-3, respectively. The learning rate is adjusted by the CosineAnnealingLR strategy [32]. Our network is run on a single Nvidia RTX3090 GPU with 24G memory.

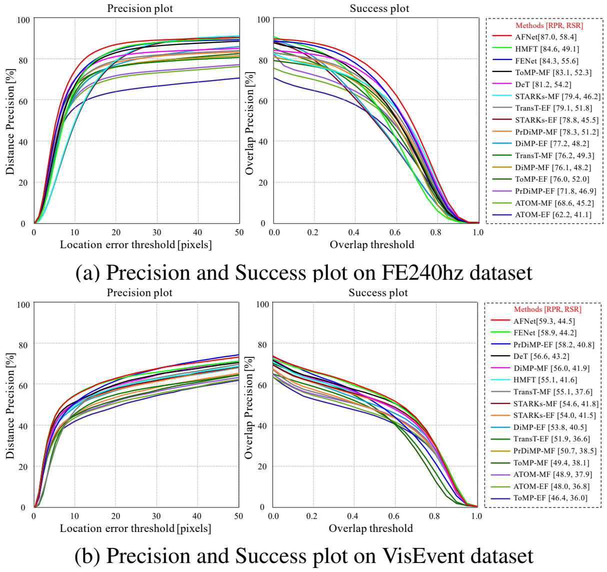
4. Experiments

4.1. Datasets

We evaluate our AFNet on two event-frame-based datasets: FE240hz [61] and VisEvent [51]. The FE240hz dataset has annotation frequencies as high as 240 Hz and consists of more than 143K images and corresponding recorded events. With this dataset, our method can accomplish a high frame rate tracking of 240Hz. Compared with FE240hz, VisEvent provides a low annotation frequency, about 25Hz. However, it contains various rigid and non-rigid targets both indoors and outdoors. Following [59], there are 205 sequences for training and 172 for testing.

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

To demonstrate the effectiveness of our method, we compare AFNet with the nine state-of-the-art trackers. Specifically, ATOM [9], DiMP [3], PrDiMP [10], STARKs [56], TransT [7] and ToMP [35] are conventional frame-based trackers. For a fair comparison, we extend them to multi-modality trackers via the following two fusion strategies: (i) Early Fusion (EF), we first add the aggregated events and corresponding frame as unified data, and then feed it into trackers; (ii) Middle Fusion (MF), we first use the backbone of these trackers to extract the frame and event features separately before feeding the sum of these features into the regressor. We also compared three original multi-modality methods: DeT [57], HMFT [62], and FENet [61] are frame-depth, frame-thermal, and frame-event trackers, respectively. All approaches are re-trained and tested on the FE240hz and VisEvent datasets. Following [61], we use RSR and RPR to evaluate all trackers. RSR and RPR focus on the overlap and center distance between the ground truth and the predicted bounding box, respectively.



(a) Precision and Success plot on FE240hz dataset

(b) Precision and Success plot on VisEvent dataset

Figure 3. Results on FE240hz [61] and VisEvent [51] datasets.

Figure 3 (a) shows the overall evaluation results on the FE240hz [61] dataset, which demonstrates the proposed AFNet offers state-of-the-art high frame rate tracking performance and outperforms other compared approaches in terms of both precision and success rate. In particular, our proposed AFNet achieves an 87.0% overall RPR and 58.4% RSR, outperforming the runner-up by 2.7% and 2.8%, respectively. We further validate the robustness of our AFNet under five common challenging scenarios: high dynamic range (HDR), low-light (LL), fast motion (FM), no motion (NM), and severe background motion (SBM). Among them, the first three conditions present challenges for tracking in the conventional frame modality, while the last two scenarios provide difficulties for the event modality. As shown in Table 1, we can see that AFNet surpasses other approaches in all conditions. These results validate the effectiveness of our proposed approach on high frame rate object track-

Methods	Fusion	HDR		LL		FM		NM		SBM		All	
	Type	RSR	RPR	RSR	RPR	RSR	RPR	RSR	RPR	RSR	RPR	RSR	RPR
ATOM [9]	EF	26.6	42.6	44.6	67.2	56.4	83.2	46.7	78.7	28.9	41.9	41.1	62.2
	MF	29.1	48.4	52.7	78.1	45.4	68.9	60.8	92.4	40.1	60.3	45.2	68.6
DiMP [3]	EF	38.5	61.0	58.1	86.8	53.7	90.4	50.1	86.5	47.8	74.8	48.2	77.2
	MF	39.1	61.4	55.4	83.6	59.4	93.0	42.6	76.6	50.4	78.7	48.2	76.1
PrDiMP [10]	EF	22.3	32.7	64.0	92.2	53.1	85.0	56.9	91.8	35.0	52.9	46.9	71.8
	MF	39.3	64.1	63.0	89.3	60.4	95.7	55.0	92.2	47.9	73.8	51.2	78.3
STARKs [56]	EF	42.2	73.1	55.0	90.5	41.6	75.1	26.4	53.0	51.9	84.5	45.5	78.8
	MF	44.1	75.7	54.8	90.0	40.7	73.1	25.5	50.5	53.2	85.2	46.2	79.4
TransT [7]	EF	47.4	74.2	58.8	84.7	64.4	95.3	43.9	70.5	54.7	84.0	51.8	79.1
	MF	49.5	74.7	49.1	73.7	57.4	87.1	28.6	49.3	54.7	83.7	49.3	76.2
ToMP [35]	EF	32.0	50.6	61.8	89.5	56.3	79.5	31.1	47.7	43.0	60.9	52.0	76.0
	MF	47.7	76.8	56.6	86.4	61.8	94.4	44.8	84.8	55.5	87.3	52.3	83.1
DeT [57]	-	52.5	78.8	57.3	86.7	65.9	96.0	58.2	95.4	56.4	82.5	54.2	81.2
HMFT [62]	-	40.2	67.7	51.4	86.7	52.6	87.7	46.9	82.5	54.9	90.3	49.1	84.6
FENet [61]	-	53.1	83.5	58.2	83.9	62.5	94.7	47.2	72.4	57.8	88.5	55.6	84.3
AFNet (Ours)	-	**55.5**	**84.9**	**64.7**	**93.8**	**66.3**	**96.4**	**62.0**	**98.8**	**60.1**	**90.3**	**58.4**	**87.0**

Table 1. Attribute-based RSR/RPR scores(%) on FE240hz [61] dataset against state-of-the-art trackers.

Methods	Fusion	Rigid		Non-Rigid		All	
	Type	RSR	RPR	RSR	RPR	RSR	RPR
ATOM [9]	EF	45.2	58.1	22.4	30.6	36.8	48.0
	MF	47.9	61.1	20.7	27.8	37.9	48.9
DiMP [3]	EF	49.3	63.6	25.4	36.8	40.5	53.8
	MF	50.1	65.5	27.8	39.5	41.9	56.0
PrDiMP [10]	EF	46.5	65.3	31.0	45.8	40.8	58.2
	MF	47.2	60.9	23.6	33.1	38.5	50.7
STARKs [56]	EF	50.0	63.7	26.7	37.2	41.5	54.0
	MF	50.1	64.0	27.4	38.3	41.8	54.6
TransT [7]	EF	43.1	59.6	25.4	38.5	36.6	51.9
	MF	43.9	63.6	26.7	40.3	37.6	55.1
ToMP [35]	EF	45.2	57.3	20.2	27.7	36.0	46.4
	MF	46.7	59.5	23.0	31.8	38.1	49.4
DeT [57]	-	48.9	62.8	33.3	45.5	43.2	56.6
HMFT [62]	-	50.0	64.0	27.2	39.7	41.6	55.1
FENet [61]	-	**51.0**	65.9	32.3	46.7	44.2	58.9
AFNet (Ours)	-	50.8	**66.1**	**33.4**	**47.6**	**44.5**	**59.3**

Table 2. State-of-the-art comparison of rigid and non-rigid targets on the VisEvent [51] dataset.

ing. The extended multi-modal methods [3, 7, 9, 10, 35, 56] lack a well-designed fusion module, preventing them from efficiently combining the complementary information of the two domains. While original multi-modality trackers DeT [57], HMFT [62] and FENet [61] do not address the misalignment between frame and event data at different measurement rates, causing ambiguity when locating targets. Figure 4 further qualitatively shows the effectiveness of our AFNet in different challenging conditions.

Even though the VisEvent dataset [51] has a low frame rate annotation, it provides various non-rigid targets that are absent from the FE240hz dataset. Thus, we also compare our AFNet against other state-of-the-art methods on VisEvent. As shown in Figure 3 (b), our AFNet obtains 44.5% and 59.3% in terms of RSR and RPR, respectively, surpassing all previous methods. Table 2 reports the evaluation of various trackers on rigid and non-rigid targets, showing that AFNet outperforms other competing trackers on these two attributes, except the RSR on rigid targets. These results validate that our proposed multi-modality approach still remains effective for low frame rate frame-event tracking.

4.3. Ablation Study

Impact of Input Modalities. To validate the effectiveness of fusing frame and event modalities, we design comparative experiments based only on a single modality: (i) tracking with low frame rate conventional frames, then linearly interpolating the results to 240Hz; (ii) tracking with aggregated events of 240Hz. As shown in the rows *A* and *B* of Table 3, when using only frame or event modality as input, the performance of trackers is 16.2%/26.9% and 43.6%/66.9% at PSR/PPR, respectively. These results are significantly worse than our AFNet, which demonstrates the necessity of multi-modality fusion for high frame rate tracking.

Influence of Event-guided Cross-modality Alignment (ECA). Our proposed ECA module has two key components: style transformer (ST) and deformable alignment (DA). We thus conduct the following experiments to validate the effectiveness of ECA: (i) without ECA; Inside ECA, (ii) without ST (ECA w/o ST); (iii) without DA (ECA w/o DA). We retrain these three modified models, and the corresponding results are shown in the rows *C*-*E* of Table 3, respectively. We can see that the proposed ECA module and its components all contribute to the tracking performance of AFNet. When the ST is removed, the

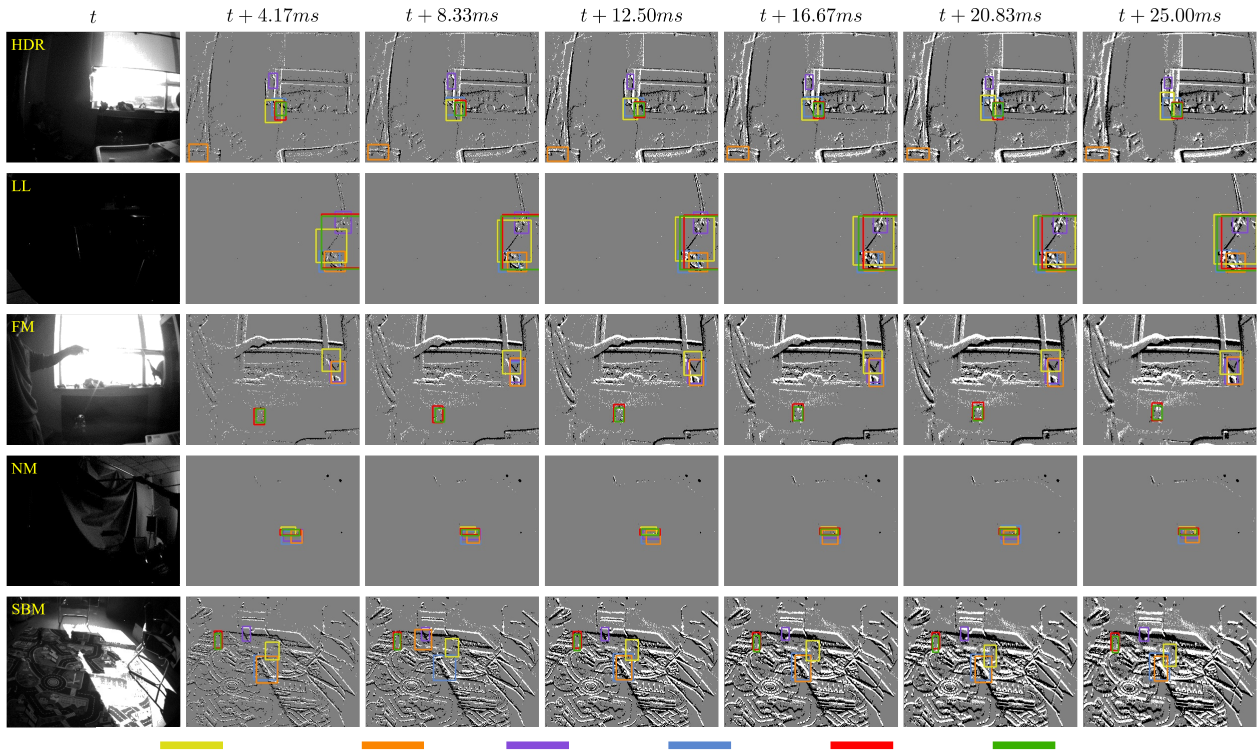


Figure 4. Qualitative comparison of AFNet against SOTA trackers on the FE240hz dataset [61] under five challenging conditions. All trackers locate the target at time $t + \Delta t$ with conventional frames at time t and aggregated events at time $t + \Delta t$ as inputs.

	Models	RSR↑	OP$_{0.50}$ ↑	OP$_{0.75}$ ↑	RPR↑
A.	Frame Only	16.2	15.8	3.4	26.9
B.	Event Only	43.6	53.4	18.6	66.9
C.	w/o ECA	55.1	69.3	29.1	82.4
D.	ECA w/o ST	55.8	69.8	31.2	83.0
E.	ECA w/o DA	55.5	70.0	30.9	82.8
F.	w/o CF	55.9	69.2	31.5	83.7
G.	CF w/o $\mathcal{K}$	56.2	69.5	31.6	84.3
H.	Ours	**58.4**	**73.5**	**32.6**	**87.0**

Table 3. Ablation study results.

PSR and PPR drop significantly by 2.6% and 4.0%, respectively. This illustrates that combining the frame modality's content with the event modality's style plays a key role in multi-modality alignment. The performance drops by 2.9%/4.2% at PSR/PPR when the DA is removed. This drop demonstrates that cross-frame-rate alignment between conventional frames and events indeed decreases target location ambiguity and enhances the discrimination ability of our tracker. To further verify cross-modality and cross-frame-rate alignment capabilities of ECA, we visualize the feature heatmaps of the frame modality prior to and following ECA. As shown in Figure 5, the first example shows a target that is moving upwards. We can see that the frame features shift the attention to the location of the aggregated events by our ECA. The second illustration shows that frame features suffered from the HDR scenario. With our ECA, target location ambiguity is eliminated. The aligned frame features will be fused with event features to improve the discriminative ability of our tracker further.

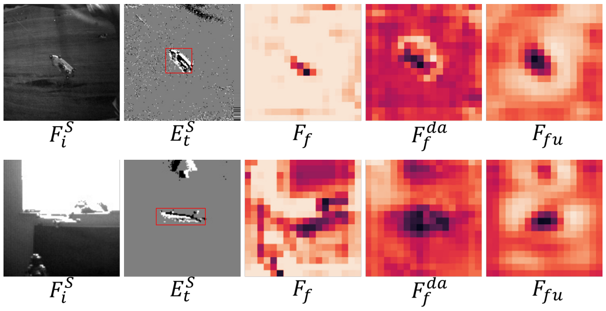


Figure 5. Visualization of features from the frame modality before (*i.e.*, F_f) and after (*i.e.*, F_f^{da}) alignment by our ECA. F_i^S and E_t^S are the frame modality input and event modality input of the search branch, respectively. F_{fu} denotes the final fused feature.

Influence of Cross-correlation Fusion (CF). We assess the influence of our CF module by replacing it with a concatenation operation in our AFNet. As shown in the row *F* of Table 3 , the performance drops on PSR and PPR by 2.5% and 3.3% illustrate that a well-designed multi-modality fusion strategy is essential. We further validate the impact of cross-correlation between two modalities by removing the dynamic filter. The results in the row *G* of Table 3 demonstrate that complementing one modality with information from another indeed enhances the feature representation.

Event Frame Rate (Hz)	40	80	120	160	200	240
DeT [57]	49.7	52.2	51.3	53.5	54.3	54.2
FENet [61]	52.4	54.4	55.6	52.8	54.7	55.6
AFNeT	**56.1**	**56.5**	**58.0**	**57.4**	**57.9**	**58.4**

Table 4. RSR of various event frame rates on the top-3 trackers.

Event Representation. We provide ablation on the way events are converted to frames from two perspectives: (i) The frame rate of accumulated event frames. We conduct experiments with different event frame rates on the FE240hz dataset. The results in Table 4 indicate that AFNet performs the best at all six event frame rates. (ii) The starting point of accumulation. We report the performance of accumulating events since the last event frame (a) and since the last intensity frame (b), see Table 5. The results of (a) on the top-3 methods are clearly lower than (b). This is because the accumulation method (a) leads to too sparse event frames, while (b) provides more motion cues for tracking.

High Frame Rate Tracking Based on Interpolation. One question in our mind is whether interpolation on results or conventional frames still yields satisfactory high frame rate tracking results. To answer this question, we conduct two interpolation strategies: (i) We first aggregate events at the frame rate of conventional frames. Then, these aggregated events and frames are utilized for training and testing trackers to predict low frame rate results, which are further linearly interpolated to generate high frame rate bounding boxes. (ii) We employ the video interpolation approach SuperSloMo [20] on conventional frames to predict high frame rate sequences for evaluation. Take note that the input of the event branch of all multi-modality trackers is replaced with interpolated frames. As shown in Figure 6, the results of interpolating on low frame rate results and on conventional frames are both noticeably inferior to using high frame rate aggregated events. These results demonstrate that designing multi-modality alignment and fusion networks to fully exploit the high temporal resolution of events for achieving high frame rate tracking is a feasible and significant manner.

	(a) since last event frame			**(b)** since last intensity frame		
	DeT [57]	FENet [61]	AFNet	DeT [57]	FENet [61]	AFNet
RSR	49.8	51.8	54.9	54.2	55.6	58.4
RPR	75.5	78.4	83.2	81.2	84.3	87.0

Table 5. Comparison of start times for event accumulation.

4.4. Discussion

Ideally, the tracking frame rate of our AFNet can reach the measurement rate of an event-based camera. Constrained by the existing annotated rates, we verify the effectiveness of our proposed AFNet on FE240hz at 240Hz and VisEvent at 25Hz. Our current focus is on exploiting multi-modality alignment and fusion schemes for effective and robust high frame rate tracking in various challenging conditions. However, we have not developed a lightweight network or a simple regression mechanism to speed up the evaluation of our approach. As shown in Table 6, we report the RPR and RSR with respect to the evaluation speed of the four multi-modality approaches on the FE240hz [61] dataset. We can see that, at nearly equal assessment speeds, our AFNet offers the best tracking accuracy.

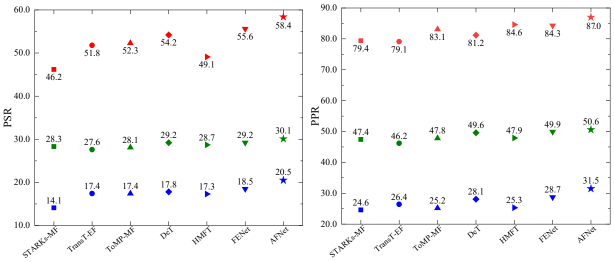


Figure 6. Comparison of whether to interpolate on the top-7 trackers. The blue denotes linearly interpolated performance on low frame rate tracking results; The green is tracking results on high frame rate conventional frames interpolated by SuperSloMo [20]; While red represents the results of utilizing aggregated events that have a higher frame rate than conventional frames.

Methods	DeT [57]	HMFT [62]	FENet [61]	AFNet
RSR	54.2	49.1	55.6	**58.4**
RPR	81.2	84.6	84.3	**87.0**
Speed (FPS)	**36.68**	34.83	35.5	36.21

Table 6. Comparison of accuracy and efficiency of multi-modality approaches on the FE240hz [61] dataset.

5. Conclusion

In this paper, we propose a multi-modality architecture for high frame rate single object tracking, which is comprised of two key components: event-guided cross-modality alignment (ECA) module and cross-correlation fusion (CF) module. The novel-designed ECA scheme is able to effectively establish cross-modality and cross-frame-rate alignment between conventional frames and aggregated events at the feature level. After alignment, the CF module focuses on fusing the advantages of both modalities by complementing one modality with information from another. Extensive experiments and ablation validation demonstrate the effectiveness and robustness of our AFNet in various challenging conditions. The proposed AFNet is the first in a line of work that jointly exploits frame and event modalities for high frame rate object tracking.

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 Grant 61972067/U21A20491/U1908214, the HiSilicon(Shanghai) Technologies Co.,Ltd (No. TC20210510004), and the Distinguished Young Scholars Funding of Dalian (No. 2022RJ01).

References

[1] Luca Bertinetto, Jack Valmadre, Stuart Golodetz, Ondrej Miksik, and Philip HS Torr. Staple: Complementary learners for real-time tracking. In *CVPR*, 2016. 2

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

[3] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *ICCV*, 2019. 3, 5, 6

[4] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *CVPR*, 2010. 2

[5] 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

[6] Haosheng Chen, Qiangqiang Wu, Yanjie Liang, Xinbo Gao, and Hanzi Wang. Asynchronous tracking-by-detection on adaptive time surfaces for event-based object tracking. In *ACMMM*, 2019. 2

[7] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, 2021. 1, 5, 6

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 2, 3, 4

[9] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *CVPR*, 2019. 1, 5, 6

[10] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *CVPR*, 2020. 1, 5, 6

[11] Jianing Deng, Li Wang, Shiliang Pu, and Cheng Zhuo. Spatio-temporal deformable convolution for compressed video quality enhancement. In *AAAI*, 2020. 2, 3

[12] Jianchuan Ding, Bo Dong, Felix Heide, Yufei Ding, Yunduo Zhou, Baocai Yin, and Xin Yang. Biologically inspired dynamic thresholds for spiking neural networks. In *NeurIPS*, 2022. 2

[13] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *CVPR*, 2019.

[14] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *TPAMI*, 2020. 1

[15] Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *ECCV*, 2022. 1

[16] Ankur Handa, Richard A Newcombe, Adrien Angeli, and Andrew J Davison. Real-time camera tracking: When is high frame-rate best? In *ECCV*, 2012. 1

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[18] Joao F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In *ECCV*, 2012. 2

[19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 3

[20] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *CVPR*, 2018. 8

[21] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *ICCV*, 2017. 1

[22] Hanme Kim, Stefan Leutenegger, and Andrew J Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *ECCV*, 2016. 1

[23] Adarsh Kowdle, Christoph Rhemann, Sean Fanello, Andrea Tagliasacchi, Jonathan Taylor, Philip Davidson, Mingsong Dou, Kaiwen Guo, Cem Keskin, Sameh Khamis, et al. The need 4 speed in real-time dense visual tracking. *TOG*, 2018. 1

[24] Matej Kristan et al. The visual object tracking vot2017 challenge results. In *ICCVW*, 2017. 1

[25] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E Shi, and Ryad B Benosman. Hots: a hierarchy of event-based time-surfaces for pattern recognition. *TPAMI*, 2016. 2

[26] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, 2018. 2

[27] Alejandro Linares-Barranco, Francisco Gómez-Rodríguez, Vicente Villanueva, Luca Longinotti, and Tobi Delbrück. A usb3. 0 fpga event-based filtering and tracking framework for dynamic vision sensors. In *ISCAS*, 2015. 2

[28] Martin Litzenberger, Christoph Posch, D Bauer, Ahmed Nabil Belbachir, P Schon, B Kohn, and H Garn. Embedded vision system for real-time object tracking using an asynchronous transient vision sensor. In *DSPW & SPEW*, 2006. 2

[29] Qianhui Liu, Dong Xing, Huajin Tang, De Ma, and Gang Pan. Event-based action recognition using motion information and spiking neural networks. In *IJCAI*, 2021. 2

[30] Yuanyuan Liu, Chengjiang Long, Zhaoxuan Zhang, Bokai Liu, Qiang Zhang, Baocai Yin, and Xin Yang. Explore contextual information for 3d scene graph generation. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 2

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2

[32] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv*, 2016. 5

[33] Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang. Hierarchical convolutional features for visual tracking. In *ICCV*, 2015. 2

[34] Juan Marín-Vega, Michael Sloth, Peter Schneider-Kamp, and Richard Röttger. Drhdr: A dual branch residual network for multi-bracket high dynamic range imaging. In *CVPR*, 2022. 2

[35] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *CVPR*, 2022. 1, 5, 6

[36] Haiyang Mei, Bo Dong, Wen Dong, Jiaxi Yang, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, and Xin Yang. Glass segmentation using intensity and spectral polarization cues. In *CVPR*, 2022. 2

[37] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *CVPR*, 2021. 2

[38] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *ECCV*, 2016. 1

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

[40] 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 *NeurIPS*, 2019. 5

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

[42] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13676–13685, 2020. 2

[43] Yu Qiao, Jincheng Zhu, Chengjiang Long, Zeyao Zhang, Yuxin Wang, Zhenjun Du, and Xin Yang. Cpral: Collaborative panoptic-regional active learning for semantic segmentation. In *AAAI*, 2022. 2

[44] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, 2018. 2

[45] Zhihao Shi, Xiaohong Liu, Kangdi Shi, Linhui Dai, and Jun Chen. Video frame interpolation via generalized deformable convolution. *TMM*, 2021. 2, 3

[46] Hubert Shum and Takaaki Komura. Tracking the translational and rotational movement of the ball using high-speed camera movies. In *ICIP*, 2005. 1

[47] Xin Tao, Hongyun Gao, Renjie Liao, Jue Wang, and Jiaya Jia. Detail-revealing deep video super-resolution. In *ICCV*, 2017. 2

[48] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, 2020. 2, 3

[49] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *CVPR*, 2021. 1

[50] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, 2019. 2, 4

[51] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. Visevent: Reliable object tracking via collaboration of frame and event flows. *arXiv*, 2021. 5, 6

[52] Yang Wang, Bo Dong, Ke Xu, Haiyin Piao, Yufei Ding, Baocai Yin, and Xin Yang. A geometrical approach to evaluate the adversarial robustness of deep neural networks. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023. 2

[53] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *TPAMI*, 2015. 1

[54] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *IJCV*, 2019. 2

[55] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. *arXiv*, 2022. 2

[56] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. *ICCV*, 2021. 1, 5, 6

[57] Song Yan, Jinyu Yang, Jani Kapyla, Feng Zheng, Ales Leonardis, and Joni-Kristian Kamarainen. Depthtrack: Unveiling the power of rgbd tracking. In *ICCV*, 2021. 5, 6, 8

[58] Haiwei Zhang, Jiqing Zhang, Bo Dong, Pieter Peers, Wenwei Wu, Xiaopeng Wei, Felix Heide, and Xin Yang. In the blink of an eye: Event-based emotion recognition. In *SIGGRAPH*, 2023. 2

[59] Jiqing Zhang, Bo Dong, Haiwei Zhang, Jianchuan Ding, Felix Heide, Baocai Yin, and Xin Yang. Spiking transformers for event-based single object tracking. In *CVPR*, 2022. 2, 5

[60] Jiqing Zhang, Chengjiang Long, Yuxin Wang, Haiyin Piao, Haiyang Mei, Xin Yang, and Baocai Yin. A two-stage attentive network for single image super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 2

[61] Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong. Object tracking by jointly exploiting frame and event domain. In *ICCV*, 2021. 2, 3, 5, 6, 7, 8

[62] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *CVPR*, 2022. 5, 6, 8

[63] He Zheng, Xin Li, Fanglong Liu, Lielin Jiang, Qi Zhang, Fu Li, Qingqing Dang, and Dongliang He. Adaptive spatial-temporal fusion of multi-objective networks for compressed video perceptual enhancement. In *CVPR*, 2021. 2

[64] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *CVPR*, 2022. 2

[65] Zikun Zhou, Jianqiu Chen, Wenjie Pei, Kaige Mao, Hongpeng Wang, and Zhenyu He. Global tracking via ensemble of local trackers. In *CVPR*, 2022. 2