This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

TAPFormer: Robust Arbitrary Point Tracking via Transient Asynchronous Fusion of Frames and Events

Jiaxiong Liu[1] Zhen Tan[1] Jinpu Zhang[1] Yi Zhou[2] Hui Shen[1]
Xieyuanli Chen[1] Dewen Hu[1*]
[1]National University of Defense Technology [2]Hunan University

Abstract

Tracking any point (TAP) is a fundamental yet challenging task in computer vision, requiring high precision and long-term motion reasoning. Recent attempts to combine RGB frames and event streams have shown promise, yet they typically rely on synchronous or non-adaptive fusion, leading to temporal misalignment and severe degradation when one modality fails. We introduce TAPFormer, a transformer-based framework that performs asynchronous temporal-consistent fusion of frames and events for robust and high-frequency arbitrary point tracking. Our key innovation is a Transient Asynchronous Fusion (TAF) mechanism, which explicitly models the temporal evolution between discrete frames through continuous event updates, bridging the gap between low-rate frames and high-rate events. In addition, a Cross-modal Locally Weighted Fusion (CLWF) module adaptively adjusts spatial attention according to modality reliability, yielding stable and discriminative features even under blur or low light. To evaluate our approach under realistic conditions, we construct a novel real-world frame-event TAP dataset under diverse illumination and motion conditions. Our method outperforms existing point trackers, achieving a 28.2% improvement in average pixel error within threshold. Moreover, on standard point tracking benchmarks, our tracker consistently achieves the best performance. Project website: tapformer.github.io

1. Introduction

Tracking any point (TAP), which estimates the motion trajectory of any point in a video, is a fundamental task in computer vision. A robust tracker capable of maintaining accurate trajectories under diverse conditions, from bright daylight to low-light nighttime scenes and from slow to fast motions, is important for applications such as augmented reality and autonomous systems [5, 38]. However, existing methods relying on conventional frame-based cameras often experience severe degradation under challenging conditions, due to motion blur, overexposure, or limited frame rates that fail to capture rapid dynamics [6, 16, 22, 42].

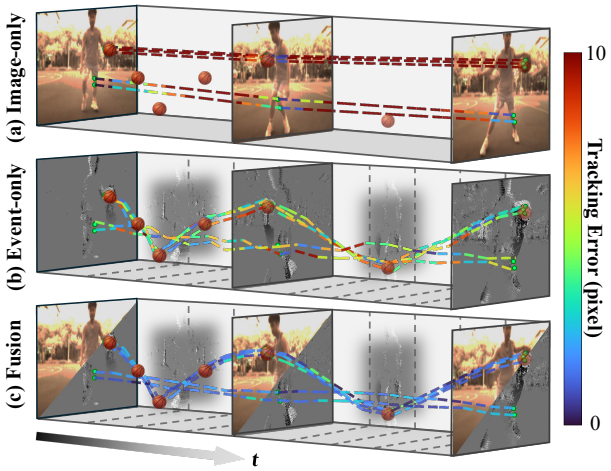


Figure 1. A qualitative comparison of tracking performance in our dataset shows that (a) frame-based tracking suffers from insufficient temporal information, and (b) event-based tracking fails to capture fine spatial details. In contrast, our (c) fusion approach can recover long-term, high-accuracy point trajectories. The rightmost color bar visualizes the per-point tracking error (in pixels).

Event cameras, in contrast, asynchronously record per-pixel brightness changes with microsecond temporal precision, offering wide dynamic range. These properties make them suitable for high-speed or high-contrast environments. However, event data are inherently coupled with motion, meaning the generated event stream depends on both the scene and the camera motion. Even the same scene can produce drastically different event patterns under varying motions, making it difficult to form consistent feature maps across motion conditions. Moreover, in slow-motion or static scenes, the event stream becomes sparse, and the lack of texture and color further limits tracking performance.

These complementary strengths and weaknesses naturally motivate fusion-based approaches [30, 41]. In biological vision, robust perception is achieved through the cooperation of parallel visual pathways. The ventral visual stream

*Corresponding author: dwhu@nudt.edu.cn

mainly processes static attributes such as color and texture, while the dorsal stream encodes motion and spatial relationships [45]. Analogously, RGB and event cameras provide complementary visual information. The former excels at capturing detailed spatial structure, while the latter encodes fine-grained temporal dynamics. Inspired by this analogy, we develop a unified fusion framework that integrates both modalities to achieve stable, high-precision arbitrary point tracking across diverse real-world conditions.

To achieve this goal, two major challenges should be addressed. The first challenge lies in the frequency mismatch between image frames and event data. Directly fusing asynchronous event streams with the most recent image frame can result in severe spatial misalignment due to the time lag between the two modalities. Existing fusion methods [28, 37] often circumvent this problem by downsampling event data to match frame rates, sacrificing temporal fidelity and limiting potential gains. Although asynchronous fusion techniques have been explored in related tasks [9, 47], they are not well suited for TAP, as it requires pixel-level precision, long-term consistency, and robust handling of fine-grained features. To address this, we introduce a Transient Asynchronous Fusion (TAF) mechanism. Instead of aligning event streams to discrete frames, TAF treats the scene as a temporally continuous latent representation. Each time a frame arrives, the fusion state is initialized by combining frame and recent events; between frames, this representation is continuously updated by incoming events through a lightweight cross-attention updater. This formulation explicitly models the temporal evolution of visual information, enabling high-frequency feature updates far beyond the frame rate.

The second challenge lies in constructing robust cross-modal feature representations under challenging conditions. In realistic scenes, the reliability of image and event modalities fluctuates spatially and temporally. Cross-modal Locally Weighted Fusion (CLWF) module introduces an adaptive local attention mechanism where event tokens query spatial neighborhoods of frame tokens, learning to emphasize the modality that is currently more informative. This results in robust, spatially coherent features even when one modality degrades due to motion blur or sparsity. Moreover, once frame and event inputs are tokenized and fused, both scale and temporal information can be naturally incorporated into our architecture. In this way, these modules establish a temporally consistent, modality-adaptive representation that bridges the temporal gap between asynchronous sensors while maintaining fine spatial precision.

Beyond algorithmic contributions, we also introduce two datasets to systematically study multimodal TAP. We built a new multimodal tracking benchmark that includes a synthetic frame-event training dataset and manually annotated real-world test sequences. Comprehensive experiments are conducted on both TAP and feature point tracking tasks. Across all benchmarks, our method consistently achieves the best performance among frame-based, event-based, and fusion-based trackers, demonstrating its strong generalization and robustness under diverse real-world conditions. The contribution of our work consist of:

- We propose a unified framework that fully exploits the complementary strengths of RGB frames and event streams, enabling robust arbitrary point tracking under diverse challenging real-world conditions.
- The first real-world TAP benchmark covering challenging conditions with synchronized frame–event data.
- State-of-the-art performance across multiple datasets in TAP and feature point tracking tasks, demonstrating the effectiveness and generality of our approach.

By shifting frame–event fusion from synchronous aggregation to asynchronous temporal modeling, TAPFormer offers a new perspective for building continuous, event-driven perception systems.

2. Related Work

2.1. Frame-Based Point Tracking Methods

TAP task has been revisited and reformulated as a particle video problem by querying trajectories across the entire image sequence [15], which handles occlusions by leveraging temporal context. TAP-Vid [7] further provides both synthetic training data and standardized evaluation benchmarks. PointOdyssey [48] also provides a large-scale synthetic dataset for long-term TAP. CoTracker [21] models the spatial correlations among target points by leveraging their physical coherence and jointly tracks them within a sliding temporal window. Inspired by DETR [4], TAPTR [25] employs a transformer-based architecture as an iterative optimizer and achieves strong performance on benchmark datasets. CoTracker3 [22] further simplifies the architecture by removing redundant components, achieving higher efficiency with minimal accuracy loss. Chrono [23] introduces temporal adapters between the transformer blocks of DINOv2 [33], demonstrating that temporal coherence plays a key role in feature extraction.

Another related line of research is optical flow estimation [20, 36], which focuses on dense correspondence estimation between consecutive frames. However, such methods often suffer from drift in long-term tracking. Despite recent advances in frame-based tracking, the inherent limitations of standard cameras, such as fixed frame rates and limited dynamic range, restrict their applicability under fast-motion or low-light conditions.

2.2. Event-Based Point Tracking Methods

Early event-based tracking approaches relied on geometric or optimization-based models[1, 2, 49]. For instance, Zhu et

al. [49] treat event streams as point clouds and estimate trajectories using ICP algorithms. HASTE [2] tracks features at the event level by hypothesizing multiple motion patterns and selecting the best match through template comparison. To enhance the performance and robustness of trackers, data-driven approaches have been proposed [3, 14, 19, 31]. However, these methods are limited to predefined feature points and cannot generalize to arbitrary point tracking.

More recently, ETAP [13] adapted trajectory optimization networks from the TAP task to the event domain and proposed an event-based TAP synthetic training dataset. MATE [14] introduces a motion-augmented approach that significantly enhances the temporal consistency and robustness of event-based point tracking.

Despite these advances, event-based tracking often fails in low-motion or static scenes due to sparse event generation. Moreover, the lack of color and texture detail limits the tracking precision of event-only methods, making them less reliable than frame-based approaches in most scenarios.

2.3. Multimodal Fusion for Events and Frames

To achieve more robust point tracking, several studies have explored the integration of frame and event data to leverage their complementary strengths: rich textures and stable appearance from images, and high temporal resolution and dynamic range from events [10, 27, 30, 39, 43]. For example, EKLT [10] uses grayscale images as templates and matches them with brightness increment images generated from event streams to perform feature point tracking. Other works adopt data-driven approaches, significantly improving performance under challenging conditions. The recent FETAP [27] introduces the first TAP model to fuse images and events, achieving substantial performance gains. However, since it is trained on optical-flow data, it lacks the ability to explicitly model point occlusions.

Fusion frames and events have also been investigated in other vision tasks such as object tracking [9, 24, 29, 47], line segment detection [46], and depth estimation [11]. However, most existing methods are either constrained by frame rates or suffer from degradation when one modality is unreliable [37]. Moreover, these approaches are not directly applicable to TAP, which demands pixel-level precision, long-term consistency, and robust per-point feature representations that remain stable under irregular motion.

To address these limitations, we propose a unified asynchronous fusion framework featuring two key modules. First, a transient feature update mechanism integrates image and event information to construct a temporally consistent representation when an image arrives and continuously updates it with incoming events between frames, enabling high-frequency tracking beyond standard frame rates. Second, a cross-modal locally weighted fusion module adaptively selects reliable regions from each modality and fuses them at a local scale, enhancing robustness under motion blur, occlusion, and illumination changes. Together, these designs yield a stable and temporally coherent representation that enables precise and robust arbitrary point tracking under diverse real-world conditions.

3. Method

3.1. Problem Definition

Our TAP framework fuses RGB and event data. Event cameras asynchronously record pixel-wise brightness changes, triggering an event whenever the logarithmic intensity variation exceeds a threshold $c \in \mathbb{R}^+$. Each event $e_k \doteq (\boldsymbol{x}_k, t_k, p_k)$ includes pixel location $\boldsymbol{x}_k = (x_k, y_k)$, a microsecond timestamp t_k, and polarity $p_k \in \{-1, +1\}$ for the sign of brightness change. The event-triggering condition can thus be formally expressed as:

$$\Delta L(\boldsymbol{x}_k, t_k) = L(\boldsymbol{x}_k, t_k) - L(\boldsymbol{x}_k, t_k - \Delta t_k) = p_k c, \quad (1)$$

where $L = \log(I)$ denotes the logarithmic intensity of image $I \in \mathbb{R}^{H \times W \times 3}$. Δt_k denotes the short interval between two consecutive events at the same pixel. Consequently, the pixel intensity at time t_k can be recovered from its previous intensity using the event polarity and threshold:

$$I(\boldsymbol{x}_k, t_k) = \exp(p_k c) \cdot I(\boldsymbol{x}_k, t_k - \Delta t_k). \quad (2)$$

Next, we define the TAP task that fuses image and event data. For simplicity, we represent the entire target process using a single query point here. Given a time interval $\tau = [\tau_s, \tau_e] \in \mathbb{R}$, $\tau_s \leq \tau_e$, we observe a sequence of events $\mathcal{E} = \{e_k \mid t_k \in \tau\}$ and a sequence of images $\mathcal{I} = \{I_t\}_{t=0}^{N^i}$, where $N^i \in \mathbb{N}^+$ is the number of image frames. We aim to track a query point $\mathbf{q} = (t_q, x, y) \in \mathbb{R}^3$, initialized at time $t_q \in \tau$, and estimate its trajectory $\boldsymbol{x}_{\tau_t} = (x_{\tau_t}, y_{\tau_t}) \in \mathbb{R}^2$ and visibility $v_{\tau_t} \in \{0, 1\}$ over discrete time steps $\{\tau_t\}_{t=0}^{N^e-1} \subset \tau$, where $N^e \in \mathbb{N}^+$ denotes the number of event frames, and 1 indicates visibility while 0 denotes occlusion.

For efficiency and noise robustness, the tracker operates at all discrete query times $\{\tau_t\}_{t=0}^{N^e-1}$, whose update frequency $f_e = \frac{N^e}{\tau_e - \tau_s}$ (typically 100–200 Hz) is much higher than the image frame rate $f_i = \frac{N^i}{\tau_e - \tau_s}$ (20–30 Hz). In principle, using our transient asynchronous fusion mechanism, the tracker can estimate the query point location at arbitrary query times τ_t. At each discrete step τ_t, the set of events in the corresponding time bin is collected:

$$E_t = \{e_k \mid \tau_{t-1} < t_k \leq \tau_t\} \subset \mathcal{E}, \quad (3)$$

and convert them to event representations $I_t^{ev} = \mathcal{F}(E_t) \in \mathbb{R}^{H \times W \times B}$. B is the channel number. $\mathcal{F}$ is a maximal timestamp version of Stacking Based on Time (SBT) [40].

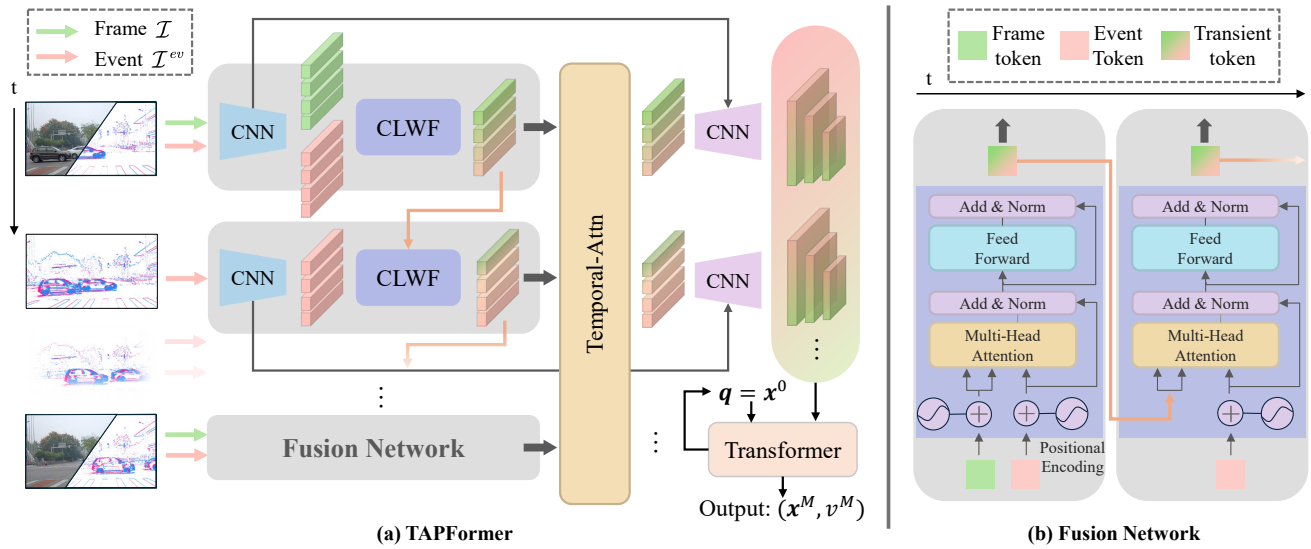


Figure 2. TAPFormer overview. (a) The overall framework: frames and events are fused by the transient asynchronous fusion mechanism and Cross-Modal Local Weighted Fusion modules to produce high-frequency transient features, refined by temporal attention and decoded into multi-scale fusion features. The resulting features, together with the initial query points position **q**, are fed into a transformer-based optimization module to iteratively predict tracking trajectories **x** and occlusion states v. M denotes the number of iterations. (b) The fusion network: image and event tokens are integrated by local weighted cross-attention to construct and update transient representations.

In summary, our goal is to use high-frequency event representations $\mathcal{I}^{ev}$, the images $\mathcal{I}$, and the initial query state $\mathbf{q}$, to estimate the pixel trajectory $\boldsymbol{x}_{\tau_t}$ and visibility state v_{τ_t} of the query point at each τ_t, as shown in Fig. 2.

3.2. Transient Asynchronous Fusion

A core challenge in frame-event fusion arises from their distinct temporal sampling characteristics: image frames are sparsely captured at low rates, whereas events provide fine-grained, asynchronous temporal updates. To bridge this temporal-frequency gap, we propose a Transient Asynchronous Fusion (TAF) mechanism that establishes and continuously updates a unified transient representation.

The Event-based Double Integral (EDI) model [26, 34] describes a blurred image $\tilde{\mathbf{B}}$ as the temporal integration of a sequence of latent sharp images $\tilde{\mathbf{L}}(t')$. It is shown as:

$$\tilde{\mathbf{L}}(t') = \tilde{\mathbf{B}} - \log\left(\frac{1}{T}\int_{t'-T/2}^{t'+T/2} \exp(c\,\mathbf{E}(t))\,dt\right), \quad (4)$$

which suggests that the observed image represents an integration of continuous latent intensity changes over time, modulated by the event data $\mathbf{E}(t)$. Building on this observation, we consider that a conventional frame and the corresponding events within its exposure interval jointly encode complementary temporal information, enabling the recovery of an instantaneous scene state. Following this idea, each incoming frame I_t is fused with events generated during its exposure period, defined as a short window $\mathcal{W}_t = (t-\delta, t]$. The exposure duration can be treated as a constant hyperparameter when the actual exposure time is unavailable. Fusion is performed through a cross-modal local weighted fusion (CLWF) module (detailed in Sec. 3.3):

$$\mathcal{R}_t = \mathcal{G}\big(\Phi_I(I_t),\, \Phi_E(\mathcal{F}(\mathcal{E} \cap \mathcal{W}_t))\big), \quad (5)$$

where Φ_I and Φ_E are modality-specific encoders. $\mathcal{G}$ denotes the CLWF operator that balances modality contributions in local spatial regions. $\mathcal{R}_t$ is transient representation.

After the initial fusion, $\mathcal{R}_t$ is asynchronously refined as new batches of events arrive. From Eq. (2), based on the event generation model [8], the image intensity at time t can, in theory, be reconstructed from the intensity in a previous frame and the accumulated events in between. However, this model is highly sensitive to sensor noise and hardware non-idealities. Instead of directly reconstructing pixel intensities, we apply an event-driven residual refinement mechanism through a cross-attention-based updater:

$$\mathcal{R}_{t+\Delta} \leftarrow \mathcal{U}\big(\mathcal{R}_{t+\Delta-1},\, \Phi_E(\mathcal{F}(E_{t+\Delta}))\big), \quad (6)$$

where $\mathcal{U}$ integrates fine-grained event cues while preserving spatial consistency from the recent frame. This design maintains temporal fidelity at event-rate updates, enabling smooth and accurate tracking in high-speed conditions.

3.3. Cross-modal Local Weighted Fusion Module

The Cross-Modal Local Weighted Fusion (CLWF) module acts as the key component in constructing the transient representation. It adaptively integrates image and event tokens, converted by a specific encoder Φ, within local regions in a modality-aware manner. The goal is to enhance local reliability by assigning higher attention weights to the more informative modality in each spatial neighborhood.

Given the image tokens $\Phi_I(I_t) \in \mathbb{R}^{N\times d}$ and event tokens $\Phi_E(\mathcal{F}(E \cap \mathcal{W}_t)) \in \mathbb{R}^{M\times d}$, CLWF performs a localized cross-attention operation. Each event token serves as a query that gathers information from image tokens within its spatial neighborhood $\mathcal{N}(j)$. The fusion weights are:

$$A_{j,i} = \frac{\exp\left(\langle q_j, k_i\rangle/\sqrt{d} + \mathcal{M}_{j,i}\right)}{\sum_{i'\in\mathcal{N}(j)} \exp\left(\langle q_j, k_{i'}\rangle/\sqrt{d} + \mathcal{M}_{j,i'}\right)}, \quad (7)$$

where $\mathcal{M}_{j,i}$ is an optional learned locality bias that encourages spatially adjacent tokens to have a stronger influence. Each event token is then updated through a residual cross-attention fusion: $\mathbf{f}_j^{tra} = \mathbf{f}_j^E + \sum_{i\in\mathcal{N}(j)} A_{j,i}\, v_i$, which preserves the event-driven temporal dynamics while injecting complementary image cues.

To enhance spatio-temporal coherence, the fused transient tokens ($\mathcal{R}_t = \{\mathbf{f}_j^{tra}\}$ are processed by a temporal self-attention block that aggregates information across adjacent time steps. Positional (spatial and temporal) embeddings are added before attention to preserve ordering information. During decoding, the resulting spatio-temporal token representation is progressively upsampled with skip connections to the corresponding encoder features, thereby producing multi-scale fused feature maps that integrate fine-grained temporal detail with global semantic context.

Finally, these multi-scale fused feature maps, together with the initial query point information, are fed into a transformer-based optimizer to iteratively estimate the query point trajectory and visibility.

4. Dataset Overview

Training Dataset Generation. To train our model, we created a high-frame-rate synthetic dataset using the Kubric [12] simulation engine, named *FE-FastKub*. Each scene was rendered at a resolution of 512×512 and a frame rate of 48 FPS over 2 s, producing temporally dense trajectories that enable high-fidelity supervision. Compared with the 12 FPS MoVi dataset, this configuration provides finer temporal granularity and allows realistic event simulation.

For each sequence, both clean sharp frames and motion-blurred frames were rendered. The former were used to generate events, and the latter served as network inputs. The final dataset contains 10,953 samples, each with 96 RGB images, 1024 ground-truth trajectories at 48 FPS, and the corresponding synthetic event stream generated by v2e [18]. During training, we sample 12 Hz image inputs while leveraging 48 Hz trajectory supervision, allowing the model to learn the transient feature change between consecutive frames. Furthermore, 2,878 high-speed samples are included to increase motion blur and improve the model's robustness to image degradation. This design helps the network rely less on high-quality RGB inputs and more on the temporal cues provided by events.

Real Dataset Construction. To evaluate real-world generalization, we constructed a novel frame–event TAP dataset that comprises two subsets: (i) InivTAP: sequences captured by a DAVIS346 camera (346×260, 20 FPS); and (ii) DrivTAP: sequences captured using our custom frame-event synchronization system, which integrates a Prophesee EVK4 (1280×720, 120 dB dynamic range) and an SG2-AR0231 RGB camera (1920×1080, 22 FPS).

The InivTAP subset contains 8 sequences of 5-7.5 s each, averaging 12.8 annotated points per sequence at 20 Hz. It covers a wide range of conditions, including fast motion, low light, overexposure, static scenes, dual-motion cases (both camera and object motion), and both indoor and outdoor environments. The DrivTAP subset includes 5 driving sequences (5-10 s each, average 12.2 points) captured at 10 Hz with ground-truth supervision at 20 Hz. Owing to the severe blur and overexposure, manually annotating a 5 s sequence typically takes about two hours, with more detailed in the supplementary material.

In total, our real-world dataset provides 20,450 annotated points (10,850+9,600 across two subsets) from 13 diverse sequences, covering a variety of conditions. To our best knowledge, it represents the first real-world multimodal TAP benchmark designed for evaluating frame-event fusion methods under realistic and challenging settings.

5. EXPERIMENTS

5.1. Implementation Details

Our model was trained on four NVIDIA RTX 4090 GPUs using images of 512×512 resolution and a batch size of 1. We employed the AdamW optimizer with an initial learning rate of 5×10^{-4}. Each sample contains 96 time steps and 1,024 ground-truth trajectories. During training, 150 trajectories were randomly selected from each sample, forming a sequence of 24 consecutive steps. To encourage temporal consistency learning, trajectories that persist for longer durations were assigned a higher sampling probability. All testing experiments were conducted on a laptop equipped with an NVIDIA RTX 3090 GPU.

5.2. Task 1: TAP

Evaluation Protocol. We evaluate our approach on our two challenging real-world datasets, *InivTAP* and *DrivTAP*. *InivTAP* covers complex representative challenging scenarios, while *DrivTAP* includes real driving sequences captured in both daytime and nighttime. To assess long-term tracking stability and occlusion awareness, we follow the TAP-Vid [7] benchmark and report three metrics: the average visible-point duration under a pixel error threshold ($\delta^{\text{vis}}_{\text{avg}}$), Occlusion Accuracy (OA), and Average Jaccard (AJ).

We compare our method with three groups of baselines: (i) sota frame-based trackers, including PIPs++ [48], Co-

Table 1. Comparison with state-of-the-art arbitrary point tracking methods. Training data: PO (Point Odyssey), Kub (Kubric), MF (MultiFlow), FE-FastKub (Our training dataset). speed expressed as ms per time step. Gray rows indicate our methods.

Method	Train Data	Input	Time↓	InivTAP			DrivTAP			Mean δ^{vis}_{avg} ↑
				AJ↑	δ^{vis}_{avg} ↑	OA↑	AJ↑	δ^{vis}_{avg} ↑	OA↑	
PIPs++ [48]	PO	Frame	21.1	–	31.2	–	–	37.8	–	34.5
Chrono [23]	Kub	Frame	-	31.2	44.7	79.8	15.8	30.8	68.1	37.8
CoTracker3 [22]	Kub	Frame	19.0	41.8	53.2	72.8	37.1	46.5	95.4	49.9
TAPFormer-F (Ours)	FastKub	Frame	16.5	44.6	54.5	74.7	36.6	46.7	93.8	50.6
E2Vid [35]-PIPs++ [48]	PO	Event	21.4	–	14.3	–	–	30.4	–	22.4
E2Vid [35]-Chrono [23]	Kub	Event	-	10.4	17.9	68.8	8.1	16.9	68.7	17.4
E2Vid [35]-CoTracker3 [22]	Kub	Event	22.2	12.8	19.4	56.6	18.6	27.0	89.4	23.2
ETAP [13]	E-Kub	Event	36.9	12.8	22.3	86.3	13.5	27.8	68.1	25.1
TAPFormer-E (Ours)	E-FastKub	Event	**14.0**	20.9	29.8	79.0	28.9	37.6	90.2	33.7
FETAP [27]	MF	Frame+Event	-	–	25.7	–	–	30.3	–	28.0
FETAP [27]	FE-FastKub	Frame+Event	23.4	42.2	54.9	83.9	36.8	46.4	95.2	50.7
TAPFormer (Ours)	FE-FastKub	Frame+Event	20.2	**57.0**	**69.9**	**95.2**	**48.8**	**60.1**	**97.8**	**65.0**

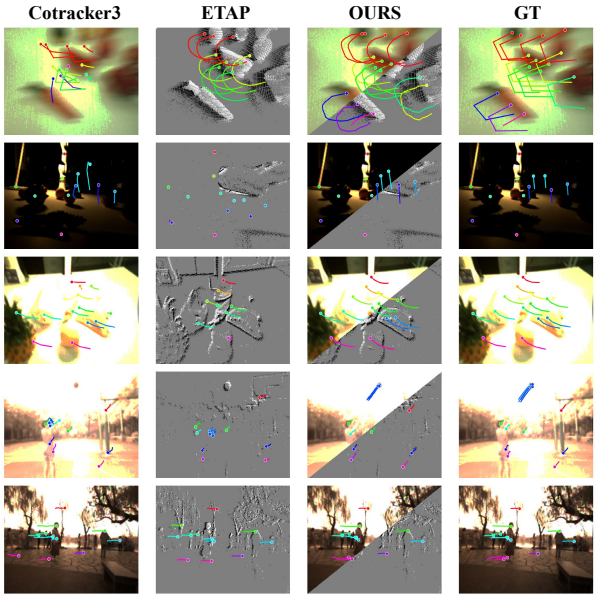


Figure 3. Task 1: TAP on *InivTAP*. Rows show different sequences, and columns correspond to frame-based, event-based, fusion-based (ours), and ground-truth results.

Tracker3 [22], and Chrono [23]; (ii) event-based trackers, including ETAP [13], as well as variants where events are reconstructed into videos via E2Vid [35] and then processed by frame-based methods; and (iii) frame-event fusion trackers, represented by FETAP [27]. Since FETAP does not explicitly model occlusions, we retrain it on our *FE-FastKub* dataset to ensure a fair comparison. All methods are carefully fine-tuned for each sequence. Additionally, we include single-modality variants of our network that process only frame or event inputs, without cross-modal fusion.

Results on InivTAP. As shown in Tab. 1, event-based methods excel in fast motion and high-dynamic-range conditions due to their high temporal resolution, while frame-based methods achieve higher accuracy in static, well-lit scenes. However, the former can lack detailed spatial structure, and the latter degrades severely under motion blur or overexposure. While naive RGB-event fusion mitigates these limitations, it fails to deal with their temporally-non-aligned nature. In contrast, our method adaptively integrates both cues, achieving consistent gains across all conditions. It attains accurate tracking under normal illumination by using detailed texture information from RGB frames, while leveraging events for rapid motion (see Fig. 3), quantitatively achieving a **36.4%** AJ improvement over the frame-based CoTracker3 and **35.1%** over the fusion-based FETAP. Our single-modality variants also outperform prior single-modality baselines, confirming the benefits of our high-frame-rate FastKubric training data and the proposed design of feature extraction.

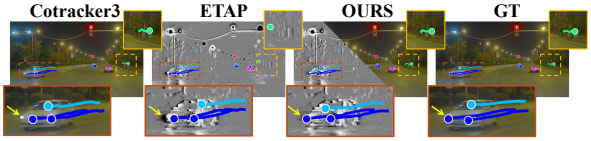


Figure 4. Red boxes show fast-moving vehicles where frame-based tracking drifts; yellow boxes highlight texture-similar regions where event-based tracking fails. Our fusion-based method achieves stable and accurate tracking in both cases.

Results on DrivTAP. To emphasize the temporal resolution of event cameras, the ground-truth trajectories in *DrivTAP* are annotated at twice the frame rate of the RGB images. This dataset contains dynamic real-world driving scenarios with rapidly changing illumination, fast oncoming traffic, and vehicles moving at similar speeds to the ego-car, resulting in low relative motion and frequent motion blur, which making the TAP task extremely challenging.

As shown in Tab. 1, the performance gap between frame-based and event-based methods becomes smaller in this dataset, since the frame rate of frame-based trackers is inherently limited by the RGB camera, while event-

Table 2. Comparison of tracking accuracy on EDS and EC datasets. Values are in percentages (%). The best results are bold.

Method	Input	EDS		EC	
		FA↑	EFA↑	FA↑	EFA↑
PIPs++[48]	Frame	75.1	63.0	82.6	82.3
CoTracker3[22]	Frame	80.2	68.8	92.5	91.9
Chrono[23]	Frame	–	–	82.9	82.3
TAPFormer-F (Ours)	Frame	80.9	69.1	92.7	92.1
EM-ICP [49]	Event	16.1	12.0	33.7	33.4
HASTE [2]	Event	9.6	6.3	44.2	42.7
AEB-Tracker [44]	Event	47.3	46.6	55.3	47.3
ETAP [13]	Event	74.5	63.9	88.1	87.6
MATE [14]	Event	–	–	88.5	87.5
TAPFormer-E (Ours)	Event	76.8	64.5	86.6	86.1
EKLT [10]	F + E	32.5	20.5	81.1	77.5
DeepEvT [30]	F + E	61.3	50.5	82.5	81.8
FETAP [27]	F + E	72.2	63.2	84.4	83.8
TAPFormer (Ours)	F + E	**82.3**	**70.4**	**93.3**	**92.6**

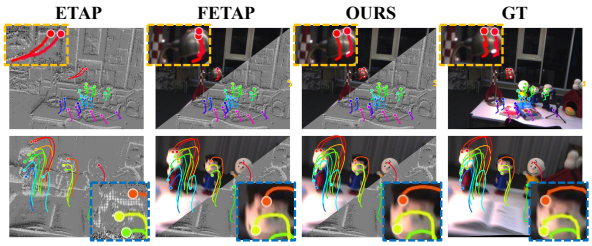


Figure 5. Task 2: Feature tracking on EDS.

based methods benefit from their high temporal resolution. Our proposed TAPFormer achieves the best overall performance, consistently surpassing both single-modality and previous fusion-based methods, see Fig. 4. Quantitatively, TAPFormer improves the AJ metric by **261.5%** over the event-based ETAP, **31.5%** over the frame-based CoTracker3, and **32.6%** over the fusion-based FETAP. Moreover, its inference speed remains comparable to mainstream trackers and even exceeds that of our single-modality baselines, demonstrating both efficiency and robustness in real-world driving scenarios.

5.3. Task 2: Feature Tracking

We further evaluate our method on the *EC* [32] and *EDS* [17] datasets for feature point tracking. Both datasets provide synchronized RGB and event streams at resolutions of 240×180 and 640×480, respectively. Following standard practice [30], we report Feature Age (FA) and Expected Feature Age (EFA), which measure the temporal consistency of tracked points under a fixed tracking error threshold. We observed that the *Peanuts Light* sequence in the EDS dataset contains several outlier trajectories that noticeably distort evaluation metrics. So, we apply a simple temporal smoothing to the ground-truth trajectories of this sequence (see supplementary material).

We compare our approach with state-of-the-art frame-only, event-only, and frame-event fusion trackers (Tab. 2). Frame-based TAP methods (PIPs++ [48], CoTracker3 [22], and Chrono [23]) perform well in static scenes but degrade under motion blur and low resolution. Our frame-only baseline achieves higher accuracy on both datasets, demonstrating the effectiveness of our feature extraction design. For event-based tracking, we include EM-ICP [49], HASTE [2], AEB-Tracker [44], ETAP [13], and MATE [14]. These methods benefit from the high temporal resolution of event data but lack semantic and spatial cues, which limit their robustness under complex textures. Fusion-based methods such as EKLT [10], DeepEvT [30], and FETAP [27] combine both modalities but do not fully exploit their complementary strengths, leading to suboptimal temporal stability.

Across both datasets, TAPFormer achieves the best overall performance. It consistently improves EFA compared with all baselines, maintaining smooth and stable trajectories even under fast motion and partial occlusions, as shown in Fig. 5. These results confirm that our method effectively integrates the complementary advantages of frames and events, achieving temporally consistent and fine-grained tracking in complex real-world scenes.

Table 3. Ablation study on the EDS dataset. FA and EFA denote feature age and expected feature age, respectively. MSSF and TAM denote multi-scale semantic features and temporal attention module, respectively.

FE-FastKub	CLWF	TAF	MSSF	TAM	FA↑	EFA↑
					0.646	0.535
✓					0.701	0.585
✓	✓				0.763	0.647
✓	✓	✓			0.803	0.685
✓	✓	✓	✓		0.814	0.698
✓	✓	✓		✓	0.807	0.691
✓	✓	✓	✓	✓	**0.823**	**0.704**

5.4. Ablation Study

Component Analysis. We conduct a detailed ablation study on the *EDS* dataset to assess the contribution of each proposed component, including the training dataset *FE-FastKub* dataset, TAF mechanism, CLWF module, the multi-scale semantic features (MSSF), and the temporal attention module (TAM). As summarized in Tab. 3, the baseline model was trained on the MultiFlow dataset without occlusion annotations. When the fusion module is removed, frame-event fusion is replaced by a channel-wise concatenation followed by a convolutional layer.

The results show that both the dataset and fusion strategy are crucial for high-quality tracking. Introducing the proposed fusion module yields a notable gain in FA and EFA, validating the effectiveness of transient asynchronous feature aggregation. Incorporating multi-scale semantic representations enhances spatial discriminability, while the temporal attention module further improves feature consistency across time. The complete model achieves the highest FA

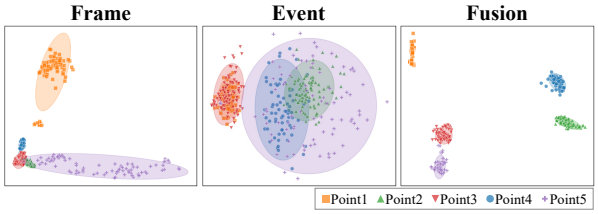


Figure 6. PCA visualization of feature embeddings from frame-, event-, and fusion-based representations. Each color denotes a tracked point; Ellipses indicate the covariance-based dispersion of feature embeddings in the PCA space, reflecting the intra-point stability and inter-point separability of learned representations.

(0.823) and EFA (0.704), demonstrating that each component contributes progressively to the overall performance.

Single-Modality Evaluation. We perform single-modality evaluations in Tab. 1 and Tab. 2. Our frame-only and event-only variants consistently outperform previous single-modality trackers, confirming the benefits of the proposed training data and feature extraction backbone. Nevertheless, each modality exhibits inherent weaknesses: the frame-only tracker degrades under fast motion or overexposure, whereas the event-only tracker lacks spatial texture cues. By jointly leveraging both modalities, our fusion-based design effectively mitigates these limitations and maintains robust tracking across diverse scenarios.

To further examine the robustness of the learned features and to validate the effectiveness of our cross-modal local weighted fusion module, we compare the temporal dynamics of point feature vectors obtained from the frame-only, event-only, and fused models. For each method, the same tracking points are sampled along ground-truth trajectories, and their feature vectors are projected onto a 2D space via principal component analysis (PCA). As shown in Fig. 6, temporally consistent and robust features are expected to form compact clusters for the same point while maintaining clear separations across different points. The fused representations exhibit both tight intra-point clustering and large inter-point margins over time, indicating superior temporal coherence and discriminative embedding quality than single-modality baselines.

Frame Rate Sensitivity. We further analyzed how the input frame rate affects tracking performance FA for our method and CoTracker3 under slow, normal, and fast motion settings. see Tab. 4 As the frame rate decreases, CoTracker3 suffers severe degradation and becomes nearly invalid at 10 Hz. In contrast, our fusion-based tracker remains stable, exhibiting only a 6.5% reduction in performance compared to a 75.3% drop for the frame-only method, as shown in Fig. 7 This experiment highlights the adaptability of our transient asynchronous fusion mechanism to low-frame-rate and high-speed conditions, demonstrating its potential for real-world deployment where frame capture rates are constrained.

Table 4. Effect of frame rate on tracking performance. Results are reported on the EDS dataset using different input frame rates.

Method	Frame rate (FPS)					
	75	37.5	25	18.75	12.5	9.375
CoTracker3	83.9	74.9	51.2	32.4	20.7	13.2
Ours	**85.1**	**83.2**	**82.7**	**80.8**	**79.6**	**75.8**

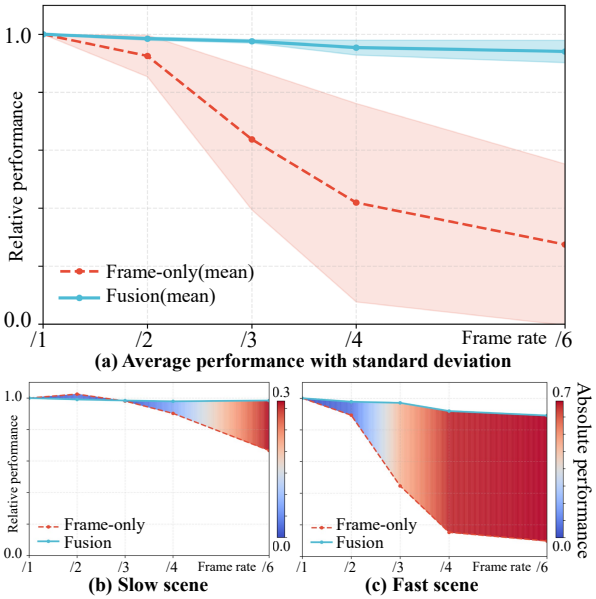


Figure 7. Frame-rate sensitivity comparison between CoTracker3 and our method. (a) shows overall performance trends with variance shading; (b–c) visualize relative performance in slow and fast motion scenarios, where color bands denote absolute performance differences (see color bar on the right).

6. Conclusion

We introduced TAPFormer, a synergistic frame-event tracker for robust and high-frequency point tracking under diverse visual conditions. Our framework builds upon two core innovations: a transient feature update mechanism that preserves the fine temporal dynamics of event streams, and a cross-modal local weighted fusion module that adaptively integrates complementary cues from frames and events to produce stable and discriminative representations. Extensive evaluations show that TAPFormer consistently achieves state-of-the-art performance on TAP and feature tracking benchmarks, outperforming all existing frame-, event-, and fusion-based baselines.

Our work also contributes two key resources to advance frame-event fusion research: a large-scale synthetic dataset for high-frame-rate supervision, and the first real-world, manually annotated TAP dataset with synchronized frame-event sequences under challenging illumination and motion. We hope these resources and the proposed framework will provide a foundation for advancing robust visual tracking in complex environments.

7. Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Grant Nos. U25B2069, 62506377, and 62403478).

References

[1] Ignacio Alzugaray and Margarita Chli. Asynchronous corner detection and tracking for event cameras in real time. *IEEE Robotics and Automation Letters*, 3(4):3177–3184, 2018. 2

[2] I. Alzugaray and M. Chli. HASTE: multi-hypothesis asynchronous speeded-up tracking of events. In *31st British Machine Vision Conference (BMVC)*, page 744, 2020. 2, 3, 7

[3] Yannick Burkhardt, Simon Schaefer, and Stefan Leutenegger. Superevent: Cross-modal learning of event-based keypoint detection. *arXiv preprint arXiv:2504.00139*, 2025. 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision. (ECCV)*, pages 213–229, 2020. 2

[5] Weirong Chen, Le Chen, Rui Wang, and Marc Pollefeys. Leap-vo: Long-term effective any point tracking for visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (CVPR)*, pages 19844–19853, 2024. 1

[6] Seokju Cho, Jiahui Huang, Jisu Nam, Honggyu An, Seungryong Kim, and Joon-Young Lee. Local all-pair correspondence for point tracking. In *European conference on computer vision. (ECCV)*, pages 306–325, 2024. 1

[7] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022. 2, 5

[8] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020. 4

[9] Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014):1034–1040, 2024. 2, 3

[10] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. EKLT: asynchronous photometric feature tracking using events and frames. *Int. J. Comput. Vis.*, 128 (3):601–618, 2020. 3, 7

[11] Daniel Gehrig, Michelle Rüegg, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. *IEEE Robotics and Automation Letters*, 6(2):2822–2829, 2021. 3

[12] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (CVPR)*, pages 3749–3761, 2022. 5

[13] Friedhelm Hamann, Daniel Gehrig, Filbert Febryanto, Kostas Daniilidis, and Guillermo Gallego. Etap: Event-based tracking of any point. In *Proceedings of the Computer Vision and Pattern Recognition Conference. (CVPR)*, pages 27186–27196, 2025. 3, 6, 7

[14] Han Han, Wei Zhai, Yang Cao, Bin Li, and Zheng-jun Zha. Mate: Motion-augmented temporal consistency for event-based point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision. (ICCV)*, pages 8340–8349, 2025. 3, 7

[15] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision. (ECCV)*, pages 59–75, 2022. 2

[16] Adam W Harley, Yang You, Xinglong Sun, Yang Zheng, Nikhil Raghuraman, Yunqi Gu, Sheldon Liang, Wen-Hsuan Chu, Achal Dave, Suya You, et al. Alltracker: Efficient dense point tracking at high resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision. (ICCV)*, pages 5253–5262, 2025. 1

[17] Javier Hidalgo-Carrió, Guillermo Gallego, and Davide Scaramuzza. Event-aided direct sparse odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.(CVPR)*, pages 5781–5790, 2022. 7

[18] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (CVPR)*, pages 1312–1321, 2021. 5

[19] Ze Huang, Li Sun, Cheng Zhao, Song Li, and Songzhi Su. Eventpoint: Self-supervised interest point detection and description for event-based camera. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5396–5405, 2023. 3

[20] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition. (CVPR)*, pages 2462–2470, 2017. 2

[21] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European conference on computer vision. (ECCV)*, pages 18–35, 2024. 2

[22] Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision. (ICCV)*, pages 6013–6022, 2025. 1, 2, 6, 7

[23] Inès Hyeonsu Kim, Seokju Cho, Jiahui Huang, Jung Yi, Joon-Young Lee, and Seungryong Kim. Exploring temporally-aware features for point tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference. (CVPR)*, pages 1962–1972, 2025. 2, 6, 7

[24] Dianze Li, Yonghong Tian, and Jianing Li. Sodformer: Streaming object detection with transformer using events and frames. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):14020–14037, 2023. 3

[25] Hongyang Li, Hao Zhang, Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, and Lei Zhang. Taptr: Tracking any point with transformers as detection. In *European Conference on Computer Vision. (ECCV)*, pages 57–75, 2024. 2

[26] Shijie Lin, Yingqiang Zhang, Dongyue Huang, Bin Zhou, Xiaowei Luo, and Jia Pan. Fast event-based double integral for real-time robotics. In *2023 IEEE International Conference on Robotics and Automation. (ICRA)*, pages 796–803, 2023. 4

[27] Jiaxiong Liu, Bo Wang, Zhen Tan, Jinpu Zhang, Hui Shen, and Dewen Hu. Tracking any point with frame-event fusion network at high frame rate. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 18834–18840. IEEE, 2025. 3, 6, 7

[28] Mengyun Liu, Na Qi, Yunhui Shi, and Baocai Yin. An attention fusion network for event-based vehicle object detection. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3363–3367, 2021. 2

[29] Zhanwen Liu, Nan Yang, Yang Wang, Yuke Li, Xiangmo Zhao, and Fei-Yue Wang. Enhancing traffic object detection in variable illumination with rgb-event fusion. *IEEE Transactions on Intelligent Transportation Systems*, 2024. 3

[30] Nico Messikommer, Carter Fang, Mathias Gehrig, and Davide Scaramuzza. Data-driven feature tracking for event cameras. In *IEEE Conference on Computer Vision and Pattern Recognition. (CVPR)*, pages 5642–5651, Vancouver, BC, Canada, 2023. 1, 3, 7

[31] Nico Messikommer, Carter Fang, Mathias Gehrig, Giovanni Cioffi, and Davide Scaramuzza. Data-driven feature tracking for event cameras with and without frames. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3

[32] Elias Mueggler, Henri Rebecq, Guillermo Gallego, Tobi Delbruck, and Davide Scaramuzza. The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam. *The International journal of robotics research*, 36(2):142–149, 2017. 7

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[34] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (CVPR)*, pages 6820–6829, 2019. 4

[35] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (CVPR)*, pages 3857–3866, 2019. 6

[36] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision. (ECCV)*, pages 402–419, 2020. 2

[37] Abhishek Tomy, Anshul Paigwar, Khushdeep S Mann, Alessandro Renzaglia, and Christian Laugier. Fusing event-based and rgb camera for robust object detection in adverse conditions. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 933–939, 2022. 2, 3

[38] Mel Vecerik, Carl Doersch, Yi Yang, Todor Davchev, Yusuf Aytar, Guangyao Zhou, Raia Hadsell, Lourdes Agapito, and Jon Scholz. Robotap: Tracking arbitrary points for few-shot visual imitation. In *2024 IEEE International Conference on Robotics and Automation.(ICRA)*, pages 5397–5403, 2024. 1

[39] Zhexiong Wan, Jianqin Luo, Yuchao Dai, and Gim Hee Lee. Event-aided dense and continuous point tracking: Everywhere and anytime. In *Proceedings of the IEEE/CVF International Conference on Computer Vision. (ICCV)*, pages 7936–7946, 2025. 3

[40] Lin Wang, Yo-Sung Ho, Kuk-Jin Yoon, et al. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (CVPR)*, pages 10081–10090, 2019. 3

[41] Lin Wang, Tae-Kyun Kim, and Kuk-Jin Yoon. Joint framework for single image reconstruction and super-resolution with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7657–7673, 2021. 1

[42] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *Proceedings of the IEEE/CVF International Conference on Computer Vision. (ICCV)*, pages 19795–19806, 2023. 1

[43] Xiangyuan Wang, Huai Yu, Lei Yu, Wen Yang, and Gui-Song Xia. Toward robust keypoint detection and tracking: A fusion approach with event-aligned image features. *IEEE Robotics Autom. Lett.*, 9(9):8059–8066, 2024. 3

[44] Ziwei Wang, Timothy Molloy, Pieter Van Goor, and Robert Mahony. Asynchronous blob tracker for event cameras. *IEEE Transactions on Robotics*, 2024. 7

[45] Zheyu Yang, Taoyi Wang, Yihan Lin, Yuguo Chen, Hui Zeng, Jing Pei, Jiazheng Wang, Xue Liu, Yichun Zhou, Jianqiang Zhang, et al. A vision chip with complementary pathways for open-world sensing. *Nature*, 629(8014):1027–1033, 2024. 2

[46] Huai Yu, Hao Li, Wen Yang, Lei Yu, and Gui-Song Xia. Detecting line segments in motion-blurred images with events. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2866–2881, 2023. 3

[47] Jiqing Zhang, Yuanchen Wang, Wenxi Liu, Meng Li, Jinpeng Bai, Baocai Yin, and Xin Yang. Frame-event alignment and fusion network for high frame rate tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (CVPR)*, pages 9781–9790, 2023. 2, 3

[48] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision. (ICCV)*, pages 19855–19865, 2023. 2, 5, 6, 7

[49] Alex Zihao Zhu, Nikolay Atanasov, and Kostas Daniilidis. Event-based feature tracking with probabilistic data association. In *IEEE International Conference on Robotics and Automation. (ICRA)*, pages 4465–4470, 2017. 2, 3, 7