

Event-aided Dense and Continuous Point Tracking: Everywhere and Anytime

Zhexiong Wan^{1,2} Jianqin Luo¹ Yuchao Dai¹ Gim Hee Lee²

¹School of Electronics and Information, Northwestern Polytechnical University &
 Shaanxi Key Laboratory of Information Acquisition and Processing

²Department of Computer Science, National University of Singapore

{wanzhexiong, luojianqin}@mail.nwpu.edu.cn, daiyuchao@nwpu.edu.cn, gimhee.lee@nus.edu.sg

Abstract

Recent point tracking methods have made great strides in recovering the trajectories of any point (especially key points) in long video sequences associated with large motions. However, the spatial and temporal granularities of point trajectories remain constrained by limited motion estimation accuracy and video frame rate. Leveraging the high temporal resolution and motion sensitivity of event cameras, we introduce event data for the first time to recover spatially dense and temporally continuous trajectories of every point at any time. Specifically, we define the dense and continuous point trajectory representation as estimating multiple control points of curves for each pixel and model the movement of sparse events triggered along continuous point trajectories. Building on this, we propose a novel multi-frame iterative streaming framework that first estimates local inter-frame motion representations from two consecutive frames with inter-frame events, then aggregates them into a global long-term motion representation to utilize input full video and event data with an arbitrary number of frames. Extensive experiments on simulated and real data demonstrate the significant improvement of our framework over state-of-the-art methods and the crucial role of introducing events to model continuous point trajectories.

1. Introduction

Estimating fine-grained motion from input videos is a crucial task in computer vision with widespread applications such as video compression [1], frame interpolation [23, 55], motion segmentation [2, 34], and dynamic scene reconstruction [16]. Early studies [21, 44] based on two-frame optical flow can model the spatially dense motion, but they suffer from the challenge of modeling long-term motion from more frames. With the proposal of tracking any point (TAP) task [5, 19], using sparse points as query indexes to

estimate pointwise long-term motions also draws attention. Despite significant progress, the sparse and independent representation of pointwise trajectories remains inherently incompatible with the spatially dense representation of input video. Thus, recently there are methods [4, 35, 46] that continue to use dense optical flow to represent long-term motion. However, the temporal frame rate of the input video is constrained by traditional shutter cameras, which makes accurately modeling complex motion trajectories difficult for video-only methods from the data acquisition perspective. Consequently, accurate estimation of fine-grained spatially dense and temporally continuous motion remains a challenging and worthwhile research problem.

Unlike traditional shutter cameras that expose the entire image at fixed frame rates, the new bio-inspired event cameras [9] can independently and asynchronously detect per-pixel brightness changes with microsecond precision. This unique design makes it inherently sensitive to visual motion changes, leading to impressive motion-related applications such as optical flow estimation [17, 62], motion segmentation [20, 61], feature tracking [33], object tracking [63], and frame interpolation [45]. However, events are typically triggered only with motion contours or rich textures, making it challenging to comprehensively perceive spatially dense motion. Recently, integrating the advantages of event camera and traditional shutter camera has become a new direction [38, 57]. Leveraging on these successes, we propose for the first time taking event data as an auxiliary input for the point tracking task, where we comprehensively model fine-grained spatially dense and temporally continuous motion from the input image and event modalities.

The introduction of event cameras offers the potential to model continuous motion from the data perspective. However, to effectively model continuous long-term motion, we need a new representation that replaces optical flow and parametrically associates the dense temporal properties of event data. BFlow [14] proposes to learn trajectories with Bézier curves from events, but is limited to input fixed frames and cannot adapt to longer sequences. CPFlow [31]

Project page: <https://npucvr.github.io/EDCPT>

proposes to learn control points represented as B-spline curves from a fixed number of image slices. Although continuous motion can be successfully modeled using these curve representations in normalized timescales, their fixed number of control points made it hard to handle complex dynamics and varying lengths of video sequences. Based on the curve representation, we propose a new streaming pipeline for accumulating multiple local curves to address these limitations. In addition, how to adequately fuse multi-frame temporal information to achieve long-term tracking is also a concern. Compared to not explicitly associating multi-frames in DOT [35] and aggregating over estimated motion vectors in AccFlow [52] and FlowTrack [4], we perform recurrent fusion in the feature space, dedicated to preserving multi-frame sequence information.

In this paper, we present the first event-aided point tracking framework for recovering spatially dense and temporally continuous point trajectories from input videos and event sequences. Specifically, we first propose a new point trajectory representation with parametric curves that accumulate multiple local curves to adapt to multi-frame input videos at any length. We then design a new framework for combining two frames with events to estimate dense point curve trajectories and extend to multi-frame iterative accumulation. In addition, we establish the association between continuous curve trajectory and event triggering as a part of the learning objective for continuous motion modeling. Extensive experiments on both simulated and real-world data demonstrate that the proposed framework significantly outperforms state-of-the-art methods. Particularly, our ablation studies illustrate the effectiveness of the proposed global aggregation and highlight the crucial role of incorporating event data in continuous trajectory modeling.

Our main contributions are summarized as follows:

- We introduce a new setup that, for the first time, enables long-term spatially dense and temporally continuous point tracking by integrating the strengths of both images and events.
- We present a novel global curve representation of continuous point trajectories through multi-frame aggregation, establishing a connection between event triggering and continuous motion.
- We propose a novel event-aided streaming framework that iteratively accumulates the local tracks from two frames with inter-frame events, resulting in global, long-term dense and continuous trajectories through iterative temporal global motion aggregation.

2. Related Works

2.1. Image-based query point tracking

The goal of point tracking is to recover the corresponding positions of query points in each frame, which has attracted

wide attention with the proposal of the TAP benchmark and the baseline model TAPNet [5]. PIPs [19] proposes to extract independent point representation for 8-frame tracking, then PIPs++ [58] extends to long-term trajectories. TAPIR [6] proposes a two-stage matching framework that fuses TAPNet and PIPs, and BootsTAPIR [7] performs self-supervised training on more data. Unlike these methods that track only one query at a time, CoTracker [24] and Context-PIPs [51] use additional tracks and pixel features to improve global tracking. SpatialTracker [54] introduces the triplane representation for group pixels in 3D.DINOTracker [46] performs test-time finetuning on the pre-trained DINO-ViT [37] model and achieves fine-grained tracking per video. When applying these query points-based methods to achieve dense tracking, full points need to be processed individually or in batches, which brings computational hurdles and limits their downstream applications [35].

2.2. Image-based dense point tracking

Recent studies turn to tracking every point within a frame in a single run, aiming to enhance neighborhood relationships while reducing computational requirements. OmniMotion [49] performs pixel-wise tracking between local and canonical space to maintain the global consistency of the motion, and then FastOmniTrack [42] and DecoMotion [26] improve from the perspectives of computation cost and object motion decomposition. CPFlow [31] proposes to estimate spatio-temporally dense motion curves, but it can only input 4 images and needs pre-sampling for longer video. AccFlow [52] proposes the forward and backward aggregation pipeline, extending inter-frame dense optical flow to multi-frame long ranges. MFT [36] select chaining multi-frame candidates and FlowTrack [4] automatically apply error compensation in instances of tracking inaccuracies. DOT [35] unifies point tracking and optical flow, upgrading sparse tracks to dense flow fields between every frame. However, limited by the video frame rate, these methods are struggling to model challenging dynamics. In this work, we propose to introduce continuous event data into the input video to enable temporally continuous point tracking.

2.3. Event-based motion estimation

Thanks to the motion-sensitive nature of event cameras, extensive motion estimation studies in recent years have highlighted their potential applications to challenging dynamics. The feature tracking methods [11, 27, 33, 50] show the benefits of event cameras for low-latency tracking, but can only track sparse, specific textured locations. Recently estimating optical flow from events has become mainstream. Using only sparse event data [13, 32, 62] allows to estimate satisfactory dense optical flow, while introducing data from other sensors such as images [47, 59] and point clouds [48, 60] achieves significant performance

gains. BFlow [14] and MotionPriorCMax [18] exploit the continuous property of events to estimate parametric Bézier trajectories, but can only estimate motion within a fixed consecutive frame interval and cannot be directly adapted to long-term sequences. Recently, FE-TAP [29] proposes to recover point trajectories from a fixed number of images and events based on TAPVid [5], but does not take full advantage of the continuous nature of events. We propose to combine the advantages of images and events to enable temporally continuous point tracking by modeling long-term global motion with an arbitrary number of frames.

3. Method

Overview. To address the limitations of sparse events and low video frame rates in modeling fine-grained motion, we present the first framework that recovers dense and continuous point trajectories from a video with corresponding event sequences. Our framework consists of four parts: 1) A parametric multi-frame continuous point trajectories representation; 2) A model of event triggering along the point trajectories; 3) A two-frame basis motion estimation model; 4) A multi-frame motion aggregation and streaming framework.

Problem Formulation. Conventional shutter camera captures a video with N_v frames $\{I_i\}_{i=1}^{N_v}$ at a fixed frame rate. Event camera triggers an unbounded event sequence $\{e_i\}_{i=1}^{N_e}$, where N_e is the number of events. Each event $e_i = \{\mathbf{x}_i, t_i, p_i\}$ consists of the pixel position $\mathbf{x}_i = (x, y)$, timestamp t_i with microsecond precision, and the brightness change polarity p_i in log domain. Our goal is to combine these two modalities to recover the spatially dense and temporally continuous trajectories $\mathbf{T}_{1 \rightarrow t}$ of all points starting from the first frame at time 1 to last frame at time t .

3.1. Motion model

Trajectory representation. Previous point tracking methods typically estimate two-channel motion vectors in the XY directions, which is the optical flow when representing dense trajectories [4, 35]. To learn the curve trajectory from the deep network, we instead learn the multiple control points as curved trajectories [31]. Specifically, we choose the B-spline curve as our curve representation, which is defined by N_c control points $\{\mathbf{P}_i\}_{i=1}^{N_c}$ and basis functions $\{B_{i,p}(t)\}_{i=1}^{N_c}$ with degree p . The continuous trajectory $\mathbf{T}(t)$ represented by b-spline curve in time variable t is a collection of piecewise polynomial functions $\mathbf{T}(t) = \sum_{i=1}^{N_c} B_{i,p}(t)\mathbf{P}_i$. Similar to learning dense flow, each pixel has an independent curve with control points denoted as $\mathbb{P} \in \mathbb{R}^{N_c \times 2 \times H \times W}$, where $H \times W$ is the image size. This enables learnable motion modeling of dense and continuous point trajectories \mathbf{T} with parametric curve representation. More details are provided in the Supp.

Multi-frame global trajectories accumulation. Existing parametric motion modeling methods are fixed in the number of frames they can handle, *e.g.*, BFlow [14] is limited to between two frames, and CPFlow [31] struggles to benefit from more than 4 frame inputs, resulting in suboptimal long-term trajectory modeling. Inspired by the practice of multi-frame optical flow accumulation [36, 52], we propose a new multi-frame curve trajectories accumulation strategy to handle long-term videos with arbitrary frames.

Our accumulation framework works on a streaming pipeline, where the previous global trajectory $\mathbf{T}_{1 \rightarrow t}$ with $(t-1) \times N_c$ control points has been accumulated from the previous $t-1$ local trajectories $\{\mathbf{T}_{i \rightarrow i+1}\}_{i=1}^{t-1}$ when processing the t -th step. When we get updated t -th local trajectory $\mathbf{T}_{t \rightarrow t+1}$ with N_c control points from time t to $t+1$, a simple approach is to directly accumulate the new global trajectory with $t \times N_c$ control points from time 1 to $t+1$ by $\mathbf{T}_{1 \rightarrow t+1}(\mathbf{x}) = [\mathbf{T}_{1 \rightarrow t}(\mathbf{x}), \text{Warp}(\mathbf{T}_{t \rightarrow t+1}, \mathbf{T}_{1 \rightarrow t})(\mathbf{x})]$, where Warp represents backward warping, the operator $[\cdot]$ combines the control points of two sub-curves and creates a more complex curve. However, there are two problems with just warping operation: 1) It suffers from numerical error as integer sampling with floating-point coordinates, *i.e.*, for warping vectors from \mathbf{b} to \mathbf{a} , $\text{Warp}(\mathbf{a}, \mathbf{b})(\mathbf{x}) = \mathbf{a}(\mathbf{x} + \mathbf{b}(\mathbf{x}))$, \mathbf{x} is integer coordinates but not $\mathbf{x} + \mathbf{b}(\mathbf{x})$. 2) Some points may be occluded at time t , resulting in failing to find the corresponding points. More details are provided in the Supp.

Our framework iteratively maintains and learns to integrate from a global motion representation $\mathbf{M}_{1 \rightarrow t}^{global}$ in the streaming process (*cf.* Sec. 3.2). For the first numerical problem, we estimate a start point offsets $\mathbf{O}_t \in \mathbb{R}^{2 \times H \times W}$ learned from $\mathbf{M}_{1 \rightarrow t}^{global}$ and normalized to the range $[-1, 1]$. Sampling compensation is achieved by adding this offset directly during warping. For the second occlusion problem, we introduce an occlusion solving strategy for occluded pixels. We additionally estimate the visibility map $\mathbf{V}_{1 \rightarrow t}$ of each point from the initial frame to the t -th frame as well as the trajectory update $\Delta \mathbf{T}_t$. Aggregation is based on a warp with offset when the point x is visible. When point x is occluded, a learnable module Fusion is introduced to regress the point’s coarse motion trajectory in $t \rightarrow t+1$ from $\mathbf{M}_{1 \rightarrow t}^{global}$. Finally, the trajectory update $\Delta \mathbf{T}_t$ as a uniformly refinement serves the global trajectory accumulation goal. The refinement process can be modeled as follows:

$$\mathbf{T}' = \begin{cases} \text{Warp}(\mathbf{T}_{t \rightarrow t+1}, \mathbf{T}_{1 \rightarrow t}, \mathbf{O}_t) + \Delta \mathbf{T}_t & \text{if } \mathbf{V}_{1 \rightarrow t}(\mathbf{x}) = 1, \\ \text{Fusion}(\mathbf{T}_{t \rightarrow t+1}, \mathbf{T}_{1 \rightarrow t}, \mathbf{M}_{1 \rightarrow t}^{global}) + \Delta \mathbf{T}_t & \text{if } \mathbf{V}_{1 \rightarrow t}(\mathbf{x}) = 0, \end{cases} \quad (1)$$

where \mathbf{T}' is the local curve in $t \rightarrow t+1$ warped to the starting coordinates at time 1. Then we direct combine it with the cached previous global curve to obtain the accumulated current global curve $\mathbf{T}_{1 \rightarrow t+1}^{accum}(\mathbf{x}) = [\mathbf{T}_{1 \rightarrow t}(\mathbf{x}), \mathbf{T}'(\mathbf{x})]$.

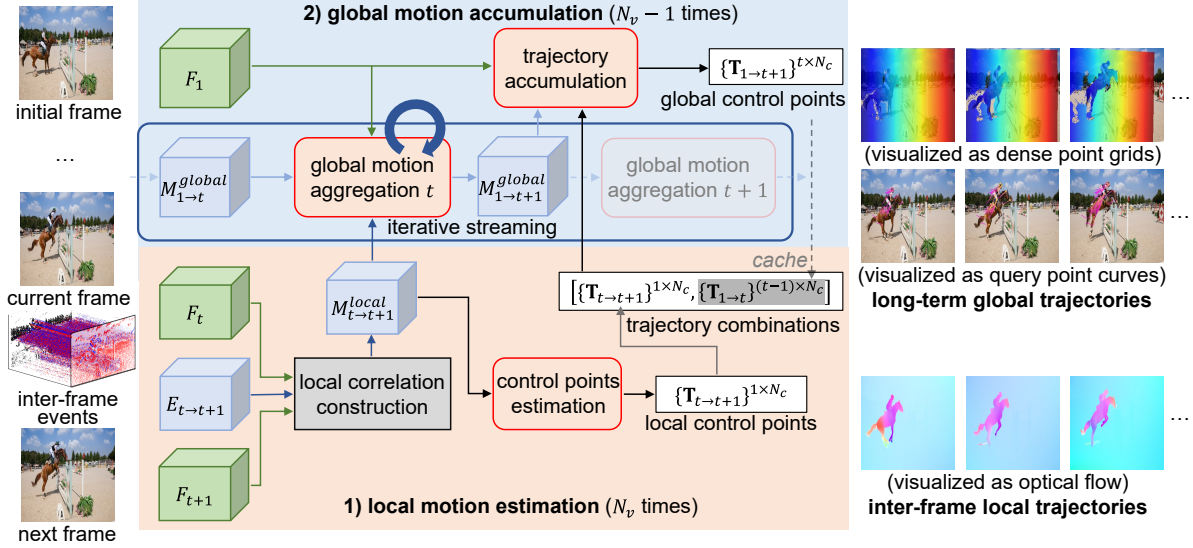


Figure 1. Our proposed event-aided dense and continuous point tracking framework consists of two main steps. 1) *Local motion estimation*: estimating short-term curve trajectories with N_c control points from two consecutive images and inter-frame events, while concurrently updating the local motion representation. 2) *Global motion accumulation*: iteratively fusing the latest local motion representation with the previous global motion representation in a streaming manner for aggregating the latest global motion representation. Subsequently, the global long-term curve trajectories with $t \times N_c$ control points are optimized on trajectory combinations.

Events along the trajectory. Following the contrast maximization framework [8], we assume that events are triggered along with the pixel motion trajectories at the moving boundary. For a motion trajectory \mathbf{T} starting from pixel \mathbf{x}_1 at time t_1 , the generated event’s coordinates satisfy the trajectory, *i.e.*, $\mathbf{x}_1 = \mathbf{T}(t_1)$, $\mathbf{x}_i - \mathbf{x}_1 = \mathbf{T}(t_i) - \mathbf{T}(t_1)$. We can thus use the motion trajectory to transform the events $e_i \doteq \{\mathbf{x}_i, t_i, p_i\}$ back to time t_1 :

$$e_i' \doteq \{\mathbf{x}_i' = \text{Warp}(\mathbf{x}_i; \mathbf{T}(t_i) - \mathbf{T}(t_1)), t_1, p_i\}. \quad (2)$$

Assuming the trajectory \mathbf{T} is accurate, this process transforms the event e_i to the starting point position \mathbf{x}_1 of the trajectory, *i.e.*, $\mathbf{x}_1 = \text{Warp}(\mathbf{x}_i; \mathbf{T}(t_i - t_1))$. Based on the correlated motion modeling of events and point trajectories, we build additional self-supervised training objectives in Sec. 3.3 to alleviate the lack of continuous trajectory annotations in the training datasets.

3.2. Framework

Two-frame basis model. The two-frame basis model is designed to recover inter-frame short-term trajectories $\mathbf{T}_{t \rightarrow t+1}$ from the encoded features of input two consecutive frames F_t, F_{t+1} and inter-frame events $E_{t \rightarrow t+1}$. In the feature extraction phase, we first convert the raw event data into a dense grid representation [41], followed by two feature encoders for two images and an event grid. Subsequently, we construct the local correlation between two frame features by matrix multiplication [44]. By leveraging the local correlation and events, we learn the local motion representation $M_{t \rightarrow t+1}^{local}$ by a motion extractor which allows

recovery of local dense trajectories $\mathbf{T}_{t \rightarrow t+1}$ by a trajectory decoder. Specifically, the former utilizes a transformer fusion module in GMA [22], while the latter is based on the typical FPN decoder in RAFT [43]. The trajectory decoder estimates the coordinates of N_c control points $\mathbb{P}_{t \rightarrow t+1}$ and a single-channel visibility map $\{\mathbf{V}\}_{t \rightarrow t+1}$, which is essential for establishing multi-frame global accumulation in Eq. (1).

Global motion aggregation module. For processing a video comprising N_v frames, the above two-frame basis model needs to be streamed sequentially $N_v - 1$ times yielding multiple local motion representations $\{M_{i \rightarrow i+1}^{local}\}_{i=1}^{t-1}$ and curve trajectories $\{\mathbf{T}_{i \rightarrow i+1}\}_{i=1}^{t-1}$. To facilitate the accumulation of global multi-frame trajectories according to Sec. 3.1, the cached global motion representation $\mathbf{h}_{t-1} \doteq M_{1 \rightarrow t}^{global}$ from the previous t -frames is utilized as the query, while the current local motion representation $\mathbf{h}_t^l \doteq M_{t \rightarrow t+1}^{local}$ serves as the key and value. We first perform the linear projections and compute the cross-attention:

$$\begin{aligned} \text{CA}(\mathbf{h}_{t-1}, \mathbf{h}_t^l, W_{Q,K,V}) &= \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \\ &= \text{softmax}\left(\frac{(W_Q \cdot \mathbf{h}_{t-1})(W_K \cdot \mathbf{h}_t^l)^T}{\sqrt{d_k}}\right)(W_V \cdot \mathbf{h}_t^l), \end{aligned} \quad (3)$$

where d_k is the channel size, \cdot is the linear projection and $W_{Q,K,V}$ are the corresponding weights. We then conduct iterative fusion based on the gated activation unit (GRU) [3], where the update gate is $\mathbf{z}_t =$

$\text{sigmoid}(\text{CA}(\mathbf{h}_{t-1}, \mathbf{h}_t^l, W_{Q,K,V}))$, the reset gate is $\mathbf{r}_t = \text{sigmoid}(\text{CA}(\mathbf{h}_{t-1}, \mathbf{h}_t^l, W'_{Q,K,V}))$, and the hidden state is $\mathbf{s}_t = \tanh(\text{CA}(\mathbf{r}_t \odot \mathbf{h}_{t-1}, \mathbf{h}_t^l, W''_{Q,K,V}))$, \odot is the element-wise multiplication. The superscript of $W_{Q,K,V}$ denotes the different projection weights taken independently in each attention calculation. Finally, we iteratively update the current global motion representation at the feature level by:

$$\mathbf{M}_{1 \rightarrow t+1}^{global} \doteq \mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \mathbf{s}_t. \quad (4)$$

The simple and effective recurrent temporal aggregation we take is naturally compatible with the streaming pipeline, and we verify its effectiveness compared to previous fusion solutions in ablation experiments (*cf.* Sec. 4.4).

Multi-frame iterative streaming framework. As depicted in Fig. 1, our framework iteratively processes the input video and event data through local motion estimation and global motion accumulation. We aggregate the local motion representations from each frame interval to the global motion representation at the feature level through the above global aggregation module in Sec. 3.2. Subsequently, the multi-frame accumulation step described in Sec. 3.1 sequentially combines each inter-frame short-term curve into a global long-term motion trajectory at the trajectory level, providing the dense and continuous point tracking representation as the model output. On the right side of Fig. 1, local motion is visualized with dense optical flow, and global motion is represented with deformations of dense point grid and curve trajectories of sparse query points.

3.3. Objective

Temporal discrete trajectory supervision. The available point tracking datasets provide only temporally discrete point tracks with no ground truth for continuous inter-frame trajectories. Following DOT [35], we first adopt supervised losses based on the temporal discrete ground-truth point tracks provided by the dataset, which consists of the L1 loss L_{traj} for sampled discrete trajectory prediction and the binary cross-entropy loss L_{vis} for visibility map. We then randomly select different frame intervals for augmented training. Local correlation is not constructed when the frames are skipped, therefore the corresponding event features are taken into streaming for iteratively updating the global motion representation. There are cases where some images are not used as input when the frame interval is greater than 1, but the corresponding input events and ground-truth tracks can be regarded as inter-frame motion contributing to curve trajectory learning. Such sampling-based augmentation ensures the model learning through diverse long- and short-term motions, capitalizing on the continuity of events to estimate continuous trajectories.

Event consistency with continuous trajectory. Since events are usually generated along motion trajectories, we propose to leverage the continuous property of events for self-supervised continuous trajectory learning in conjunction with discrete supervision of point trajectories. However, events are computationally intensive to process one by one and are generally accompanied by noise. We thus first introduce event temporal chunking to process events in batches within a fixed duration to reduce the noise impact and computation. For the b -th interval of B chunks, we isolate the events within that b -th chunk and aggregate them after warping them to t_b as Eq. (2). For each chunk, the events then are summed into an image of warped events (IWE) [8], *i.e.*, $\mathbf{EB}(\mathbf{x}_i, b) \doteq \sum_{i=1}^{N_e} \mathcal{N}(\mathbf{x}_i; x'_i, \sigma^2)$, where $t_b \leq t_i < t_{b+1}$ and σ is the neighboring range which is usually chosen as 1 pixel. This IWE essentially counts the number of warped events e'_i per pixel and chunk. The chunking intervals are chosen randomly to exploit the continuous nature of events. Thus we can establish consistent connections between event chunks and continuous trajectories:

$$L_{ec} = \sum_{\mathbf{x}} \sum_{b_1 \neq b_2}^{\Omega, B} \rho \left(\mathbf{EB}(\mathbf{x}, b_1), \text{Warp}(\mathbf{EB}(\mathbf{x}, b_2); \mathbf{T}(t_{b_2}) - \mathbf{T}(t_{b_1})) \right), \quad (5)$$

where ρ is the consistency measure with L1 norm. Since events are spatially sparse, we only establish connections with the dense trajectories at locations Ω with valid events.

Image consistency with discrete trajectory. Similar to the unsupervised optical flow task [30], we can also establish the discrete consistency of the motion trajectory with the images at discrete times, to compensate for the spatial sparsity issue of ground-truth point tracks in the training data. In addition, in our sampling-based augmented training, skipped images can be used as additional continuity training objectives.

For the accumulated continuous global trajectory $\mathbf{T}_{1 \rightarrow t}$, we sample the the discrete optical flow $\mathbf{F}_{i \rightarrow j}$ from I_i to I_j via timestamps. Similar to Eq. (5), the consistency of images can be modeled as:

$$L_{ic} = \sum_{\mathbf{x}} \sum_{i \neq j}^{\Omega, N_v} \rho \left(I_i(\mathbf{x}), \text{Warp}(I_j(\mathbf{x}); \mathbf{F}_{i \rightarrow j}(\mathbf{x})) \right). \quad (6)$$

Total objective. Our total training objective is a weighted combination of the above objectives, *i.e.*, $L = L_{traj} + \lambda_1 L_{vis} + \lambda_2 L_{ec} + \lambda_3 L_{ic}$, λ are manual hyperparameters. Our ablations verify that combined self-supervised training can compensate for the temporal continuity of trajectories.

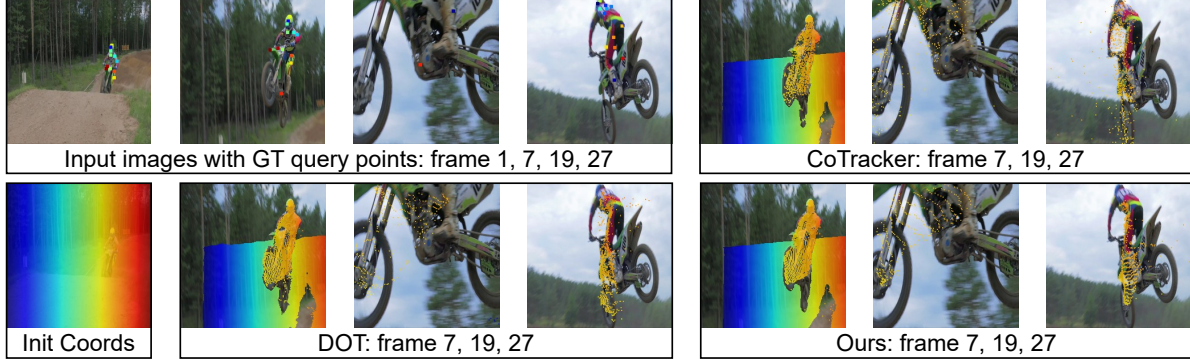


Figure 2. Visual comparisons of long-term dense point tracking on the TAP-Vid-DAVIS [5] dataset, with the ground-truth sparse query points with input images. Due to space limitations, we provide more sequences and event data visualization in the Supp. and demo video.

Table 1. Quantitative results of dense evaluation on the CVO test [52] and extended set [35].

Method	CVO (Clean)		CVO (Final)		CVO (Extended)		
	EPE _{all/vis/occ} ↓	OA ↑	EPE _{all/vis/occ} ↓	OA ↑	EPE _{all/vis/occ} ↓	OA ↑	
Query	PIPs++ [58]	9.05 / 6.62 / 21.5	33.3	9.49 / 7.06 / 22.0	32.7	18.4 / 10.0 / 32.1	58.7
	TAPIR [6]	3.80 / 1.49 / 14.7	73.5	4.19 / 1.86 / 15.3	72.4	19.8 / 4.74 / 42.5	68.4
	CoTracker [24]	1.51 / 0.88 / 4.57	75.5	1.52 / 0.93 / 4.38	75.3	5.20 / 3.84 / 7.70	70.4
Dense	GMA [22]	2.42 / 1.38 / 7.14	60.5	2.57 / 1.52 / 7.22	59.7	21.8 / 15.7 / 32.8	65.6
	MFT [36]	2.91 / 1.39 / 9.93	19.4	3.16 / 1.56 / 10.3	19.5	21.4 / 9.20 / 41.8	37.6
	AccFlow [52]	1.69 / 1.08 / 4.70	48.1	1.73 / 1.15 / 4.63	47.5	36.7 / 28.1 / 52.9	36.5
	DOT [35]	1.32 / 0.74 / 4.12	80.4	1.38 / 0.82 / 4.10	80.2	5.07 / 3.67 / 7.34	71.0
	EDCPT (Ours)	1.23 / 0.71 / 3.83	82.1	1.31 / 0.76 / 3.86	81.9	4.88 / 3.44 / 7.46	71.9

Table 2. Quantitative results of sparse evaluation on the TAP-Vid-DAVIS [5] point tracking benchmark.

Method	DAVIS (First)			DAVIS (Strided)			
	AJ ↑	$<\delta_{avg}^x$ ↑	OA ↑	AJ ↑	$<\delta_{avg}^x$ ↑	OA ↑	
Query	TAP-Net [5]	33.0	48.6	78.8	38.4	53.1	82.3
	Context-PIPs [51]	42.7	60.3	79.5	48.9	64.0	83.4
	TAPIR [6]	56.2	70.0	86.5	61.3	73.6	88.8
	CoTracker [24]	61.1	74.6	89.1	63.5	79.8	87.8
	SpatialTracker [54]	61.1	76.3	89.5	-	-	-
Dense	CPFlow [31]	9.6	14.6	-	-	-	-
	MFT [36]	47.3	66.8	77.8	56.1	70.8	86.9
	DecoMotion [26]	53.0	69.9	84.2	60.2	74.4	87.2
	DinoTracker [46]	-	-	-	62.3	78.2	87.5
	FlowTrack [4]	-	-	-	63.2	76.3	89.2
	DOT [35]	61.6	75.5	89.5	66.7	80.6	90.4
	EDCPT (Ours)	63.8	76.3	90.6	67.5	80.5	91.1

4. Experiments

4.1. Experimental details

We train our model on MOVI-F [15] dataset and directly evaluate it on CVO [52] and TAP-Vid-DAVIS [5] benchmarks, using the vid2e [10] simulator to generate events. Following DOT [35], we train the model for 500k steps on $4 \times$ NVIDIA L40 48G GPUs, using the Adam optimizer and OneCycle learning rate decay with a maximum of 10^{-4} . We choose 3 frames as training samples, along with the random selection of up to 10 frames in different frame intervals. The loss hyperparameters are set to 1.0,

0.1, 0.1. For dense evaluation, we report the dense absolute error $EPE_{all/vis/occ}$ for all/visible/occluded points, as well as occlusion IoU accuracy OA for estimated visible mask. For sparse evaluation, we follow TAPNet [5] and report average Jaccard AJ, position accuracy $<\delta_{avg}^x$, and occlusion accuracy OA. Additionally, we adopt the real-captured event-based optical flow benchmark, DSEC [12, 13], to verify our adaptation capacity. More details are provided in the Supp.

4.2. Standard spatially dense point tracking

Performance on CVO and TAP-Vid-DAVIS benchmarks. We conduct comprehensive evaluations on two common standard point tracking benchmarks. Consistent with DOT [35], we report the quantitative results of the spatially dense optical flow from the last to the first frame of CVO [52] dataset in Tab. 1. Our proposed new EDCPT framework achieves significant performance improvements, whether comparing methods that only predict the partial *Query* points or directly estimating spatially *Dense* trajectories within a single inference. Particularly, we achieve 0.19 EPE_{all} and 0.9 OA improvements on the extended set of 476 videos with 48 frames when compared to DOT [35].

In contrast, the real TAP-Vid-DAVIS dataset [5] only provides ground-truth trajectories for selected query points. As a result, we only sparsely evaluate these points for a fair comparison despite the output trajectories of our model and some compared methods are spatially dense. The



Figure 3. Visual comparisons of dense and continuous point trajectories on TAP-Vid-DAVIS [5]. Zoom in for detailed curve trajectories.

Table 3. Quantitative results on the real DSEC optical flow benchmark [13]. The first two methods are categorized as self-supervised learning and the remaining as supervised learning.

Method	Input	EPE ↓	AE ↓	%Out ↓
Taming [39]	Events	2.33	10.56	17.77
MPCMax [18]	Events	3.20	8.53	15.21
E-RAFT [13]	Events	0.79	2.85	2.68
TMA [28]	Events	0.74	2.68	2.30
IDNet [53]	Events	0.72	2.72	2.04
ECDDP [56]	Events	0.70	2.58	1.96
BFlow [14]	Events	0.75	2.68	2.44
BFlow [14]	Images+Events	0.69	2.42	1.88
EDCPT (Ours)	Images+Events	0.63	2.17	1.52

quantitative results in Tab. 2 demonstrate the superiority of our framework, outperforming the existing state-of-the-art methods DOT [35] and SpatialTracker [54] with up to 2.7 AJ and 1.1 OA. We argue that ours $< \delta_{\text{avg}}^x$ is close to DOT’s because the sparse to dense upgrade strategy from sparse point matches is capable of modeling the rigid object motion and thus benefiting the proportion evaluation. We also perform qualitative visual comparisons in Fig. 2. Combining the above quantitative and qualitative comparisons with previous image-based methods, the new attempts of incorporating events by our framework significantly improve the accuracy with the standard dense point tracking setting.

Performance on DSEC benchmark. We further conduct experiments on the DSEC benchmark [13] with real captured event data. Because the DSEC online leaderboard only measures the optical flow between two consecutive frames, we therefore finetune the local motion estimation on the DSEC training set from our pre-trained long-term model. The submission results are shown in Tab. 3. Notably, while BFlow [14] can estimate curve trajectories, they only submitted another optical flow version. Our framework fuses image and event data, yielding state-of-the-art performance with improvements of 0.05 EPE and 0.25 AE.

4.3. Temporally continuous point tracking

We adapt the above standard evaluation to input only a portion of the full video frames into the model to evaluate the temporal continuity with the ground truths of skipped

frames. For the CVO final set with 7 frames per video, we skip 1 frame (*half*) and 2 frames (*one-third*) as model inputs, because the longer extended set lacks multi-frame ground-truth tracks. For the TAP-Vid-DAVIS dataset with ~ 100 frames, we report skip 1 frame (*half*) and skip 3 frames (*quarter*). The compared image-based methods cannot model inter-frame motion, thus we take linear motion interpolation to generate trajectory when frames are skipped. We retrain AccFlow [52] and mark it with * as its public version does not support forward point tracking.

As reported in Tab. 4, our proposed new framework with global continuous trajectory accumulation significantly outperforms existing methods. Especially in nonlinear motion scenarios of TAP-Vid-DAVIS datasets, the larger frame intervals lead to greater performance gaps. In addition to the ablation of different motion assumptions in Tab. 6, the B-spline representation we adopt achieves better performance. We also provide visual comparisons in Fig. 3 and a demo video of continuous trajectory visualization in the Supp., that includes four sequences with diverse motion complexity and event sparsity from TAP-Vid-DAVIS and real-captured ERF-X170FPS [25] dataset. With extensive experiments on both simulated events (CVO and TAP-Vid-DAVIS) and real events (DSEC and ERF-X170FPS), we fully validate the robustness of the proposed global motion accumulation in modeling complex continuous trajectories.

4.4. Ablation experiments and discussions

To perform progressive ablations in Tab. 5, Tab. 6, and Tab. 7, the underlined settings are those utilized in the previous table, and the bolded ones represent the choices for our final framework. To validate the capability for continuous point tracking, the ablation results are reported with CVO third and DAVIS quarter settings as in Sec. 4.3 and Tab. 4.

Global motion aggregation. One of our key contributions is the global aggregation of local motion representations in the *streaming* pipeline. In Tab. 5, N/A indicates do not explicitly model sequential motion as in DOT [35]. *post* is the forward aggregation in AccFlow [52], which performs post-processing only on the output optical flow and ignores the internal long-term motion information. We aggregate

Table 4. Continuous point tracking evaluation results on the CVO extended set [35] and TAP-Vid-DAVIS dataset [5].

Method	CVO (Final) – EPE _{all/vis/occ} ↓			DAVIS (First) – AJ / δ_{avg}^x ↑		
	full	half	third	full	half	quarter
RAFT	2.09 / 0.81 / 8.02	2.44 / 0.95 / 9.07	3.02 / 1.16 / 11.42	33.9 / 46.6	28.6 / 40.1	22.4 / 34.2
GMA	1.99 / 0.77 / 7.57	2.35 / 0.89 / 8.45	2.92 / 1.09 / 10.97	39.3 / 52.5	31.7 / 44.3	26.5 / 38.3
AccFlow*	2.28 / 0.60 / 11.18	2.39 / 0.80 / 10.74	2.79 / 1.02 / 11.37	47.2 / 62.3	37.5 / 49.2	30.9 / 42.4
CoTracker	1.89 / 0.63 / 7.05	2.11 / 0.82 / 8.02	3.17 / 1.65 / 11.13	61.1 / 74.6	54.3 / 68.8	48.9 / 63.9
DOT	1.83 / 0.59 / 6.95	2.10 / 0.73 / 7.88	2.69 / 0.97 / 10.84	61.6 / 75.5	55.6 / 70.1	50.4 / 65.3
EDCPT (Ours)	1.76 / 0.55 / 6.73	1.97 / 0.66 / 7.61	2.16 / 0.73 / 8.76	63.8 / 76.3	59.7 / 73.1	56.2 / 70.9

Table 5. Ablations on global motion aggregation.

	EPE _{all/vis/occ} ↓	AJ / δ_{avg}^x ↑
N/A	2.54 / 0.89 / 10.33	51.9 / 66.4
post	2.49 / 0.85 / 9.73	52.5 / 66.9
solo	2.45 / 0.82 / 9.89	52.7 / 67.0
–offsets	2.43 / 0.82 / 9.73	53.0 / 67.2
stream	2.42 / 0.80 / 9.68	53.3 / 67.4

Table 6. Ablations on parametric curve representation.

	EPE _{all/vis/occ} ↓	AJ / δ_{avg}^x ↑
<u>linear</u>	2.42 / 0.80 / 9.68	53.3 / 67.4
quad	2.49 / 0.84 / 9.46	54.2 / 68.1
$N_c = 3$	2.32 / 0.79 / 9.33	54.4 / 68.6
$N_c = 4$	2.23 / 0.76 / 8.97	55.4 / 69.7
$N_c = 5$	2.26 / 0.75 / 9.20	55.0 / 69.3

Table 7. Ablations on input data and supervision.

	EPE _{all/vis/occ} ↓	AJ / δ_{avg}^x ↑
Images	2.50 / 0.86 / 9.92	52.5 / 66.7
+Events	2.23 / 0.76 / 8.97	55.4 / 69.7
+ L_{ic}	2.20 / 0.75 / 8.90	55.8 / 70.2
+ L_{ec}	2.16 / 0.73 / 8.76	56.2 / 70.9

motion representations at the feature level instead of dealing directly with motion vectors. *solo* is the short and long term fusion module in SOLOFusion [40]. Unlike its temporal fusion with a fixed number of frames and low resolution, we adopt sequential modeling with an unspecified number of frames in a streaming pipeline. Our proposed motion aggregation framework fuses image correspondence and event features from local to global with recurrent *streaming* structure, which better preserves multi-frame motion information and achieves optimal performance. In addition, we also verified that removing the additional offset estimation to address the numerical problem in Warping leads to a slight performance degradation, as this would require the subsequent refinement to handle it simultaneously.

Curve representation. Based on our streaming framework, we compare the performance improvement of taking the B-spline curve representation compared to interpolating with *linear* and *quadratic* motion assumptions in Tab. 6. For the control points number of B-spline curves, we chose $N_c = 4$, since further increasing the number of control points does not significantly improve the performance under CVO third and DAVIS quarter settings.

Input data and supervision. Since previous methods usually use only image data, we verify the advantages of incorporating event data for high-precision continuous point tracking by removing events, as depicted in Tab. 7. Moreover, the comparison results between this image-only variant and DOT [35] in Tab. 4 demonstrates that our proposed streaming aggregation and curve representation are beneficial even in the absence of event data. Furthermore, we validate that training using the proposed image and event-to-point trajectory consistencies as additional supervision complements the lack of continuous inter-frame tracks in

the training data and can further improve performance.

Limitations. On the same RTX 3090 PC, our framework processes a 48-frame video at 512x512 resolution in 12.6 seconds, significantly faster than CoTracker’s 11 minutes, while only processing partial query points in a single run. However, it is slower than the two-frame optical flow method GMA of only 2.1 seconds and slightly slower than the multi-frame point tracking method DOT of 9.5 seconds. In addition, there is currently a lack of point-tracking datasets with real events, and it is difficult to obtain long-term tracking annotations. We evaluate on standard benchmarks with simulated events, as well as visualize point tracking on real events, and also evaluate optical flow estimation on a real event benchmark. Our future work plans to improve on model efficiency and the evaluation setup.

5. Conclusion

In this paper, we propose a new long-term dense point tracking framework for integrating image and event data to estimate continuous motion trajectories. Specifically, we streamingly process two adjacent frames and inter-frame events to generate local motion representations, and accumulate previous caches to global motion representation at the feature level to produce new global trajectories at the trajectory level. We utilize multi-frame parametric curve accumulation to represent continuous trajectories with any number of frames, complemented by image and event-to-trajectory consistencies for model training. We validate the effectiveness of our framework by conducting extensive benchmark experiments as well as qualitative verification on real challenge data. We believe this work provides new insights into the point tracking task from the perspective of event modality and continuous global curve representations.

Acknowledgments

This research/project is supported by the National Natural Science Foundation of China (62271410, 12150007), the National Research Foundation, Singapore, under its NRF- Investigatorship Programme (Award ID. NRF-NRFI09-0008), and the Tier 1 grant T1-251RES2305 from the Singapore Ministry of Education. Zhexiong Wan was also supported by the Program of China Scholarship Council (202306290193), and the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University (CX2023013).

References

- [1] Eirikur Agustsson, David Minnen, Nick Johnston, Johannes Balle, Sung Jin Hwang, and George Toderici. Scale-space flow for end-to-end optimized video compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8503–8512, 2020. 1
- [2] Adam Bielski and Paolo Favaro. Move: Unsupervised movable object segmentation and detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:33371–33386, 2022. 1
- [3] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014. 4
- [4] Seokju Cho, Jiahui Huang, Seungryong Kim, and Joon-Young Lee. Flowtrack: Revisiting optical flow for long-range dense tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19268–19277, 2024. 1, 2, 3, 6
- [5] 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. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, pages 13610–13626, 2022. 1, 2, 3, 6, 7, 8
- [6] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10061–10072, 2023. 2, 6
- [7] Carl Doersch, Yi Yang, Dilara Gokay, Pauline Luc, Skanda Koppula, Ankush Gupta, Joseph Heyward, Ross Goroshin, João Carreira, and Andrew Zisserman. Bootstap: Bootstrapped training for tracking-any-point. In *Asian Conference on Computer Vision (ACCV)*, 2024. 2
- [8] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3867–3876, 2018. 4, 5
- [9] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(1):154–180, 2022. 1
- [10] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3586–3595, 2020. 6
- [11] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. EKLt: Asynchronous photometric feature tracking using events and frames. *International Journal of Computer Vision (IJCV)*, 128(3):601–618, 2020. 2
- [12] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. DSEC: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters (RA-L)*, 2021. 6
- [13] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-RAFT: Dense optical flow from event cameras. In *International Conference on 3D Vision (3DV)*, pages 197–206, 2021. 2, 6, 7
- [14] Mathias Gehrig, Manasi Muglikar, and Davide Scaramuzza. Dense continuous-time optical flow from event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 46(7):4736–4746, 2024. 1, 3, 7
- [15] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanaprasgam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3749–3761, 2022. 6
- [16] Xiang Guo, Jiadai Sun, Yuchao Dai, Guanying Chen, Xiaoqing Ye, Xiao Tan, Errui Ding, Yumeng Zhang, and Jingdong Wang. Forward flow for novel view synthesis of dynamic scenes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16022–16033, 2023. 1
- [17] Jesse Hagenaaers, Federico Paredes-Vallés, and Guido De Croon. Self-supervised learning of event-based optical flow with spiking neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7167–7179, 2021. 1
- [18] Friedhelm Hamann, Ziyun Wang, Ioannis Asmanis, Kenneth Chaney, Guillermo Gallego, and Kostas Daniilidis. Motion-prior contrast maximization for dense continuous-time motion estimation. In *European Conference on Computer Vision (ECCV)*, 2024. 3, 7
- [19] 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. Springer, 2022. 1, 2
- [20] Xueyan Huang, Yueyi Zhang, and Zhiwei Xiong. Progressive spatio-temporal alignment for efficient event-based motion estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1537–1546, 2023. 1
- [21] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2462–2470, 2017. 1
- [22] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9772–9781, 2021. 4, 6
- [23] Xin Jin, Longhai Wu, Jie Chen, Youxin Chen, Jayoon Koo, and Cheul-hee Hahm. A unified pyramid recurrent network for video frame interpolation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1578–1587, 2023. 1
- [24] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-tracker: It is better to track together. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 6
- [25] Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, and Kuk-Jin Yoon. Event-based video frame interpolation with cross-modal asymmetric bidirectional motion fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18032–18042, 2023. 7
- [26] Rui Li and Dong Liu. Decomposition betters tracking everything everywhere. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 6
- [27] Siqi Li, Zhikuan Zhou, Zhou Xue, Yipeng Li, Shaoyi Du, and Yue Gao. 3d feature tracking via event camera. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18974–18983, 2024. 2
- [28] Haotian Liu, Guang Chen, Sanqing Qu, Yanping Zhang, Zhijun Li, Alois Knoll, and Changjun Jiang. Tma: Temporal motion aggregation for event-based optical flow. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9685–9694, 2023. 7
- [29] 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. *arXiv preprint arXiv:2409.11953*, 2024. 3
- [30] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6489–6498, 2020. 5
- [31] Jianqin Luo, Zhexiong Wan, Yuxin Mao, Bo Li, and Yuchao Dai. Continuous parametric optical flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 23520–23532, 2023. 1, 2, 3, 6
- [32] Xinglong Luo, Ao Luo, Zhengning Wang, Chunyu Lin, Bing Zeng, and Shuaicheng Liu. Efficient meshflow and optical flow estimation from event cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19198–19207, 2024. 2
- [33] Nico Messikommer, Carter Fang, Mathias Gehrig, and Davide Scaramuzza. Data-driven feature tracking for event cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5642–5651, 2023. 1, 2
- [34] Etienne Meunier, Anaïs Badoual, and Patrick Bouthemy. Em-driven unsupervised learning for efficient motion segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(4):4462–4473, 2023. 1
- [35] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. Dense optical tracking: Connecting the dots. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3, 5, 6, 7, 8
- [36] Michal Neoral, Jonáš Šerých, and Jiří Matas. Mft: Long-term tracking of every pixel. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 6837–6847, 2024. 2, 3, 6
- [37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024. 2
- [38] Liyuan Pan, Miaomiao Liu, and Richard Hartley. Single image optical flow estimation with an event camera. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1669–1678, 2020. 1
- [39] Federico Paredes-Vallés, Kirk YW Scheper, Christophe De Wagter, and Guido CHE De Croon. Taming contrast maximization for learning sequential, low-latency, event-based optical flow. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9695–9705, 2023. 7
- [40] Jinyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris M. Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. In *International Conference on Learning Representations (ICLR)*, 2023. 8
- [41] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. Events-to-video: Bringing modern computer vision to event cameras. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3857–3866, 2019. 4
- [42] Yunzhou Song, Jiahui Lei, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Track everything everywhere fast and robustly. In *European Conference on Computer Vision (ECCV)*, 2024. 2
- [43] Xiuchao Sui, Shaohua Li, Xue Geng, Yan Wu, Xinxing Xu, Yong Liu, Rick Goh, and Hongyuan Zhu. CRAFT: Cross-attentional flow transformer for robust optical flow. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17602–17611, 2022. 4
- [44] 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. 1, 4
- [45] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time Lens: Event-based video frame interpolation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16155–16164, 2021. 1

- [46] Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. Dino-tracker: Taming dino for self-supervised point tracking in a single video. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 6
- [47] Zhexiong Wan, Yuchao Dai, and Yuxin Mao. Learning dense and continuous optical flow from an event camera. *IEEE Transactions on Image Processing (TIP)*, 31:7237–7251, 2022. 2
- [48] Zhexiong Wan, Yuxin Mao, Jing Zhang, and Yuchao Dai. RPEFlow: Multimodal fusion of RGB-pointcloud-event for joint optical flow and scene flow estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10030–10040, 2023. 2
- [49] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19795–19806, 2023. 2
- [50] Xiangyuan Wang, Huai Yu, Lei Yu, Wen Yang, and Gui-Song Xia. Towards robust keypoint detection and tracking: A fusion approach with event-aligned image features. *IEEE Robotics and Automation Letters (RA-L)*, 9(9):8059–8066, 2024. 2
- [51] BIAN Weikang, Zhaoyang Huang, Xiaoyu Shi, Yitong Dong, Yijin Li, and Hongsheng Li. Context-pips: Persistent independent particles demands spatial context features. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 55285–55298, 2023. 2, 6
- [52] Guangyang Wu, Xiaohong Liu, Kunming Luo, Xi Liu, Qingqing Zheng, Shuaicheng Liu, Xinyang Jiang, Guangtao Zhai, and Wenyi Wang. Accflow: backward accumulation for long-range optical flow. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12119–12128, 2023. 2, 3, 6, 7
- [53] Yilun Wu, Federico Paredes-Vallés, and Guido CHE De Croon. Lightweight event-based optical flow estimation via iterative deblurring. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 14708–14715. IEEE, 2024. 7
- [54] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 6, 7
- [55] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1647–1656, 2019. 1
- [56] Yan Yang, Liyuan Pan, and Liu Liu. Event camera data dense pre-training. In *European Conference on Computer Vision (ECCV)*, pages 292–310. Springer, 2024. 7
- [57] Jiqing Zhang, Yuanchen Wang, Wenxi Liu, Meng Li, Jinpeng Bai, Baocai Yin, and Xin Yang. Frame-event alignment and fusion network for high frame rate tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9781–9790, 2023. 1
- [58] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19855–19865, 2023. 2, 6
- [59] Hanyu Zhou, Yi Chang, Haoyue Liu, Wending Yan, Yuxing Duan, Zhiwei Shi, and Luxin Yan. Exploring the common appearance-boundary adaptation for nighttime optical flow. In *International Conference on Learning Representations (ICLR)*, 2024. 2
- [60] Hanyu Zhou, Yi Chang, and Zhiwei Shi. Bring event into rgb and lidar: Hierarchical visual-motion fusion for scene flow. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26477–26486, 2024. 2
- [61] Yi Zhou, Guillermo Gallego, Xiuyuan Lu, Siqi Liu, and Shaojie Shen. Event-based motion segmentation with spatio-temporal graph cuts. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 34(8):4868–4880, 2021. 1
- [62] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 989–997, 2019. 1, 2
- [63] Zhiyu Zhu, Junhui Hou, and Dapeng Oliver Wu. Cross-modal orthogonal high-rank augmentation for RGB-event transformer-trackers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22045–22055, 2023. 1