Event Stream-based Visual Object Tracking: A High-Resolution Benchmark Dataset and A Novel Baseline

Xiao Wang[1], Shiao Wang[1], Chuanming Tang[2,3], Lin Zhu[4], Bo Jiang [1]*, Yonghong Tian[5,6,7], Jin Tang[1]
[1]School of Computer Science and Technology, Anhui University, Hefei, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Institute of Optics and Electronics, CAS, Chengdu, China
[4]Beijing Institute of Technology, Beijing, China
[5]Peng Cheng Laboratory, Shenzhen, China
[6]National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University, China
[7]School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, China
{*xiaowang, jiangbo, tangjin*}*@ahu.edu.cn*, *wsa1943230570@126.com*,
tangchuanming19@mails.ucas.ac.cn, {*linzhu, yhtian*}*@pku.edu.cn*
https://github.com/Event-AHU/EventVOT_Benchmark

Abstract

Tracking with bio-inspired event cameras has garnered increasing interest in recent years. Existing works either utilize aligned RGB and event data for accurate tracking or directly learn an event-based tracker. The former incurs higher inference costs while the latter may be susceptible to the impact of noisy events or sparse spatial resolution. In this paper, we propose a novel hierarchical knowledge distillation framework that can fully utilize multi-modal / multi-view information during training to facilitate knowledge transfer, enabling us to achieve high-speed and low-latency visual tracking during testing by using only event signals. Specifically, a teacher Transformer-based multi-modal tracking framework is first trained by feeding the RGB frame and event stream simultaneously. Then, we design a new hierarchical knowledge distillation strategy which includes pairwise similarity, feature representation, and response maps-based knowledge distillation to guide the learning of the student Transformer network. In particular, since existing event-based tracking datasets are all low-resolution (346×260), we propose the first large-scale high-resolution (1280×720) dataset named EventVOT. It contains 1141 videos and covers a wide range of categories such as pedestrians, vehicles, UAVs, ping pong, etc. Extensive experiments on both low-resolution (FE240hz, VisEvent, COESOT), and our newly proposed high-resolution EventVOT dataset fully validated the effectiveness of our proposed method.

*✉ Corresponding Author: Bo Jiang

1. Introduction

Visual Object Tracking (VOT) targets predicting the locations of target object initialized in the first frame. Existing trackers are usually developed based on RGB cameras and deployed for autonomous driving, drone photography, intelligent video surveillance, and other fields. Due to the influence of challenging factors like fast motion, illumination, background distractor, and out-of-view, the tracking performance in complex scenarios is still unsatisfactory. The video frames with these challenges are unevenly distributed in the tracking video, making it difficult to improve the overall tracking results by investing more labeled data.

To address these challenges, some researchers have started to improve the effectiveness of input data by introducing new sensors. As a new type of bio-inspired sensor, event cameras are different from traditional video frame sensors in that they can output event pulses asynchronously and capture motion information through the detection of events (e.g., changes in light intensity). Event camera performs better than traditional RGB cameras in capturing fast-moving objects due to dense temporal resolution. It also works well on high dynamic range, low energy consumption, and low latency [14]. Event cameras can be used for a wide range of applications, including surveillance, robotics, medical imaging, and sports analysis.

Although few, there have been some studies that ex-

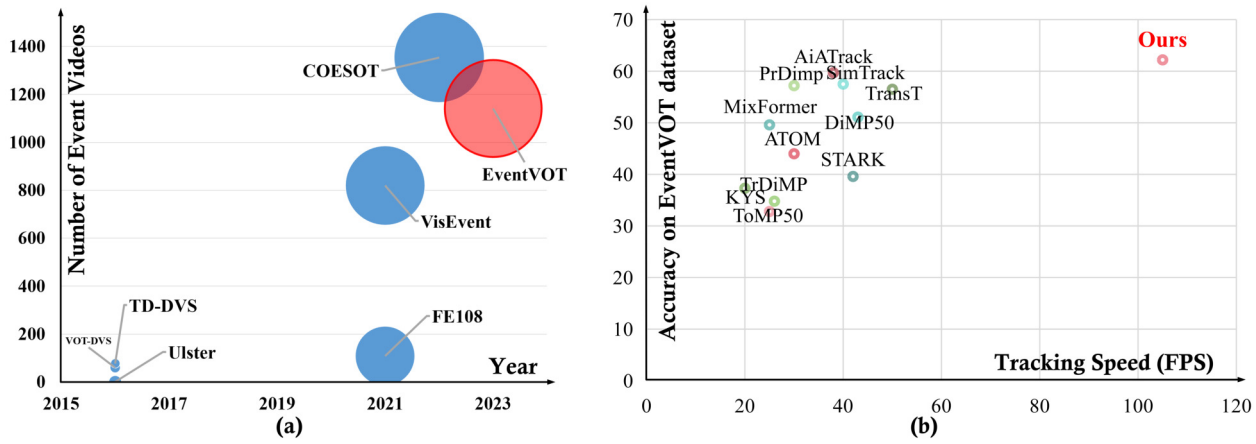

Figure 1. (a). Comparison between our newly proposed EventVOT and other event-based tracking datasets; (b). Comparison between our tracker and existing SOTA trackers on the tracking speed and accuracy on the EventVOT dataset.

ploit event cameras for visual object tracking. For example, Zhang et al. propose AFNet [36] and CDFI [34] to combine the frame and event data via multi-modality alignment and fusion modules. STNet [35] is proposed to connect the Transformer and spiking neural networks for event-based tracking. Zhu et al. [41] attempt to mine the key events and employ a graph-based network to embed the irregular spatio-temporal information of key events into a high-dimensional feature space for tracking. These works attempt to obtain stronger tracking algorithms through multi-modal fusion or pure event training and tracking methods. Although good performance can be achieved, however, these algorithms are still easily influenced by the following issues: **Firstly**, the spatial signal of event cameras is very sparse in slow-moving scenes, and the contours of target objects are not clear enough, which may lead to tracking failures. Tracking using RGB-Event data can better compensate for this deficiency, but additional modalities will increase the inference cost. **Secondly**, existing event-based tracking datasets are collected using the DVS346 camera, which has an output resolution of 346×260. It has not been explored or validated whether the event representation methods designed for low-resolution event stream are still effective for high-resolution event data. Therefore, it is natural to raise the following open question: *Can we transfer knowledge from multi-modal or multi-view data during the training phase and achieve robust tracking only using the event data during the testing phase?*

In this work, we propose a novel event-based visual tracking framework by designing a new cross-modality hierarchical knowledge distillation scheme. As shown in Fig. 2, we first train a **teacher** Transformer network by feeding the RGB frame and event stream. It crops the template patch and search region of dual-modality from the initialized and subsequent frames respectively and adopts a projection layer to transform them into token representations. Then, a couple of Transformer blocks are used to fuse the tokens as a unified backbone. Finally, the tracking head is adopted to predict the response maps for target localization. Once we obtain the teacher Transformer network, the hierarchical knowledge distillation strategy is conducted to guide the learning of the **student** Transformer network, which is fed only with event data. To be specific, the similarity matrix, feature representation, and response maps based knowledge distillation are simultaneously considered for cross-modality knowledge transfer. Note that, since only the event data are fed into the student network, it can achieve not only accurate but also low-latency and high-speed object tracking in the testing stage.

In particular, in addition to evaluating our tracker on existing event-based tracking datasets, we also propose a new first high-resolution event-based tracking dataset, termed **EventVOT**, to fully validate the effectiveness of our method as well as other related works. Different from existing datasets with limited resolution (e.g., FE240hz, VisEvent, COESOT are 346×260) as shown in Fig. 1 (a), our videos are collected by using the Prophesee camera EVK4–HD which outputs event stream in 1280×720. It contains 1141 videos and covers a wide range of target objects, including pedestrians, vehicles, UAVs, ping pong, etc. To build a comprehensive benchmark dataset, we provide the tracking results of multiple baseline trackers for future works to compare. We hope our newly proposed EventVOT dataset can open up new possibilities for event tracking research.

To sum up, our contributions can be concluded as the following three aspects:

- We propose a novel hierarchical cross-modality knowledge distillation approach for event-based tracking problem. To our knowledge, it is the first work to exploit the knowledge transfer from multi-modal

(RGB-Event) / multi-view (Event Image-Voxel) to an unimodal event-based tracker, termed HDETrack.

- We propose the first high-resolution benchmark dataset for event-based tracking, termed EventVOT. We also provide experimental evaluations of recent strong trackers to build a comprehensive event-based tracking benchmark.
- Extensive experiments on four large-scale benchmark datasets, i.e., FE240hz, VisEvent, COESOT, and EventVOT, fully validate the effectiveness of our proposed tracker.

2. Related Work

RGB Camera based Tracking. The mainstream visual trackers are developed based on RGB videos and boosted by deep learning techniques in recent years. The convolutional neural networks are first adopted for feature extraction and learning. Specifically, the MDNet series [22] extract the deep features using three convolutional layers and learn domain-specific layers for tracking. Xu et al. [31] proposes a spatial-time discrimination model based on affine subspace for visual object tracking. The SiamFC [1] and SINT [26] first utilize the Siamese fully convolutional neural networks and Siamese instance matching for tracking, respectively. Besides, A topology-aware universal adversarial attack method against 3D object tracking is proposed by [9]. Later, the Siamese network based trackers become the mainstream gradually and many representative trackers are proposed, like SiamRPN++ [20], SiamMask [29], SiamBAN [8], Ocean [37], LTM [38], ATOM [11], DiMP [2], PrDiMP [12], etc.

Inspired by the success of self-attention and Transformer networks in natural language processing, some researchers also exploit Transformers for visual object tracking [4, 7, 15, 21, 25, 28, 33, 35]. For example, Wang et al. [28] proposed TrDiMP, which integrates Transformer with tracking tasks, exploits temporal context for robust visual tracking. Chen et al. [7] proposed TransT, a novel attention-based feature fusion network and a Siamese structured tracking approach that integrates a fusion network have been designed using Transformer. Other works like ToMP [21] proposed a Transformer-based model prediction module, enabling it to learn more powerful target prediction capabilities, due to the powerful inductive bias of Transformer in capturing global relationships. Gao et al. [15] proposed AiATrack that introduce a universal feature extraction and information propagation module based on Transformer. A simplified tracking architecture called SimTrack [4] has been proposed by Chen et al. which utilize the Transformer as backbone for joint feature extraction and interaction. Ye et al. [33] propose OSTrack, they design a one-stream tracking framework to replace the complex dual-stream framework. Zhang et al. [35] combine spiking neural networks with Transformer for event-based tracking. CEUTrack [25] is proposed by Tang et al., who explore a Transformer-based dual-modal framework for RGB-Event tracking. Different from these works, we exploit event cameras to achieve reliable tracking even under challenging scenarios, like low illumination and fast motion.

Event Camera based Tracking. Tracking based on event cameras is a newly arising research topic and has drawn more and more attention in recent years. Specifically, early event-based trackers ESVM (event-guided support vector machine) [18] is proposed by Huang et al. for high-speed moving object tracking. AFNet [36] proposed by Zhang et al. incorporates event-guided cross-modality alignment and cross-correlation fusion module, which effectively aligns and fuses RGB and event streams. Chen et al. [5] propose an Adaptive Time-Surface with Linear Time Decay (ATSLTD) event-to-frame conversion algorithm for asynchronous retinal event-based tracking. EKLT [17] fuse the frame and event streams to track visual features with high temporal resolution. Zhang et al. [34] adopt self- and cross-domain attention schemes to enhance the RGB and event features for robust tracking. STNet [35] is proposed to capture the global spatial information and temporal cues by using Transformer and spiking neural network (SNN). Wang et al. [30] fuse the RGB and event data using cross-modality Transformer module. Zhu et al. [41] sample the key-events using a density-insensitive downsampling strategy and embed them into high-dimensional feature space for tracking. Tang et al. [25] conduct RGB-Event tracking through a unified backbone network to simultaneously realize multi-modal feature extraction, correlation, and fusion. Zhu et al. [40] introduce prompt tuning to drive the pre-trained RGB backbone for multi-modal tracking. AFNet [36] is proposed to combine both modalities at different measurement rates by using multi-modality alignment and fusion modules. Zhu et al. [42] randomly masks tokens of a specific modality and proposes an orthogonal high-rank loss function to enforce the interaction between different modalities. Different from existing works, we propose to conduct knowledge distill from multi-modal or multi-view in the training phase and only utilize the event data for efficient and low-latency tracking.

Knowledge Distillation. Learning a student network using knowledge distillation for efficient and accurate inference is widely studied. Deng et al. [13] provide explicit feature-level supervision for the learning of event stream by using knowledge distilled from the image domain. For the tracking task, Shen et al. [23] propose to distill large Siamese trackers using a teacher-students knowledge distillation model for small, fast, and accurate trackers. Chen et al. [6] attempt to learn a lightweight student correlation filter-based tracker by distilling a pre-trained deep convo-

lutional neural network. Zhuang et al. [43] introduce Ensemble learning (EL) into the Siamese tracking framework and treat two Siamese networks as students and enabling them to learn collaboratively. Sun et al. [24] conduct cross-modal distillation for TIR tracking from RGB modality on unlabeled paired RGB-TIR data. Wang et al. [27] distill the CNN model pre-trained from the image classification dataset into a lightweight student network for fast correlation filter trackers. Zhao et al. [39] propose a distillation-ensemble-selection framework to address the conflict between the tracking efficiency and model complexity. Ge et al. [16] propose channel distillation for correlation filter trackers which can accurately mine better channels and alleviate the influence of noisy channels. Different from these works, our proposed hierarchical knowledge distillation enables message propagation from multi-modality or multi-view to event-tracking networks.

3. Methodology

3.1. Overview

To achieve efficient and low-latency visual tracking, in this paper, we exploit tracking using an event camera only. To ensure its tracking performance, we resort to the knowledge distillation (KD) from multi-modal or multi-view data. Therefore, we first train a large-scale teacher Transformer using the RGB frames and event stream, as shown in Fig. 2. To be specific, the template patch and search patch of dual modalities are extracted and transformed into tokens by using the projection layer. These tokens are directly concatenated and fed into a unified Transformer backbone network for simultaneous feature extraction, interactive learning, and fusion. For the event student tracking network, we take the event images or voxels as the input and optimize the parameters based on tracking loss function and knowledge distillation (KD) functions. More in detail, the similarity matrix based KD, feature based KD, and response map based KD are considered for a higher tracking performance. We will introduce the more details about the network architecture and hierarchical knowledge distillation strategies in the following subsections.

3.2. Input Representation

In this work, we denote the RGB frames as $\mathcal{I} = \{I_1, I_2, ..., I_N\}$, where I_i denotes each video frame and N is the number of video frames. We treat event stream as $\mathcal{E} = \{e_1, e_2, ..., e_M\}$, where e_j denotes each event point asynchronously launched and M is the number of event points for the input data.

For the video frames $\mathcal{I}$, we utilize the standard processing approach for Siamese tracking and extract the template patch T_I and search patch S_I as the input. For the event stream $\mathcal{E}$, we stack/split them into event images/voxels which can fuse more conveniently with existing RGB modality. More in detail, the event images are obtained by aligning with the exposure time of RGB modality. Event voxels are obtained by splitting the event stream along with the spatial (width W and height H) and temporal dimensions (T_i). The scale of each voxel grid is denoted as (a, b, c), thus, we can get $\frac{W}{a} \times \frac{H}{b} \times \frac{T_i}{c}$ voxel grids. Similarly, we can obtain the template and search regions of event data, i.e., T_E and S_E.

3.3. Network Architecture

We propose a novel hierarchical knowledge distillation framework for event-based tracking. As shown in Fig. 2, it primarily consists of the Multi-modal/Multi-view Teacher Transformer and Unimodal Student Transformer network.

Multi-modal/multi-view Teacher Tracker. We feed the RGB frame and event stream or different event data (e.g., event image and voxel) into the teacher Transformer network. The template and search patches of both modalities/views are concatenated and fed into a projection layer for feature embedding. Following the unified backbone based trackers [25, 33], we propose a teacher network consisting of Transformer layers for multi-modal feature learning and fusing. Then, the tokens corresponding to the search region are selected for target object localization by using the tracking head.

Unimodal Student Tracker. To achieve efficient and low-latency visual tracking, we don't conduct tracking using multi-modal data. A lightweight student Transformer based tracker is proposed, as shown in Fig. 2. Note that, only event data is fed into the student Transformer for tracking. Due to the influence of challenging factors of event-based tracking, such as sparse event points and clutter background, we introduce a hierarchical knowledge distillation strategy to enhance its tracking performance.

3.4. Hierarchical Knowledge Distillation

The tracking loss functions used in OSTrack [33] (i.e., focal loss $\mathcal{L}_{focal}$, $\mathcal{L}_1$ loss, and GIoU loss $\mathcal{L}_{GIoU}$) and three knowledge distillation functions are used to optimize our visual tracker. Generally speaking, the overall loss can be denoted as:

$$\begin{aligned}\mathcal{L}_{total} = \lambda_1\mathcal{L}_{focal} + \lambda_2\mathcal{L}_1 + \lambda_3\mathcal{L}_{GIoU} + \\ \eta_1\mathcal{L}_{simKD} + \eta_2\mathcal{L}_{featKD} + \eta_3\mathcal{L}_{resKD}\end{aligned} \quad (1)$$

For the first three loss functions for tracking, we refer the readers to check OSTrack [33] for better understanding. In the following paragraphs, we will describe the hierarchical knowledge distillation loss functions in detail.

Similarity Matrix based Distillation. The similarity matrix computed in the multi-head self-attention layers incorporates abundant long-range and cross-modal relation in-

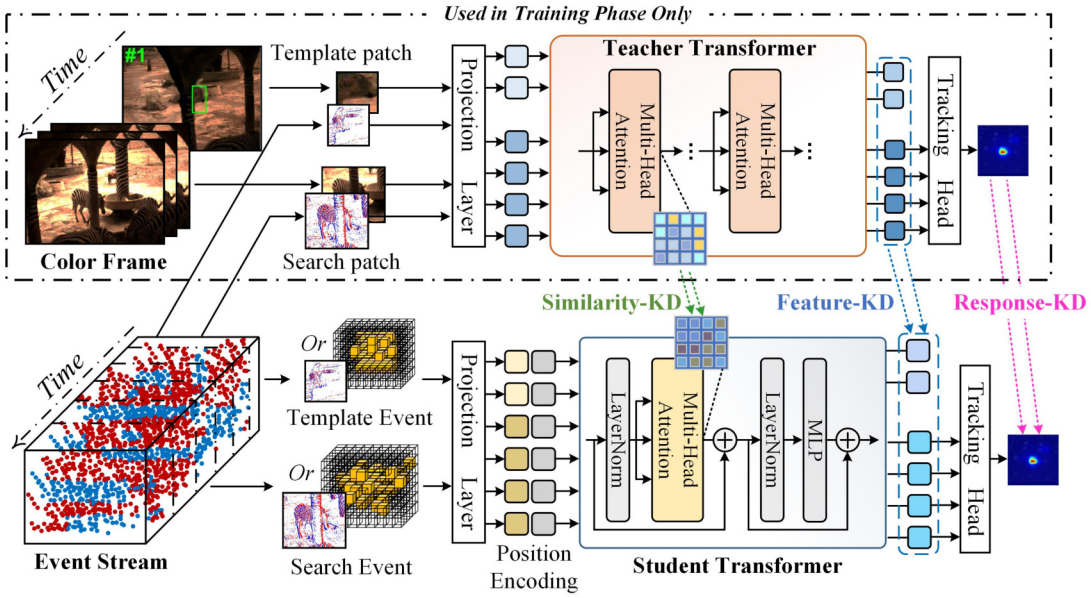

Figure 2. **An overview of our proposed Hierarchical Knowledge Distillation Framework for Event Stream based Tracking, termed HDETrack.** It contains the teacher and student Transformer network which takes multi-modal/multi-view data and event data only as the input respectively. The two networks share the same architecture, i.e., tracking using a unified Transformer backbone network similar to OSTrack [33] and CEUTrack [25]. Our tracker achieves a better tradeoff between accuracy and model complexity, as shown in Fig. 1(b).

formation. In this work, we exploit the knowledge transfer from the similarity matrix learned by the teacher Transformer to the student Transformer. Specifically, we denote the similarity matrix of the i^{th} teacher Transformer layer as $S_t^i \in \mathbb{R}^{640\times640}$. The similarity matrix of the j^{th} student Transformer is denoted as $S_s^j \in \mathbb{R}^{320\times320}$. We repeat the S_s^j to make it have the same dimension as S_t^i. In addition to tracking loss functions, the learning of similarity matrix S_s^j also depends on distilling loss $\mathcal{L}_{simKD}$ as follows,

$$\mathcal{L}_{simKD} = \mathcal{L}_2(S_s^j, S_t^i). \tag{2}$$

Feature based Distillation. The feature distillation from the robust and powerful teacher Transformer network is the second strategy. We denote the token representation of the teacher and student network as F_t and F_s. Then, the distilling loss between them can be represented as,

$$\mathcal{L}_{featKD} = \|F_t - F_s\|_F^2 \tag{3}$$

Response based Distillation. The response maps output from tracking networks are used for target object localization. Obviously, if we can directly mimic this response map R_t, the obtained tracking results will be better. In this paper, the weighted focal loss function [19] is adopted to achieve this target. We denote the ground truth target center and the corresponding low-resolution equivalent as $\hat{p}$ and $\bar{p} = [\bar{p}_x, \bar{p}_y]$, respectively. The Gaussian kernel is used to generate the ground truth heatmap $\hat{\mathbf{P}}_{xy} = exp(-\frac{(x-\bar{p}_x)^2+(y-\bar{p}_y)^2}{2\delta_p^2})$, where δ denotes the object size-adaptive standard deviation [19]. Thus, the Gaussian Weighted Focal (GWF) loss function is formulated as:

$$\mathcal{L}_{GWF} = -\sum_{xy} \begin{cases} (1-\mathbf{P}_{xy})^{\alpha} log(\mathbf{P}_{xy}), & if\ \hat{\mathbf{P}}_{xy} = 1 \\ (1-\hat{\mathbf{P}}_{xy})^{\beta}(\mathbf{P}_{xy})^{\alpha} log(1-\mathbf{P}_{xy}), & otherwise \end{cases} \tag{4}$$

where α and β are two hyper-parameters and which are set to 2 and 4 respectively in our experiments, as suggested in OSTrack [33]. In our implementation, we normalize the response maps of both the teacher and student networks by dividing them via the temperature coefficient τ (empirically set to 2), followed by inputting them into the focal loss for response distillation, i.e., $\mathcal{L}_{resKD} = \mathcal{L}_{GWF}(R_s/\tau, R_t/\tau)$.

4. EventVOT Dataset

4.1. Criteria for Collection and Annotation

To construct a dataset with a diverse range of target categories, as shown in Fig. 4, capable of reflecting the distinct features and advantages of event tracking, this paper primarily considers the following aspects during data collection. *1). Diversity of target categories:* Many common and meaningful target objects are considered, including UAVs, pedestrians, vehicles, ball sports, etc. *2). Diversity of data collection environments:* The videos in our dataset are recorded in day and night time, and involved venue information includes playgrounds, indoor sports arenas, main

Table 1. Event camera based datasets for visual tracking. # denotes the number of corresponding items.

Datasets	Year	#Videos	#Frames	#Class	#Att	#Resolution	Aim	Absent	Frame	Reality	Public
VOT-DVS	2016	60	-	-	-	240 × 180	Eval	✗	✗	✗	✓
TD-DVS	2016	77	-	-	-	240 × 180	Eval	✗	✗	✗	✓
Ulster	2016	1	9,000	-	-	240 × 180	Eval	✗	✗	✓	✗
EED	2018	7	234	-	-	240 × 180	Eval	✗	✗	✓	✓
FE108	2021	108	208,672	21	4	346 × 260	Train & Eval	✗	✗	✓	✓
VisEvent	2021	820	371,127	-	**17**	346 × 260	Train & Eval	✓	✓	✓	✓
COESOT	2022	**1354**	478,721	**90**	**17**	346 × 260	Train & Eval	✓	✓	✓	✓
EventVOT	2023	1141	**569,359**	19	14	**1280 × 720**	Train & Eval	✓	✗	✓	✓

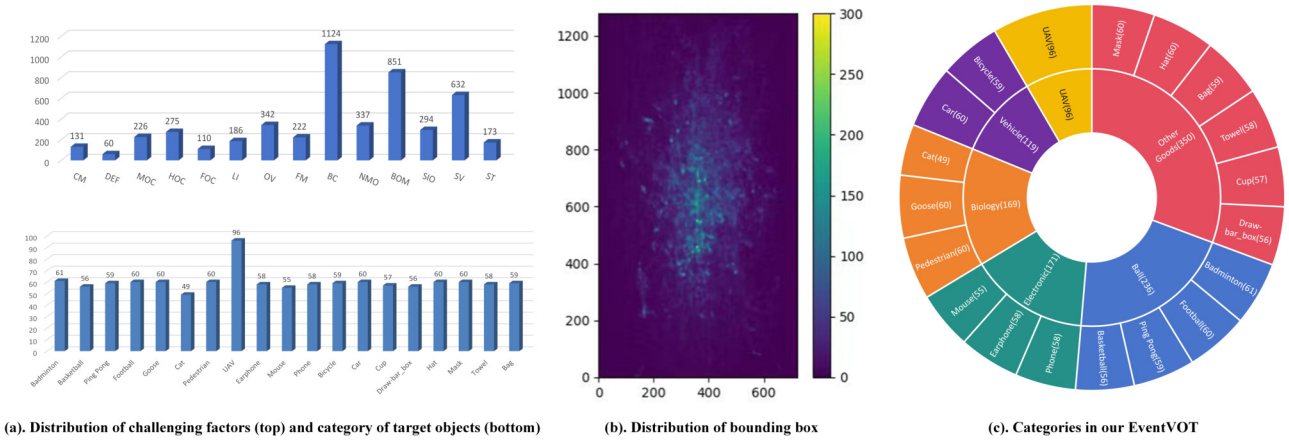

Figure 3. Distribution visualization of challenging factors, category of the target object, and bounding box.

streets and roads, cafeteria, dormitory, etc. *3). Recorded specifically for event camera characteristics:* Different motion speeds, such as high-speed, low-speed, momentary stillness, and varying light intensity, etc. 14 challenging factors are reflected by our EventVOT dataset. *4). High-definition, wide-field event signals:* The videos are collected using a Prophesee EVK4–HD event camera, which outputs event stream with 1280 × 720. This high-definition event camera excels in supporting pure event-based object tracking, thereby avoiding the influences of the RGB cameras and showcasing its features and advantages in various aspects such as high-speed, low-light, low-latency, and low-power consumption. *5). Data annotation quality:* All data samples are annotated by a professional data annotation company and has undergone multiple rounds of quality checks and iterations to ensure the accuracy of the annotations. For each event stream, we first stack into a fixed number (499 in our case) of event images for annotation. *6). Data size:* Collect a sufficiently large dataset to train and evaluate robust event-based trackers. A comparison between the newly proposed dataset and existing tracking datasets is summarized in Table 1.

4.2. Statistical Analysis

In the EventVOT dataset, we have defined 14 challenging factors, involving 19 classes of target objects. The number of videos corresponding to these attributes and categories is visualized in Fig. 3 (a, c). We can find that BC, BOM, and SV are top-3 major challenges which demonstrates that our dataset is relatively challenging. The balance between different categories is also well-maintained, with the number of samples roughly distributed between 50 to 60. Among them, UAVs (Unmanned Aerial Vehicles) are a special category of targets, with a total count of 96. The distribution of the center points of the annotated bounding boxes is visualized in Fig. 3 (b).

4.3. Benchmarked Trackers

To construct a comprehensive benchmark dataset for event-based visual tracking, we consider the following visual trackers: **1). Siamese or Discriminate trackers:** DiMP50 [2], PrDiMP [12], KYS [3], ATOM [11], **2). Transformer trackers:** OSTrack [33], TransT [7], SimTrack [4], AiATrack [15], STARK [32], ToMP50 [21], MixFormer [10], TrDiMP [28]. Note that, we re-train these trackers using their default settings on our training dataset, instead of directly testing on the testing subset. Our EventVOT dataset is split into training/validation/testing

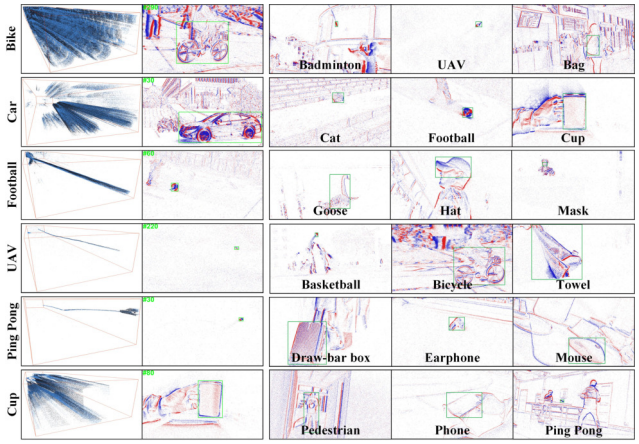

Figure 4. Representative samples of our proposed EventVOT dataset. The 1^{th} column is the 3D event point stream and the 2^{th} columns are sampled event images. 3^{th}-5^{th} columns are more samples of our EventVOT dataset.

subset which contains 841, 18, and 282 videos, respectively. We believe that these retrained tracking algorithms can play a crucial role in future comparisons of their performance.

5. Experiment

5.1. Dataset, Metric, Implementation Details

In addition to our newly proposed **EventVOT** dataset, we also compare our tracker with other SOTA visual trackers on existing event-based tracking datasets, including **FE240hz** [34], **VisEvent** [30], and **COESOT** [25] dataset.

For the evaluation metrics, we adopt the widely used Precision Rate (PR), Normalized Precision Rate (NPR), and Success Rate (SR). The efficiency is also an important metric for a practical tracker, in this work, we adopt FPS (Frames Per Second) to measure the speed of each tracker. More details of datasets, evaluation metrics, and implementation details can be found in our *supplementary material*.

5.2. Comparison on Public Benchmarks

As shown in Table 2, we re-train and report multiple SOTA trackers on the EventVOT dataset. We can find that our baseline tracker OSTrack achieves $55.4, 60.4, 71.1$ on the SR, PR, and NPR, respectively. When adopting our proposed hierarchical knowledge distillation framework in the training phase, these results can be improved to $57.8, 62.2, 73.5$ which fully validates the effectiveness of our proposed method for event-based tracking. Our results are also better than other SOTA trackers, including the Siamese trackers and Transformer trackers (STARK, MixFormer, PrDiMP, etc.). These experimental results fully demonstrate the effectiveness of our proposed hierarchical knowledge distillation from multi-modal to event-based tracking networks. Similar conclusions can also be drawn from the experimental results on FE240hz (Table 3), VisEvent (Table 4), and COESOT (Table 5).

Table 2. Overall tracking performance on EventVOT dataset.

Trackers	Source	SR	PR	NPR	Params	FPS
Ours	–	57.8	62.2	73.5	92.1	105
TrDiMP	CVPR21	39.9	34.8	48.7	26.3	26
ToMP50	CVPR22	37.6	32.8	47.4	26.1	25
OSTrack	ECCV22	55.4	60.4	71.1	92.1	105
AiATrack	ECCV22	57.4	59.7	72.8	15.8	38
STARK	ICCV21	44.5	39.6	55.7	28.1	42
TransT	CVPR21	54.3	56.5	68.8	18.5	50
DiMP50	ICCV19	52.6	51.1	67.2	26.1	43
PrDiMP	CVPR20	55.5	57.2	70.4	26.1	30
KYS	ECCV20	38.7	37.3	49.8	–	20
MixFormer	CVPR22	49.9	49.6	63.0	35.6	25
ATOM	CVPR19	44.4	44.0	57.5	8.4	30
SimTrack	ECCV22	55.4	57.5	69.9	57.8	40

Table 3. Experimental results (SR/PR) on FE240hz dataset.

STNet	TransT	STARK	PrDiMP	EFE	SiamFC++
58.5/89.6	56.7/89.0	55.4/83.7	55.2/86.8	55.0/83.5	54.5/85.3
DiMP	**ATOM**	**Ocean**	**SiamPRN**	**OSTrack**	**Ours**
53.4/88.2	52.8/80.0	50.2/76.4	41.6/75.5	57.1/89.3	59.8/92.2

Table 4. Results on VisEvent dataset. EF and MF are short for early fusion and middle-level feature fusion.

	Trackers	SR	PR	NPR
RGB + Event Input	**CEUTrack**	64.89	69.06	73.81
	LTMU (EF)	60.10	66.76	69.78
	PrDiMP (EF)	57.20	64.47	67.02
	CMT-MDNet (MF)	57.44	67.20	69.78
	ATOM (EF)	53.26	60.45	63.41
	SiamRPN++ (EF)	54.11	60.58	64.72
	SiamCAR (EF)	52.66	58.86	62.99
	Ocean (EF)	43.56	52.02	54.21
	SuperDiMP (EF)	36.21	46.99	42.84
Event Input	**STNet (Event-Only)**	39.7	49.2	-
	TransT (Event-Only)	39.5	47.1	-
	STARK (Event-Only)	34.8	41.8	-
	OSTrack (Event-Only)	34.5	50.1	41.6
	Ours (Event-Only)	37.3	54.6	44.5

5.3. Ablation Study

Analysis on Hierarchical Knowledge Distillation. In this section, we will isolate each distillation strategy for individual evaluation to assess its impact on the final tracking performance. On the COESOT dataset, we take the RGB and event image as the input of teacher network and feed the event data only into the student tracker. For the EventVOT dataset, we stack the event stream into images and voxels and conduct hierarchical knowledge distillation based on multi-view settings. As shown in Table 6, the base denotes the tracker which is trained using three tracking loss functions only as the same as methods OSTrack

Table 5. Overall tracking performance on COESOT dataset.

Trackers	Source	SR	PR	NPR
Ours	-	53.1	64.1	64.5
TrDiMP	CVPR21	50.7	59.2	58.4
ToMP50	CVPR22	46.3	55.2	56.0
OSTrack	ECCV22	50.9	61.8	61.5
AiATrack	ECCV22	50.6	59.5	59.2
STARK	ICCV21	40.8	44.5	46.1
TransT	CVPR21	45.6	54.3	54.2
DiMP50	ICCV19	53.8	64.8	65.1
PrDiMP	CVPR20	47.5	57.8	57.9
KYS	ECCV20	42.6	52.7	52.1
MixFormer	CVPR22	44.4	50.2	51.1
ATOM	CVPR19	42.1	50.4	51.3
SimTrack	ECCV22	48.3	55.7	56.6

Table 6. Component Analysis results (PR/SR) on COESOT and EventVOT dataset.

No.	Base	SKD	FKD	RKD	COESOT	EventVOT
1	✓				61.8/50.9	60.4/55.4
2	✓	✓			62.5/51.6	60.8/56.5
3	✓		✓		63.0/52.1	60.4/56.4
4	✓			✓	62.3/51.5	60.6/56.2
5	✓	✓	✓		63.3/52.2	61.3/57.2
6	✓	✓		✓	63.2/52.1	62.2/57.5
7	✓		✓	✓	63.3/52.3	62.1/57.6
5	✓	✓	✓	✓	64.1/53.1	62.2/57.8

and CEUTrack do. We can note that it achieves $61.8/50.9$, and $60.4/55.4$ on the COESOT and EventVOT datasets, respectively. When introducing new distillation loss like similarity-based, feature-based, and response-based distillation functions, the results are all improved in both settings. Note that, the feature-based distillation works better on the COESOT in contrast to the EventVOT dataset. When all these distillation strategies are used, better tracking performance can be obtained on multi-modal and multi-view settings. On the basis of all these experiments, we can draw the conclusion that all the proposed hierarchical knowledge distillation strategies can contribute to event-based tracking.

Analysis on Tracking in Specific Challenging Environment. In this work, our proposed EventVOT dataset reflects 14 core challenging factors in the tracking task. As shown in Fig. 5, we report the results of our tracker and other state-of-the-art trackers under each challenging scenario. We can note that our proposed tracker achieves better performance when facing attributes like DEF (Deformation), CM (Camera motion), SIO (Similar interferential object) and BC (Background clutter), etc. We also achieve similar tracking results in other attributes which demonstrate that our proposed hierarchical knowledge distillation strategy works well for transferring knowledge from multi-modal/multi-view data to event-based tracker.

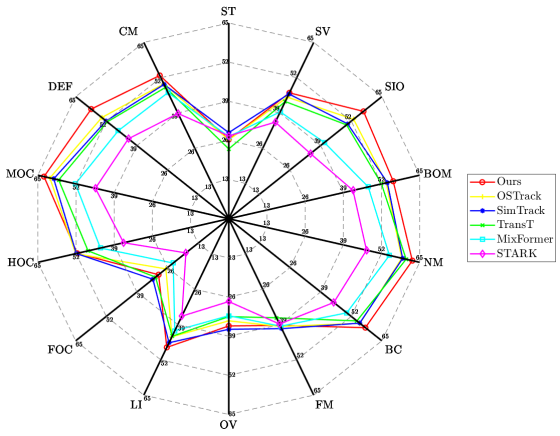

Figure 5. Tracking results (SR) under each challenging factor.

Table 7. Ablation studies on event representation on EventVOT.

Input Data	SR	PR	NPR
1. Event Frames	57.8	62.2	73.5
2. Event Voxels	8.6	7.5	10.3
3. Event Time Surface	53.3	55.1	68.7
4. Event Reconstruction Images	54.5	60.5	69.2

Analysis on Different Event Representations. In this part, we conduct tracking with multiple representations of event data and analyze the influences of different event representations. Specifically, the event image, event voxel, and time surface are considered, as shown in Table 7. We can observe that the event voxel based tracking performs worse than others on our high-resolution event stream. We believe this may be attributed to the necessity of a meticulous design for the feature representation of voxels.

6. Conclusion

In this paper, we propose a novel hierarchical knowledge distillation framework for event-based tracking. It formulates the learning of event trackers based on the teacher-student knowledge distillation framework. The teacher network takes the multi-modal or multi-view data as the input while the student network takes the event data for tracking. In the distillation phase, it simultaneously considers similarity-based, feature-based, and response-based knowledge distillation. To bridge the data gap, in this work, we also propose the first large-scale, high-resolution event-based tracking dataset, termed EventVOT. Extensive experiments on multiple datasets fully validated the effectiveness of our proposed hierarchical knowledge distillation strategy. In our future work, we will consider collecting more high-resolution event videos and pre-train a strong event-based tracker in a self-supervised learning manner.

References

[1] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. Fully-convolutional siamese networks for object tracking. In *European Conference on Computer Vision*, page 850–865, 2016.

[2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, page 6182–6191, 2019.

[3] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Know your surroundings: Exploiting scene information for object tracking. In *European Conference on Computer Vision*, page 205–221, 2020.

[4] Boyu Chen, Peixia Li, Lei Bai, Lei Qiao, Qiuhong Shen, Bo Li, Weihao Gan, Wei Wu, and Wanli Ouyang. Backbone is all your need: A simplified architecture for visual object tracking. In *European Conference on Computer Vision*, page 375–392, 2021.

[5] Haosheng Chen, Qiangqiang Wu, Yanjie Liang, Xinbo Gao, and Hanzi Wang. Asynchronous tracking-by-detection on adaptive time surfaces for event-based object tracking. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 473–481, 2019.

[6] Qihuang Chen, Bineng Zhong, Qihua Liang, Qingyong Deng, and Xianxian Li. Teacher-student knowledge distillation for real-time correlation tracking. *Neurocomputing*, 500:537–546, 2022.

[7] Xin Chen, Jiawen Yan, Bin Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 8126–8135, 2021.

[8] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6668–6677, 2020.

[9] Riran Cheng, Xupeng Wang, Ferdous Sohel, and Hang Lei. Topology-aware universal adversarial attack on 3d object tracking. *Visual Intelligence*, 1:1–12, 2023.

[10] Yutao Cui, Cheng Jiang, Limin Wang, and Wu Gangshan. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 13608–13618, 2022.

[11] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 4660–4669, 2019.

[12] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, page 7183–7192, 2019.

[13] Yongjian Deng, Hao Chen, Huiying Chen, and Youfu Li. Learning from images: A distillation learning framework for event cameras. *IEEE Transactions on Image Processing*, 30: 4919–4931, 2021.

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

[15] Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *European Conference on Computer Vision*, page 146–164, 2022.

[16] Shiming Ge, Zhao Luo, Chunhui Zhang, Yingying Hua, and Dacheng Tao. Distilling channels for efficient deep tracking. *IEEE Transactions on Image Processing*, 29:2610–2621, 2019.

[17] Daniel Gehrig, Henri Rebecq, Guillermo Gallego, and Davide Scaramuzza. Eklt: Asynchronous photometric feature tracking using events and frames. *International Journal of Computer Vision*, 128(3):601–618, 2020.

[18] Jing Huang, Shizheng Wang, Menghan Guo, and Shoushun Chen. Event-guided structured output tracking of fast-moving objects using a celex sensor. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2413–2417, 2018.

[19] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

[20] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, page 8971–8980, 2018.

[21] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 8731–8740, 2022.

[22] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4293–4302, 2016.

[23] Jianbing Shen, Yuanpei Liu, Xingping Dong, Xiankai Lu, Fahad Shahbaz Khan, and Steven Hoi. Distilled siamese networks for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8896–8909, 2021.

[24] Jingxian Sun, Lichao Zhang, Yufei Zha, Abel Gonzalez-Garcia, Peng Zhang, Wei Huang, and Yanning Zhang. Unsupervised cross-modal distillation for thermal infrared tracking. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2262–2270, 2021.

[25] Chuanming Tang, Xiao Wang, Ju Huang, Bo Jiang, Lin Zhu, Jianlin Zhang, Yaowei Wang, and Yonghong Tian. Revisiting color-event based tracking: A unified network, dataset, and metric. *arXiv preprint arXiv:2211.11010*, 2022.

[26] Ran Tao, Efstratios Gavves, and Arnold W. M. Smeulders. Siamese instance search for tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1420–1429, 2016.

[27] Ning Wang, Wengang Zhou, Yibing Song, Chao Ma, and Houqiang Li. Real-time correlation tracking via joint model compression and transfer. *IEEE Transactions on Image Processing*, 29:6123–6135, 2020.

[28] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 1571–1580, 2021.

[29] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H.S.Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1328–1338, 2019.

[30] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. Visevent: Reliable object tracking via collaboration of frame and event flows. *arXiv preprint arXiv:2108.05015*, 2021.

[31] Tianyang Xu, Xuefeng Zhu, and Xiaojun Wu. Learning spatio-temporal discriminative model for affine subspace based visual object tracking. *Visual Intelligence*, 1:1–13, 2023.

[32] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, page 10448–10457, 2021.

[33] Botao Ye, Hong Chang, Bingpeng Ma, and Shiguang Shan. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European Conference on Computer Vision*, 2022.

[34] Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong. Object tracking by jointly exploiting frame and event domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13043–13052, 2021.

[35] Jiqing Zhang, Bo Dong, Haiwei Zhang, Jianchuan Ding, Felix Heide, Baocai Yin, and Xin Yang. Spiking transformers for event-based single object tracking. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 8801–8810, 2022.

[36] 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*, pages 9781–9790, 2023.

[37] Zhipeng Zhang, Houwen Peng, Jianlong Fu, Bing Li, and Weiming Hu. Ocean: Object-aware anchor-free tracking. In *European Conference on Computer Vision*, page 771–787, 2020.

[38] Zhipeng Zhang, Yihao Liu, Xiao Wang, Bing Li, and Weiming Hu. Learn to match: Automatic matching network design for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, page 13339–13348, 2021.

[39] Shaochuan Zhao, Tianyang Xu, Xiao-Jun Wu, and Josef Kittler. Distillation, ensemble and selection for building a better and faster siamese based tracker. *IEEE transactions on circuits and systems for video technology*, 2022.

[40] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9516–9526, 2023.

[41] Zhiyu Zhu, Junhui Hou, and Xianqiang Lyu. Learning graph-embedded key-event back-tracing for object tracking in event clouds. *Advances in Neural Information Processing Systems*, 35:7462–7476, 2022.

[42] Zhiyu Zhu, Junhui Hou, and Dapeng Oliver Wu. Cross-modal orthogonal high-rank augmentation for rgb-event transformer-trackers. *arXiv preprint arXiv:2307.04129*, 2023.

[43] Junfei Zhuang, Yuan Dong, and Hongliang Bai. Ensemble learning with siamese networks for visual tracking. *Neurocomputing*, 464:497–506, 2021.