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

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

0.1. Dataset

- **FE240hz dataset**: It is collected using a gray-scale DVS346 event camera which contains 71 training videos and 25 testing videos. More than 1132K annotations on more than 143K images and corresponding events are provided. It considers different degraded conditions for tracking, such as motion blur and high dynamic range.
- **VisEvent dataset**: It is the first large-scale frame-event tracking dataset recorded using a color DVS346 event camera. A total of 820 videos are collected in both indoor and outdoor scenarios. Specifically, the authors split these videos into a training subset and a testing subset which contain 500 and 320 videos, respectively. More details can be found on GitHub https://github.com/wangxiao5791509/VisEvent_SOT_Benchmark.
- **COESOT dataset**: It is a category-wide RGB-event-based tracking dataset that contains 90 categories and 1354 video sequences (478,721 RGB frames). 17 challenging factors are formally defined in this dataset. The training and testing subset contains 827 and 527 videos, respectively. Please refer to the following GitHub for more details https://github.com/Event-AHU/COESOT.

Implementation Details The training of our tracker can be divided into two stages. We first pre-train the teacher Transformer with multimodal inputs for 50 epochs. The learning rate is 0.0001, weight decay is 0.0001, and batch size is 32. Then, the hierarchical knowledge distillation strategy is adopted for the training of the student Transformer network. The learning rate, weight decay, and batch size are set as 0.0001, 0.0001, and 32, respectively. The AdamW [1] is selected as the optimizer. Our code is implemented using Python based on PyTorch [2] framework and the experiments are conducted on a server with CPU Intel(R) Xeon(R) Gold 5318Y CPU @2.10GHz and GPU RTX3090.

Table 1. Description of 14 attributes in our EventVOT dataset.

Attributes	Description
01. CM	Abrupt motion of the camera
02. MOC	Mildly occluded
03. HOC	Heavily occluded
04. FOC	Fully occluded
05. DEF	The target is deformable
06. LI	Low illumination
07. OV	The target completely out of view
08. SV	Scale variation
09. BC	Background clutter
10. FM	Fast motion
11. NMO	No motion
12. BOM	Influence of background object motion
13. SIO	Similar interferential object
14. ST	Small target

[*] ✉ Corresponding Author: Bo Jiang

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

0.2. Comparison on Public Benchmarks

Results on FE240hz Dataset. As shown in Table 3, our baseline OSTrack achieves 57.1/89.3 on the SR/PR metric, meanwhile, ours are 59.8/92.2 which is significantly better than the baseline method. Our tracker also beats other SOTA trackers including event-based trackers (e.g., STNet and EFE), and Transformer trackers (like TransT, STARK) by a large margin. These results fully validated the effectiveness of our proposed hierarchical knowledge distillation strategy for event-based tracking.

Results on EventVOT Dataset. 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 validated 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.

Results on VisEvent Dataset. As shown in Table 4, we report the tracking results on the VisEvent dataset and compare them with multiple recent strong trackers. Specifically, our baseline OSTrack [4] achieves $34.5, 50.1, 41.6$ on SR, PR, and NPR, respectively, meanwhile, ours are $37.3, 54.6, 44.5$ on these metrics. These results demonstrate that our proposed hierarchical knowledge distillation strategy can enhance the event-based tracking results by learning from multimodal input data. Compared with other Transformer based trackers, such as the STARK [3], we can find that our results are much stronger than this tracker, with an improvement of +2.5 and +12.8 on SR and PR. We also beat the STNet and TransT on the PR metric, which fully validated the effectiveness of our proposed strategy for event-based tracking.

Results on COESOT Dataset. As shown in Table 5, we report our tracking results on the large-scale RGB-Event tracking dataset COESOT. Note that, the compared baseline methods are re-trained on the training subset of COESOT using their default settings and hyper-parameters to achieve a relatively fair comparison. It is easy to find that our baseline OSTrack achieves $50.9, 61.8, 61.5$ on the SR, PR, and NPR metrics, meanwhile, we obtain $53.1, 64.1, 64.5$ which are significantly better than theirs. Our tracking results are also better than most of the compared trackers, including TransT, AiATrack, MixFormer, etc. These experimental results fully demonstrate the effectiveness of our proposed hierarchical knowledge distillation from multi-modal to event-based tracking networks.

Analysis on Loss Function for Distillation. When conducting knowledge distillation in our training phase, multiple loss functions can be selected, such as L1, L2, MSE (Mean Squared Error), and KLD (Kullback-Leibler Divergence) loss functions. In this part, we test the effectiveness of these loss functions based on feature-level knowledge distillation and report the tracking results on the COESOT dataset. As shown in Table 6, we can find that MSE performs the best achieving 52.1 and 63.0 on SR and PR metric.

Analysis on Number of Transformer Layers. When conducting tracking using student Transformer networks, the accuracy and tracking speed are influenced by the number of Transformer layers. In this part, we set different layers to

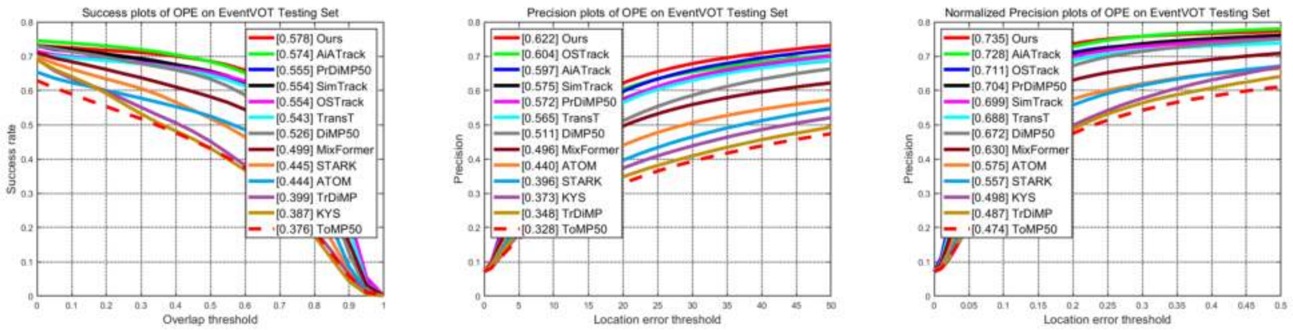

Figure 1. Visualization of tracking results of our proposed EventVOT dataset.

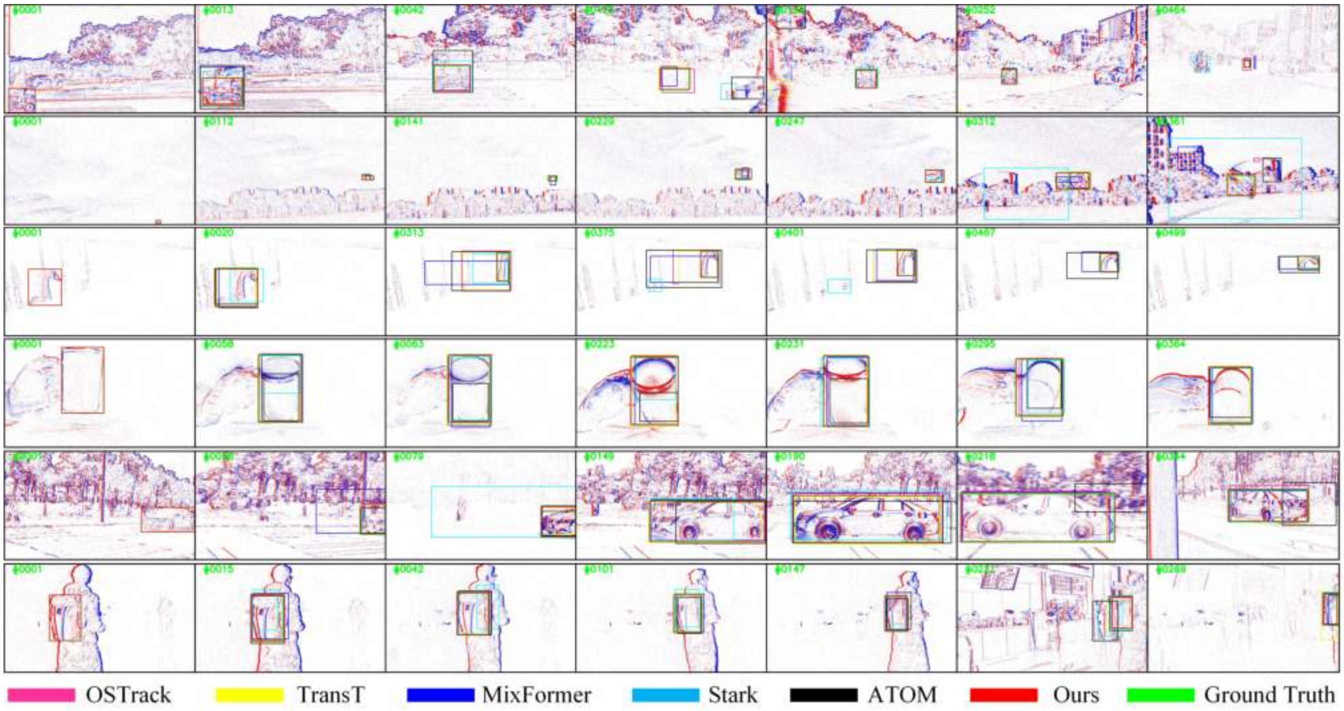

Figure 2. Visualization of the tracking results of ours and other SOTA trackers.

check their influences, i.e., 12, 8, and 4 layers. It is easy to find that more Transformer layers (more learnable parameters) will bring us better tracking results.

Analysis on Align Methods for Distillation. In the training phase, the number of teacher Transformer networks is twice as large as the student network. We also tried different alignment approaches to bridge this gap for knowledge distillation, including repeating, reshaping, and resizing, and adjusting using a fully connected layer. We can find that simple repeat features of the student network perform the best for event-based tracking.

Analysis on Tradeoff Parameters for Distillation Strategies. In the hierarchical knowledge distillation phase, we set different tradeoff parameters to achieve better tracking performance. As feature-level distillation is widely exploited and also performs well on our dataset, therefore, we default set its weight as 1. For the similarity-level and response-level distillation, we experimentally set their weights as equal ones, including 0.50, 0.65, 0.68, 0.70, 0.72, 0.75, and 1. As shown in Fig. 3, better tracking results can be obtained if we set the weights as [1, 0.7, 0.7] which achieves 0.570, 0.610, 0.726 on SR, PR, and NPR, respectively.

0.3. Visualization

In addition to the quantitative analysis mentioned above, we also conducted a visual analysis of the proposed tracking algorithm to provide readers with a better understand-

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. Ablation studies on event representation, loss functions, align methods, and the number of Transformer layers on EventVOT and COESOT dataset.

#(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
#(COESOT). Loss for Feature-level KD	**SR**	**PR**	**NPR**
5. MSE loss	52.1	63.0	62.9
6. L2 loss	51.9	62.8	62.6
7. L1 loss	51.6	61.9	62.0
8. KLD loss	50.2	60.0	59.7
#(COESOT). Align Method for Distillation	**SR**	**PR**	**NPR**
9. Repeat	52.1	63.0	62.9
10. Reshape & Resize	51.4	61.5	61.3
11. FC	50.9	61.1	61.0
#(COESOT). Number of Former Layers	**SR**	**PR**	**NPR**
12. 12 layers	53.1	64.1	64.5
13. 8 layers	49.2	58.9	59.2
14. 4 layers	42.1	47.6	49.4

ing of our tracking framework. As shown in Fig. 2, we visualize the tracking results of ours and other SOTA trackers on the EventVOT dataset, including OSTrack, TransT, MixFormer, STARK, and ATOM. We can find that our tracking using event camera is an interesting and challenging task. These trackers perform well in simple scenarios, however, there is still significant room for improvement. Besides, we also provide the response maps and similarity maps of Transformer networks on our Baseline, student and teacher networks respectively. As shown in Fig. 4, the target object regions are highlighted which means that tracker focuses on the real targets accurately. Clearly, the areas of focus of our student network guided by the teacher network are more accurate than the baseline approach, second only to the performance of the excellent teacher network.

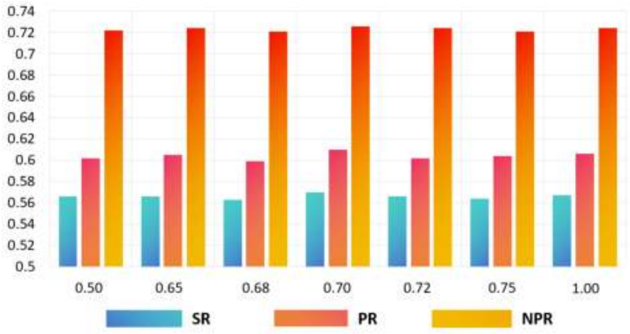

Figure 3. Results with different tradeoff parameters for distillation.

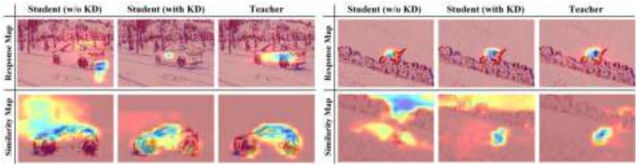

Figure 4. Visualization of the response maps and similarity maps predicted by our Baseline, student and teacher.

References

[1] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[2] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[3] 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.

[4] 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.