Enhancing Visual Tracking by Leveraging High-frequency Information within Event Signals

Yuheng Jiang[1]
jiangyuheng@mail.ustc.edu.cn
Hebei Li[1]
lihebei@mail.ustc.edu.cn
Dachun Kai[1]
dachunkai@mail.ustc.edu.cn
Yansong Peng[1]
pengyansong@mail.ustc.edu.cn
Jiahui Yuan[1]
yuanjiahui@mail.ustc.edu.cn
Peilin Xiao[1]
peilinxiao@mail.ustc.edu.cn
Yueyi Zhang[2]
yueyi.zhang@miromind.ai
Xiaoyan Sun[†1,3]
sunxiaoyan@ustc.edu.cn

[1] MoE Key Laboratory of Brain-Inspired Intelligent Perception and Cognition, University of Science and Technology of China, Hefei, China
[2] MiroMind, Singapore
[3] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China

Abstract

Traditional object trackers struggle in degraded scenarios, lacking sufficient appearance details of moving targets for precise tracking. Recent trackers have integrated high-frequency event signals to assist tracking. However, they neglect the high-temporal-resolution motion information inherent in events, limiting their performance especially in occlusion and background clutter. To address these challenges, we propose HFTrack, a novel tracker designed to fully leverage the spatio-temporal high-frequency information within event signals, thereby enhancing the tracking performance. Specifically, we introduce a frequency-based feature enhancement module, which enriches the frame feature with high-frequency components from events in frequency space, capturing detailed appearance information of moving targets. Additionally, we propose a spatio-temporal information decoder with an auto-regressive temporal query, integrating both historical motion cues from events and enhanced spatial features for robust target localization. Experimental results demonstrate that our HFTrack significantly outperforms existing trackers, showcasing its strong ability to track the target under challenging conditions.

[†]Corresponding author.

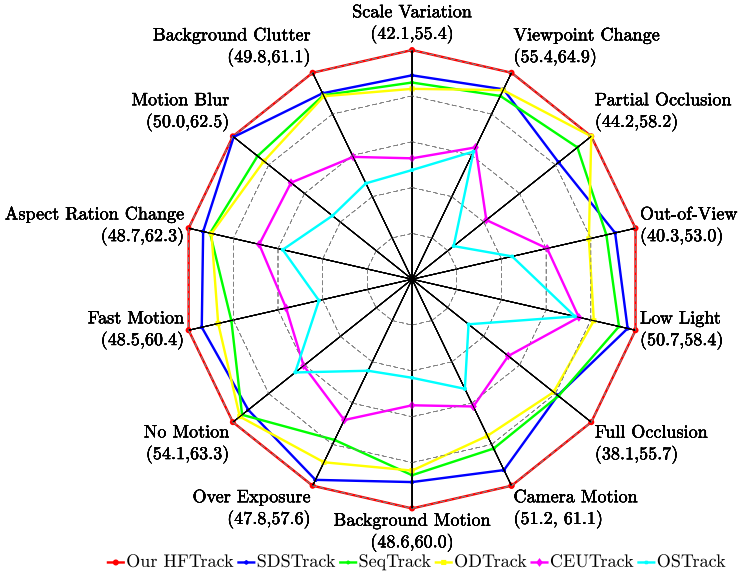

Figure 1: Success rate scores of trackers across different scenes on VisEvent [21]. Our HFTrack outperforms others on multiple challenging attributes.

1 Introduction

Visual Object Tracking (VOT) is a fundamental task in computer vision, which aims to identify and locate the target within a video sequence based on its initial status in the template. It has wide applications in fields such as smart surveillance [23], autonomous driving [5] and drone navigation [25].

Traditional VOT methods [4, 20, 22, 24, 26, 29] rely solely on RGB signals. They struggle in challenging scenarios due to the limited dynamic range and frame rate of RGB cameras [27], lacking sufficient appearance details of targets for precise tracking. Recently, event cameras, as a novel bio-inspired sensor, have been introduced to assist VOT [10, 18, 21, 27, 28, 31]. Compared to RGB signals, event signals provide richer spatio-temporal high-frequency information. In the spatial domain, events are triggered by the edges and textures of moving objects, capturing finer high-frequency appearance details [12, 14]. In the temporal domain, events record the target's motion with high temporal resolution up to 1 MHz [9], making them valuable for understanding and predicting the target's motion state.

Although there has been some progress in RGB-Event (RGB-E) trackers [10, 18, 21, 27, 28, 31], most of them compress event streams into frame-like images, neglecting the high-temporal-resolution information inherent in events. Furthermore, these trackers primarily focus on intra-frame similarity matching, failing to establish inter-frame associations. As a result, they lack the fine-grained motion cues of the target, which limits their ability to comprehend and predict the target's motion pattern, especially in occlusion and background clutter. Additionally, these trackers [10, 18, 21, 27, 28, 31] perform feature fusion only in spatial space. This implicit fusion approach cannot fully leverage the high-frequency appearance details provided by events, as demonstrated in [14].

To address the above challenges, we propose a novel RGB-E tracker named **HFTrack**,

to fully leverage the spatio-temporal **H**igh-**F**requency information within event signals for better tracking performance. In the spatial domain, we propose a frequency-based feature enhancement (FFE) module, which complements the high-frequency information missing in RGB features by incorporating events in frequency space [8], boosting the tracker's ability to capture the appearance details of moving targets. In the temporal domain, we introduce a temporal query to interact with the high-temporal-resolution event features, recording and propagating fine-grained motion cues of the target for tracking guidance. Finally, we design a spatio-temporal information decoder (SID), which integrates the enhanced spatial features and historical motion information, improving the tracker's comprehension of the target's motion pattern. As shown in Figure 1, our HFTrack significantly outperforms other trackers on multiple challenging attributes.

In summary, we make the following contributions: (1) We propose **HFTrack** to fully leverage the spatio-temporal high-frequency information within event signals. In the spatial domain, we propose an FFE module to capture high-frequency appearance details from events. (2) In the temporal domain, we design an SID module with an auto-regressive temporal query to integrate high-temporal-resolution motion cues in events and spatial features for robust tracking performance. (3) Extensive experiments demonstrate that our HFTrack outperforms other state-of-the-art (SOTA) trackers.

2 Related Works

RGB-only-based Visual Object Tracking. Deep-learning-based trackers have dominated VOT. Early CNN-based trackers [1, 6, 16] employed two weight-shared CNN branches to extract features separately from the template and search frame. The feature interaction was achieved through extra relation modeling networks. Inspired by the remarkable performance of Vision Transformer (ViT) [7], some trackers [3, 20, 24] combined CNNs for feature extraction with Transformers for relation modeling. However, this hybrid approach increased the complexity of the tracker. Recent trackers [4, 22, 26, 29] have adopted a single Transformer for both feature extraction and relation modeling, balancing tracking accuracy and inference speed. Despite these advancements, RGB-only-based trackers struggle in degraded scenarios due to the limited dynamic range and frame rate of RGB cameras, resulting in insufficient appearance details of the target for accurate tracking.

RGB-event-joint Visual Object Tracking. Compared to RGB cameras, event cameras offer high-frequency information about edges and textures of moving targets with high temporal resolution [9, 11, 13]. A series of studies have explored integrating RGB and event data for VOT. Wang *et al.* [21] introduced a transformer module to fuse the frame and event features. Zhang *et al.* [28] designed an event-guided alignment framework for high frame rate tracking. Tang *et al.* [18] proposed a unified Transformer-based tracking framework. Zhu *et al.* [31] introduced the orthogonal high-rank augmentation to improve cross-modal interaction. Hou *et al.* [10] used self-distillation symmetric adapter learning for effective multi-modal feature fusion. However, these trackers neglect the high-temporal-resolution motion cues of the target inherent in events. As a result, they struggle to comprehend the object's rapid variation of motion state under challenging conditions. In comparison, our proposed HFTrack integrates fine-grained motion information in events for tracking guidance, achieving more robust tracking performance.

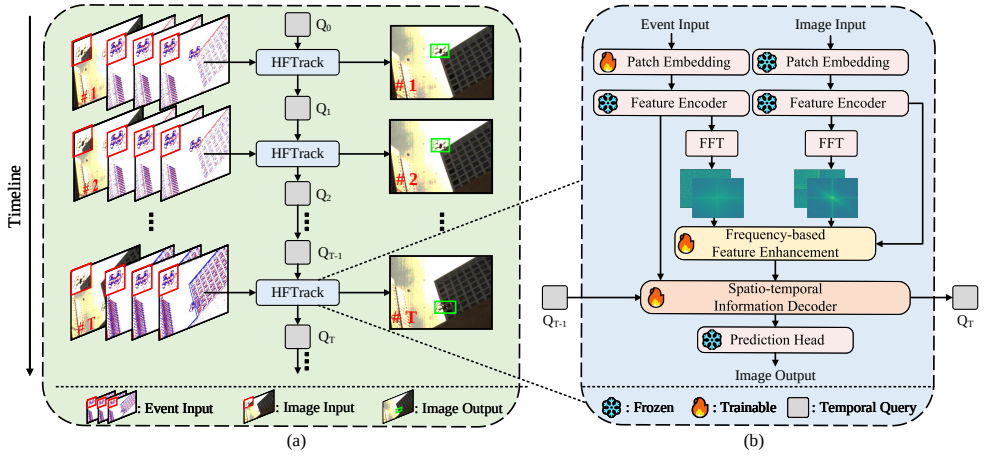

Figure 2: Overview of our HFTrack. (a) The proposed HFTrack's working process, including a temporal query Q to record and convey historical motion cues across adjacent frames. The red rectangles refer to the templates, and the green rectangles indicate the tracking results. Event images here are only used for visualization. (b) The structure of HFTrack's network.

3 Methods

3.1 Overview

We propose a novel RGB-E tracker named HFTrack. As shown in Figure 2, HFTrack operates based on a recurrent architecture, with a temporal query Q propagated across adjacent frames. The network consists of four main components: the pre-trained feature encoder, the FFE module, the SID module, and the prediction head. The RGB template and search region are fed into the feature encoder for feature extraction and relation modeling, with the same process applied to the events. The extracted RGB feature is enhanced by events through the FFE in frequency space. The SID utilizes the enhanced feature of FFE, the event feature, and the temporal query Q_{T-1} from the last frame for target location. Finally, the SID's output is fed into the prediction head for result generation.

To preserve the high-temporal-resolution information in events, we convert the event stream into the voxel grid form [30] with the bin = 3. The frozen parts of our network are the same as [26]. Their weights are pre-trained on diverse datasets and kept fixed during HFTrack's training. The patch embedding layer for events is trainable to provide event-specific tokens.

3.2 Frequency-based Feature Enhancement

Events are triggered by rapid changes in lighting intensity, providing high-frequency information about edges and textures of moving objects [2], which is crucial for precise tracking. However, most RGB-E trackers [10, 18, 21, 27, 28, 31] fuse features only in spatial space. This implicit fusion fails to fully utilize the high-frequency appearance details inherent in events, as illustrated in [14]. In order to leverage these high-frequency details from events more explicitly, we propose the FFE module to complement the high-frequency information

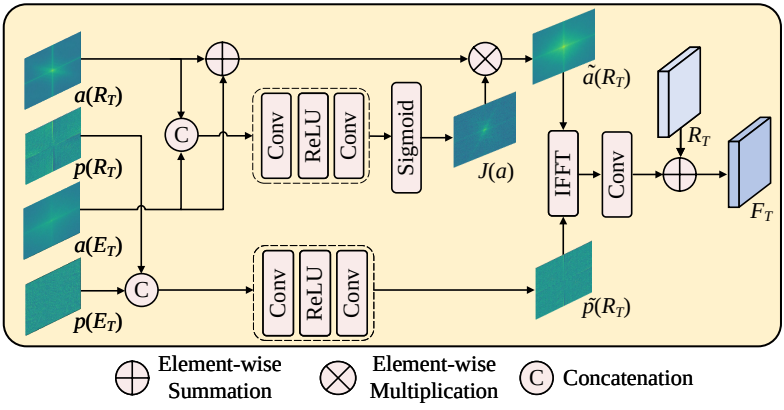

Figure 3: The structure of the proposed Frequency-based Feature Enhancement module. R_T and E_T correspond to the extracted RGB feature and event feature, $a(\cdot)$ and $p(\cdot)$ indicate the amplitude and phase respectively.

lacked in RGB features by events in frequency space, improving the tracker's capability of capturing the moving target's appearance details.

As shown in Figure 3, we first adopt fast Fourier transform (FFT) [8] to the extracted RGB feature R_T and event feature E_T, generating the inputs as:

$$a(R_T), p(R_T) = F(R_T) \quad a(E_T), p(E_T) = F(E_T), \tag{1}$$

where $F(\cdot)$ indicates FFT [8], $a(\cdot)$ and $p(\cdot)$ correspond to the amplitude and phase.

From the perspective of amplitude, the event feature contains richer high-frequency information. In comparison, the RGB feature presents redundant amplitude components. Therefore, we generate a frequency filter $J(a)$ to emphasize the high-frequency appearance details and remove the unwanted amplitude components in the RGB feature as:

$$J(a) = \mathcal{S}(\mathcal{V}(\mathcal{R}(\mathcal{V}[a(R_T); a(E_T)]))), \tag{2}$$

where $[\cdot]$ denotes channel-wise concatenation, $\mathcal{V}(\cdot)$ is the convolution operation, $\mathcal{S}(\cdot)$ and $\mathcal{R}(\cdot)$ correspond to the sigmoid and ReLU function separately.

Once gained the frequency filter $J(a)$, we get the filtered amplitude $\widetilde{a}(R_T)$ as:

$$\widetilde{a}(R_T) = J(a) \otimes (a(R_T) + a(E_T)), \tag{3}$$

where $\otimes$ denotes the element-wise multiplication.

From the perspective of phase, we use the phase of the event feature to suppress the undesired phase components in the RGB feature, generating $\widetilde{p}(R_T)$ as follows:

$$\widetilde{p}(R_T) = \mathcal{V}(\mathcal{R}(\mathcal{V}[p(R_T); p(E_T)])). \tag{4}$$

Finally, we apply the inverse fast Fourier transform (IFFT) $F^{-1}(\cdot)$ [8] to convert the fused amplitude $\widetilde{a}(R_T)$ and phase $\widetilde{p}(R_T)$ back to spatial space, gaining the enhanced spatial feature F_T as follows:

$$F_T = \mathcal{V}(F^{-1}(\widetilde{a}(R_T), \widetilde{p}(R_T))) + R_T. \tag{5}$$

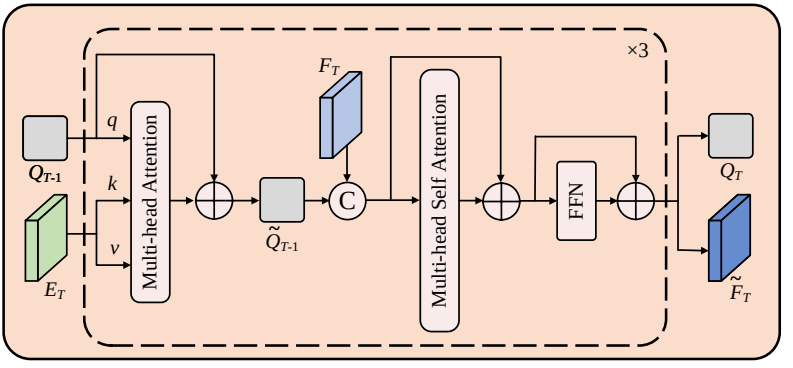

Figure 4: The structure of the proposed Spatio-temporal Information Decoder. The temporal query Q_{T-1} is propagated from the $T-1^{th}$ frame. E_T and F_T correspond to the event feature and the enhanced spatial feature extracted from the FFE respectively.

Through the FFE module, the output feature F_T is enhanced by the high-frequency information from events, enabling the tracker to better capture the target's appearance details under degraded conditions. Furthermore, learning in frequency space allows for a broader receptive field that captures global contextual information [8], complementing pixel-level details and enhancing the model's representational capability.

3.3 Spatio-temporal Information Decoder

As a temporal prediction task, VOT involves tracking targets that exist within a dynamic context. However, current RGB-E trackers [10, 18, 21, 27, 28, 31] overlook this issue, relying solely on independent template matching within each image pair and ignoring the historical motion cues from previous frames. Additionally, these trackers compress event streams into frame-like images, neglecting the high-temporal-resolution information inherent in events. Consequently, they struggle to address challenges such as occlusion and background clutter, which require the comprehension of the target's motion pattern.

Inspired by [22, 29], we propose the SID module, which employs an auto-regressive temporal query to integrate the enhanced spatial feature and historical motion cues for robust tracking. The temporal query Q captures fine-grained motion information of the target from events. It is propagated across adjacent frames to convey these temporal cues, facilitating frame-to-frame association. The SID module, consisting of three decoder layers, is designed to utilize the temporal query. The structure of each layer is shown in Figure 4.

When tracking the target in the T^{th} frame, in the SID, the temporal query Q_{T-1} from the last frame will first perform cross-attention with the event feature E_T of the current frame, generating the fine-grained query $\widetilde{Q}_{T-1}$ as follows:

$$\widetilde{Q}_{T-1} = Q_{T-1} + MHA(Q_{T-1}, E_T, E_T), \tag{6}$$

where $MHA(\cdot)$ refers to the multi-head attention mechanism in [19]. Since the event feature E_T records the target's motion during the frame's blind time of $T-1 \sim T$ with a high measurement rate, it complements the temporal query Q_{T-1} with temporally fine-grained motion

information through cross-attention. As a result, the generated query $\widetilde{Q}_{T-1}$ is better suited for predicting the target's state in the T^{th} frame.

The query $\widetilde{Q}_{T-1}$ is then concatenated with the enhanced spatial feature F_T extracted from the FFE. This concatenated input will be processed as follows:

$$\begin{aligned} K &= [\widetilde{Q}_{T-1}; F_T] + MHSA([\widetilde{Q}_{T-1}; F_T]), \\ [Q_T; \widetilde{F}_T] &= K + FFN(K), \end{aligned} \tag{7}$$

where MHSA($\cdot$) refers to the multi-head self-attention [7], K is the output of this mechanism, FFN($\cdot$) refers to the feed-forward network. This approach enables two key advantages. (1) The generated feature $\widetilde{F}_T$ will predict the target's location more accurately, guided by the fine-grained temporal query $\widetilde{Q}_{T-1}$. (2) The query $\widetilde{Q}_{T-1}$ can learn the target's state at the current timestamp T and update its information in an auto-regressive way, generating the new temporal query Q_T for tracking guidance in the next frame.

3.4 Head and Loss

The SID's output $\widetilde{F}_T$ will finally enter the prediction head for target localization. Our prediction head is the same as [26]. For the loss function, we adopt L_1 loss, generalized IoU loss L_{GIOU} [17] and the focal loss [15] as classification loss L_{cls}. The total loss L can be formulated as:

$$L = L_{cls} + \lambda_1 L_1 + \lambda_2 L_{GIOU}, \tag{8}$$

where $\lambda_1 = 5$ and $\lambda_2 = 2$ are the same as in [26].

4 Experiments

To demonstrate our HFTrack's effectiveness, we carry out comprehensive experiments on two real-world RGB-E datasets: VisEvent [21] and COESOT [18].

VisEvent: VisEvent [21] is a visible-event dataset collected from the real world. Its spatial resolution is 346 $\times$ 260. VisEvent contains various rigid and nonrigid targets in multiple scenarios. Following [18], there are 205 sequences for training and 172 for testing.

COESOT: COESOT [18] comprises 1354 color-event videos with 478,721 frames, which can be split into 827 and 527 sequences for training and testing respectively. These sequences cover the widest range of object categories (90 classes).

4.1 Implementation Details

We employ OSTrack [26] (ViT-B ver.) pre-trained on multiple datasets to initialize the weights of HFTrack's frozen parts. We utilize a template size of 128 $\times$ 128 and a search region size of 256 $\times$ 256. Additionally, the search region's center and scale jitter factors are set to 3 and 0.25, with no jitter applied to the template. HFTrack is trained and tested on 4 NVIDIA RTX 3090 GPUs with Pytorch 1.10.0. The training process takes 35 epochs at all, and each epoch contains 6×10^4 sample pairs. The AdamW optimizer with a weight decay of 10^{-4} is adopted. Following [18], we adopt the success rate (SR), the precision rate (PR), and the normalized precision rate (NPR) as evaluation metrics.

Trackers	VisEvent [21]			COESOT [18]		
	SR(%)	PR(%)	NPR(%)	SR(%)	PR(%)	NPR(%)
MDNet-MF [21]	46.3	67.5	59.8	56.3	69.0	70.0
OSTrack [26]	51.0	67.5	62.6	59.0	70.7	70.5
CEUTrack [18]	53.5	71.8	66.4	62.7	76.0	74.9
AFNet [28]	49.5	67.7	61.6	59.4	74.5	71.3
HRCEUTrack [31]	51.0	68.0	63.1	63.2	77.4	76.5
SeqTrack [4]	58.4	75.4	70.7	65.8	79.5	78.8
ODTrack [29]	58.2	76.1	71.2	64.4	77.3	76.5
SDSTrack [10]	60.2	77.3	72.7	66.6	79.5	78.6
HFTrack (Ours)	62.3	78.5	75.0	67.7	80.7	80.1

Table 1: Performance comparison of our HFTrack against other SOTA trackers on VisEvent and COESOT. The best and second best results are marked in red and blue colors.

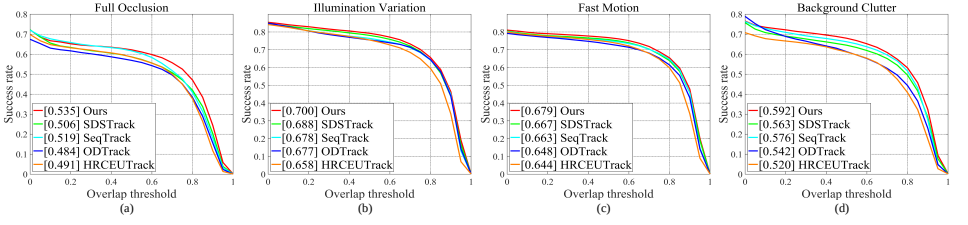

Figure 5: Success plots of trackers in four scenarios on COESOT [18]. (a) Full occlusion, (b) Illumination variation, (c) Fast motion, (d) Background clutter. **Zoom in for best view.**

4.2 Experimental Results

As shown in Table 1, our proposed HFTrack achieves 62.3%, 78.5%, 75.0% in SR, PR, and NPR respectively on VisEvent [21], outperforming the runner-up SDSTrack [10] by 2.1%, 1.2% and 2.3%. This result demonstrates that HFTrack presents stronger adaptability to tracking objects under diverse scenarios. Additionally, HFTrack outperforms others with an SR of 67.7%, PR of 80.7%, and NPR of 80.1% on COESOT [18], confirming HFTrack's superior generalization capability of tracking different kinds of targets.

To verify whether our HFTrack excels in tracking targets with varying appearance and motion patterns, we evaluate its performance in four challenging scenarios on COESOT [18]. As shown in Figure 5, our HFTrack outperforms other trackers in illumination variation and fast motion, confirming our proposed FFE's effectiveness in capturing the moving target's appearance details under degraded conditions. On top of that, when tracking objects in full occlusion or background clutter, HFTrack shows better performance in understanding and predicting the target's motion state based on the SID module.

We further conduct comparative experiments against SDSTrack [10] and SeqTrack [4] in model parameters, MACs and inference speed. As shown in Table 2, HFTrack exhibits lower computational complexity and higher inference speed with a similar amount of parameters.

To intuitively demonstrate the effectiveness of HFTrack in long-term tracking, we visualize the tracking results of HFTrack alongside other three trackers [4, 10, 29] in both outdoor and indoor scenarios. As shown in Figure 6, HFTrack focuses on the target robustly while

Trackers	Params(M)(↓)	MACs(G)(↓)	Speed(FPS)(↑)
SeqTrack [4]	**89.3**	66.2	40.0
SDSTrack [10]	102.2	108.4	20.9
HFTrack (Ours)	96.3	**42.3**	**58.8**

Table 2: Comparison of model parameters, MACs, and inference speed.

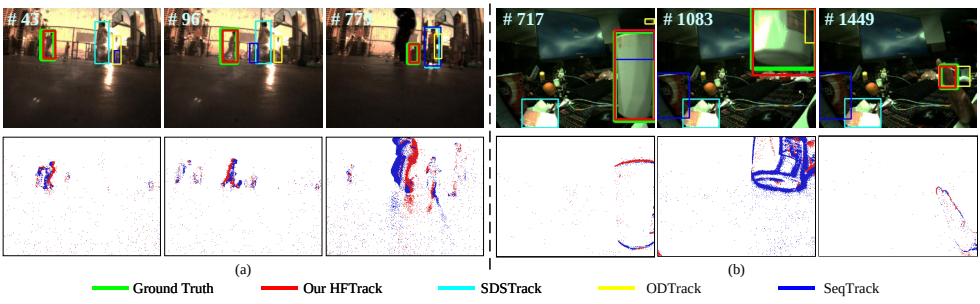

Figure 6: Visualized comparison results in long-term tracking. (a) and (b) correspond to outdoor and indoor scenarios. Event images listed in the second row are used for visualization.

Method		VisEvent [21]			COESOT [18]		
FFE	SID	SR(%)	PR(%)	NPR(%)	SR(%)	PR(%)	NPR(%)
✗	✗	59.4	76.0	72.6	66.3	79.6	79.3
✓	✗	61.9	77.6	74.3	67.2	80.2	79.7
✗	✓	60.8	76.4	72.9	66.7	79.9	79.5
✓	✓	**62.3**	**78.5**	**75.0**	**67.7**	**80.7**	**80.1**

Table 3: Ablation study for modules on VisEvent [21] & COESOT [18].

other trackers are misled by similar objects in the background. HFTrack also predicts the target's location precisely in case of occlusion. These results confirm that HFTrack can better understand and predict target's motion state based on the SID module. It is also worth noting that, as an RGB-E tracker, SDSTrack [10] fails to locate the object even when the target is highlighted in events, indicating that it does not fully utilize the event signal in the spatial domain. More visualization results can be found in the supplementary material.

4.3 Ablation Study

We conduct the ablation study on VisEvent [21] and COESOT [18] to validate the effectiveness of our proposed FFE and SID modules. As shown in Table 3, the baseline model, which excludes the FFE and SID, performs the poorest. Introducing either FFE for more appearance details or SID for more temporal motion cues can improve the performance of the baseline. Only when both FFE and SID modules are incorporated can our tracker achieve the best performance. This result indicates that both FFE and SID modules are essential for improving our HFTrack's tracking ability.

Furthermore, we visualize the attention maps of our model before and after applying the

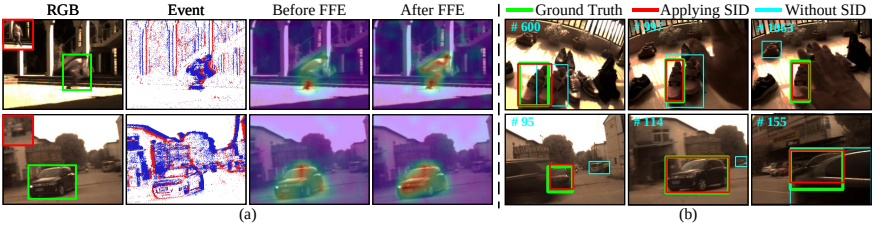

Figure 7: (a) The attention maps before and after applying the FFE module. (b) The tracking results of HFTrack with and without the SID module.

FFE module. As shown in Figure 7 (a), the FFE module significantly enhances the tracker's ability to capture the target's appearance details under degraded conditions, emphasizing features such as edges and textures which are highlighted in the event modality. Besides, we compare the tracker's tracking results with and without the SID module. As shown in Figure 7 (b), excluding the SID module which provides motion cues for tracking guidance, the tracker fails to understand the target's motion state and tends to be misled by similar objects in the background, confirming the necessity of the SID module.

5 Conclusion

In this paper, we propose HFTrack to leverage spatio-temporal high-frequency information from events for better tracking performance. We first propose an FFE module to capture appearance details from events for enhancing RGB features in frequency space. We also introduce an SID module based on an auto-regressive temporal query, which integrates the enhanced spatial features and high-temporal-resolution motion cues from events for robust tracking. Extensive experiments demonstrate the effectiveness of our HFTrack.

6 Acknowledgement

This work was in part supported by the National Natural Science Foundation of China under grants 62472399 and 62021001.

References

[1] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *ICCV*, pages 6182–6191, 2019.

[2] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341, 2014.

[3] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, pages 8126–8135, 2021.

[4] Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *CVPR*, pages 14572–14581, 2023.

[5] Xuerui Dai, Xue Yuan, and Xueye Wei. Tirnet: Object detection in thermal infrared images for autonomous driving. *Applied Intelligence*, 51(3):1244–1261, 2021.

[6] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *CVPR*, pages 4660–4669, 2019.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Matteo Frigo and Steven G Johnson. Fftw: An adaptive software architecture for the fft. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 1381–1384. IEEE, 1998.

[9] Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.

[10] Xiaojun Hou, Jiazheng Xing, Yijie Qian, Yaowei Guo, Shuo Xin, Junhao Chen, et al. Sdstrack: Self-distillation symmetric adapter learning for multi-modal visual object tracking. In *CVPR*, pages 26551–26561, 2024.

[11] Dachun Kai, Yueyi Zhang, and Xiaoyan Sun. Video super-resolution via event-driven temporal alignment. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2950–2954. IEEE, 2023.

[12] Dachun Kai, Jiayao Lu, Yueyi Zhang, and Xiaoyan Sun. Evtexture: event-driven texture enhancement for video super-resolution. In *Forty-first International Conference on Machine Learning*, 2024.

[13] Dachun Kai, Yueyi Zhang, Jin Wang, Zeyu Xiao, Zhiwei Xiong, and Xiaoyan Sun. Event-Enhanced Blurry Video Super-Resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4175–4183, 2025.

[14] Taewoo Kim, Hoonhee Cho, and Kuk-Jin Yoon. Frequency-aware event-based video deblurring for real-world motion blur. In *CVPR*, pages 24966–24976, 2024.

[15] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, pages 734–750, 2018.

[16] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, pages 8971–8980, 2018.

[17] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, pages 658–666, 2019.

[18] 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.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[20] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *CVPR*, pages 1571–1580, 2021.

[21] 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. *IEEE Transactions on Cybernetics*, 54(3):1997–2010, 2023.

[22] Jinxia Xie, Bineng Zhong, Zhiyi Mo, Shengping Zhang, Liangtao Shi, Shuxiang Song, and Rongrong Ji. Autoregressive queries for adaptive tracking with spatio-temporal transformers. In *CVPR*, pages 19300–19309, 2024.

[23] Junliang Xing, Haizhou Ai, and Shihong Lao. Multiple human tracking based on multi-view upper-body detection and discriminative learning. In *2010 20th International conference on pattern recognition*, pages 1698–1701. IEEE, 2010.

[24] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *ICCV*, pages 10448–10457, 2021.

[25] Jinyu Yang, Shang Gao, Zhe Li, Feng Zheng, and Aleš Leonardis. Resource-efficient rgbd aerial tracking. In *CVPR*, pages 13374–13383, 2023.

[26] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *ECCV*, pages 341–357, 2022.

[27] Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong. Object tracking by jointly exploiting frame and event domain. In *ICCV*, pages 13043–13052, 2021.

[28] 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 *CVPR*, pages 9781–9790, 2023.

[29] Yaozong Zheng, Bineng Zhong, Qihua Liang, Zhiyi Mo, Shengping Zhang, and Xianxian Li. Odtrack: Online dense temporal token learning for visual tracking. In *AAAI*, volume 38, pages 7588–7596, 2024.

[30] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *CVPR*, pages 989–997, 2019.

[31] Zhiyu Zhu, Junhui Hou, and Dapeng Oliver Wu. Cross-modal orthogonal high-rank augmentation for rgb-event transformer-trackers. In *ICCV*, pages 22045–22055, 2023.