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

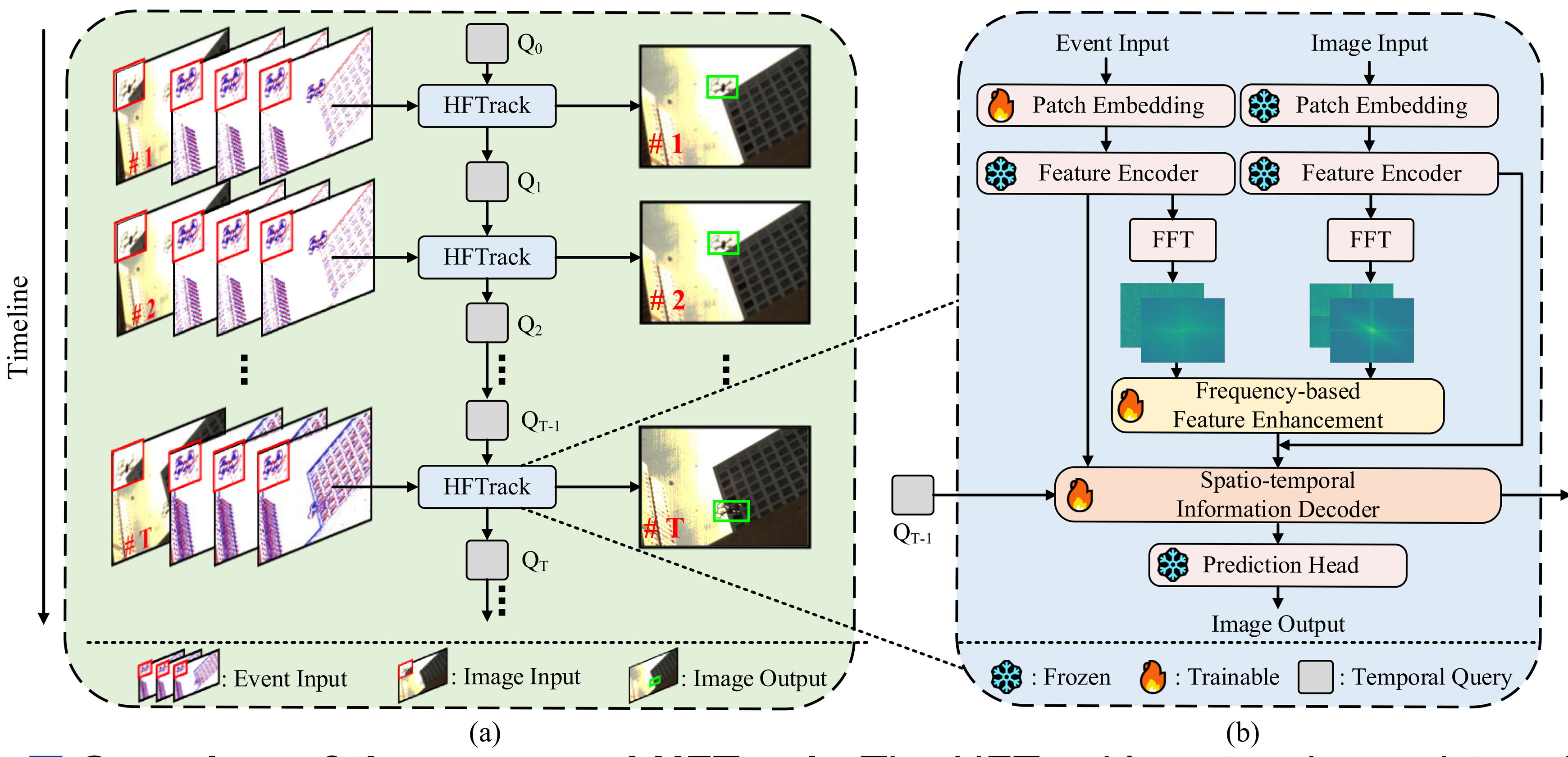
Yuheng Jiang[1], Hebei Li[1], Dachun Kai[1], Yansong Peng[1], Jiahui Yuan[1], Peilin Xiao[1], Yueyi Zhang[2], Xiaoyan Sun[1†]

[1]University of Science and Technology of China, [2]MiroMind

Introduction

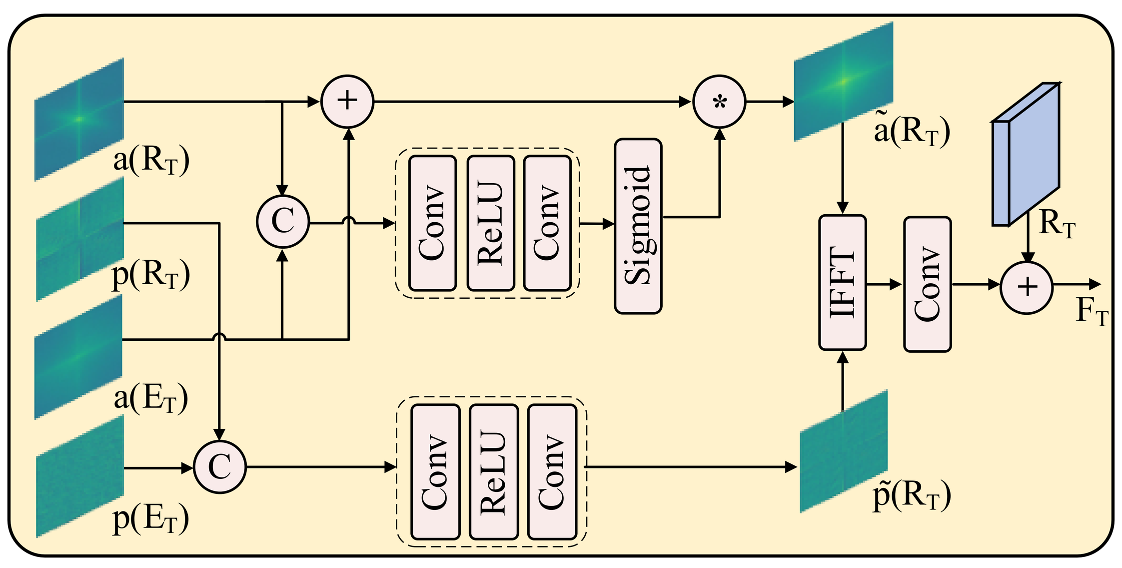
- Visual Object Tracking (VOT) aims to identify and locate the target within a video sequence based on its initial status in the template.
- Most of the existing RGB-Event trackers focus on intra-frame similarity matching, and neglect the high-temporal-resolution information inherent in events. Additionally, these trackers perform implicit feature fusion in spatial space, failing to leverage the high-frequency appearance details provided by events.
- To address the challenges, we propose HFTrack to fully leverage the spatio-temporal high-frequency information within event signals. In the spatial domain, we propose a module to capture high-frequency appearance details from events. In the temporal domain, we design a spatio-temporal information decoder to integrate high-temporal-resolution motion cues in events and spatial features for robust tracking performance.

Methodology



- **Overview of the proposed HFTrack:** The HFTrack's network consists of four components: the pre-trained feature encoder, a Frequency-based Feature Enhancement module (FFE), a Spatio-temporal Information Decoder (SID), and a pre-trained prediction head.

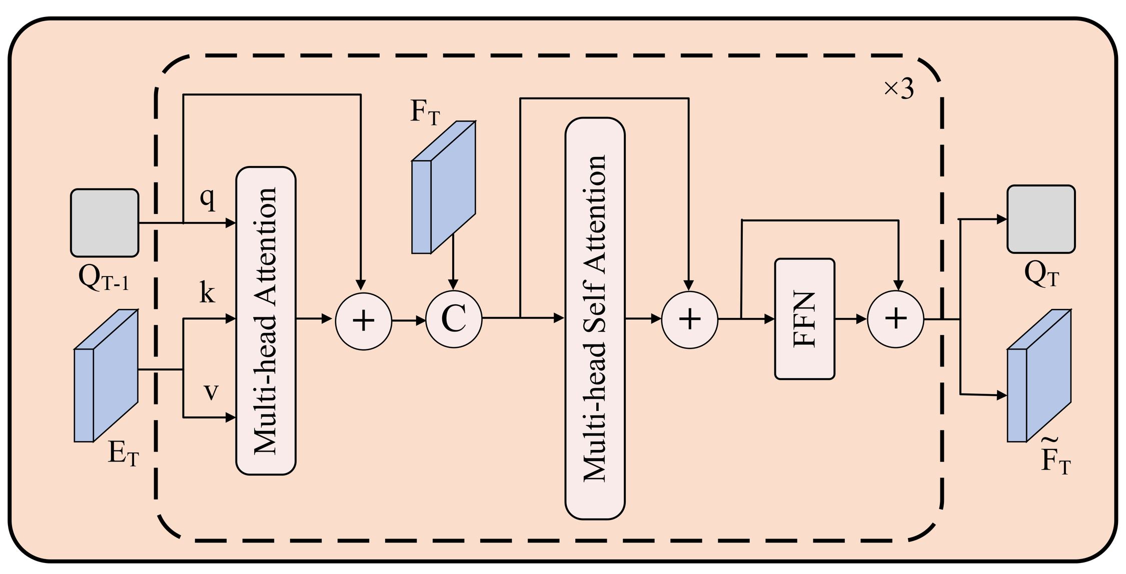
  To preserve the high-temporal-resolution information in events, we convert the event stream to the voxel grid form with the bin = 3. The frozen parts of our network are the same as OSTrack.



- **The Frequency-based Feature Enhancement Module (FFE):**

  We adopt fast Fourier transform (FFT) to the extracted RGB feature R_T and event feature E_T, generating inputs $a(R_T)$, $p(R_T)$, $a(E_T)$, $p(E_T)$.

  Through this FFE module, the output feature F_T is enhanced by the high-frequency information from events.



- **The Spatio-temporal Information Decoder (SID):**

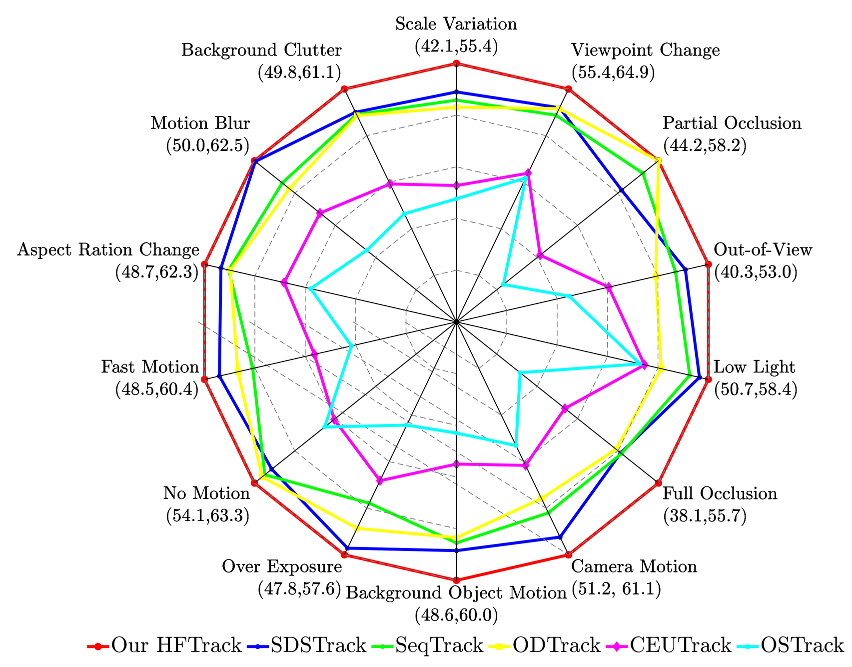
  We employ a temporal query Q to capture fine-grained motion information of the target from high-temporal-resolution event feature E_T. It is propagated across adjacent frames, facilitating frame-to-frame association.

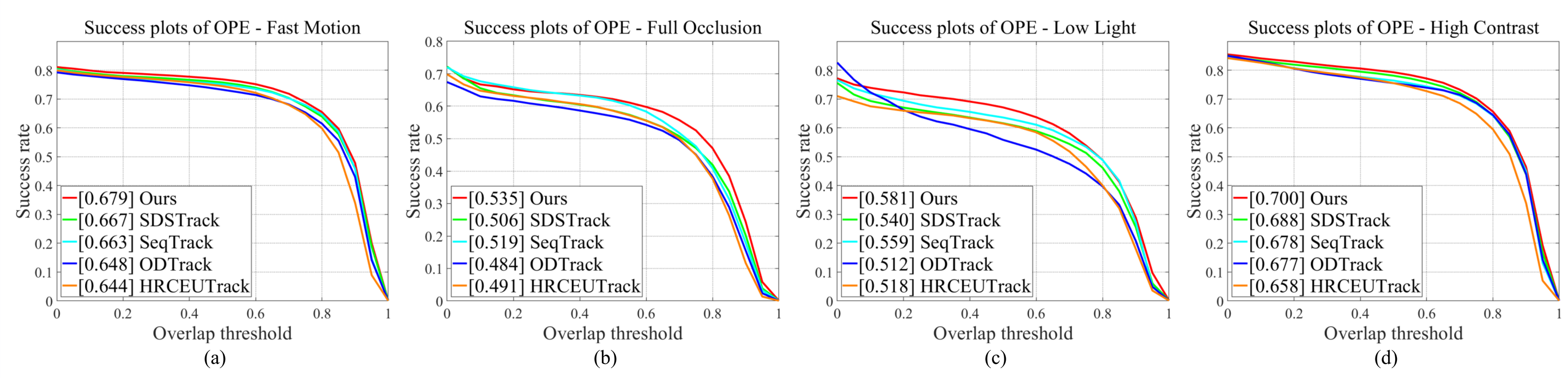
  The SID module consists of three decoder layers based on attention mechanism. It is designed to integrate the spatial appearance feature F_T and historical motion cues from the temporal query Q for robust tracking.

Experiments

Results on VisEvent and COESOT

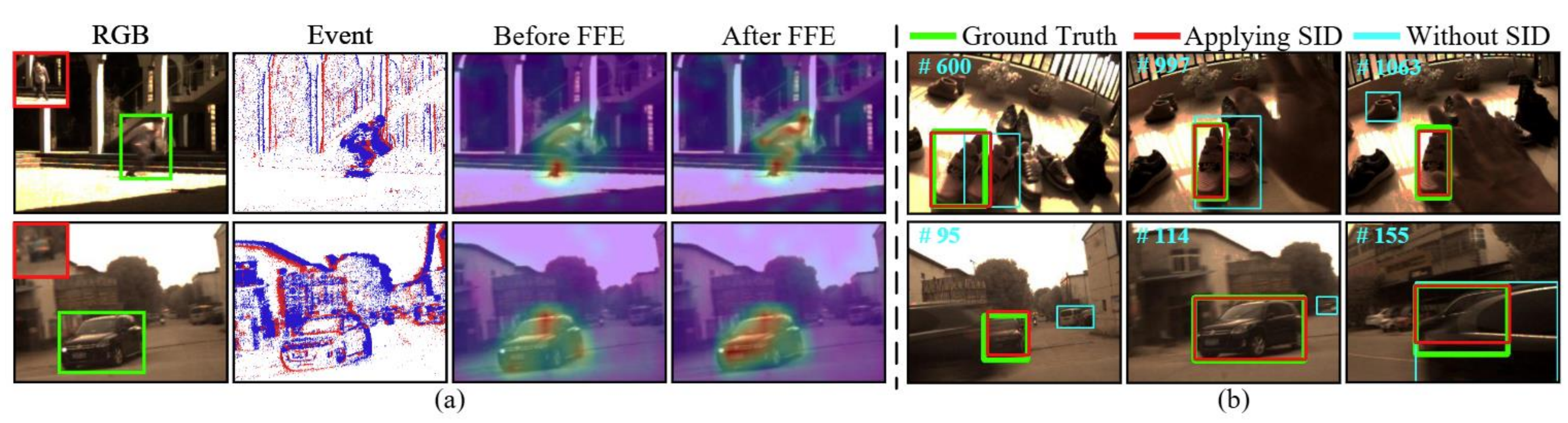
Trackers	VisEvent SR(%)	VisEvent PR(%)	VisEvent NPR(%)	COESOT SR(%)	COESOT PR(%)	COESOT NPR(%)
MDNet-MF	46.3	67.5	59.8	56.3	69.0	70.0
OSTrack	51.0	67.5	62.6	59.0	70.7	70.5
CEUTrack	53.5	71.8	66.4	62.7	76.0	74.9
AFNet	49.5	67.7	61.6	59.4	74.5	71.3
HRCEUTrack	51.0	68.0	63.1	63.2	77.4	76.5
SeqTrack	58.4	75.4	70.7	65.8	79.5	78.8
ODTrack	58.2	76.1	71.2	64.4	77.3	76.5
SDSTrack	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



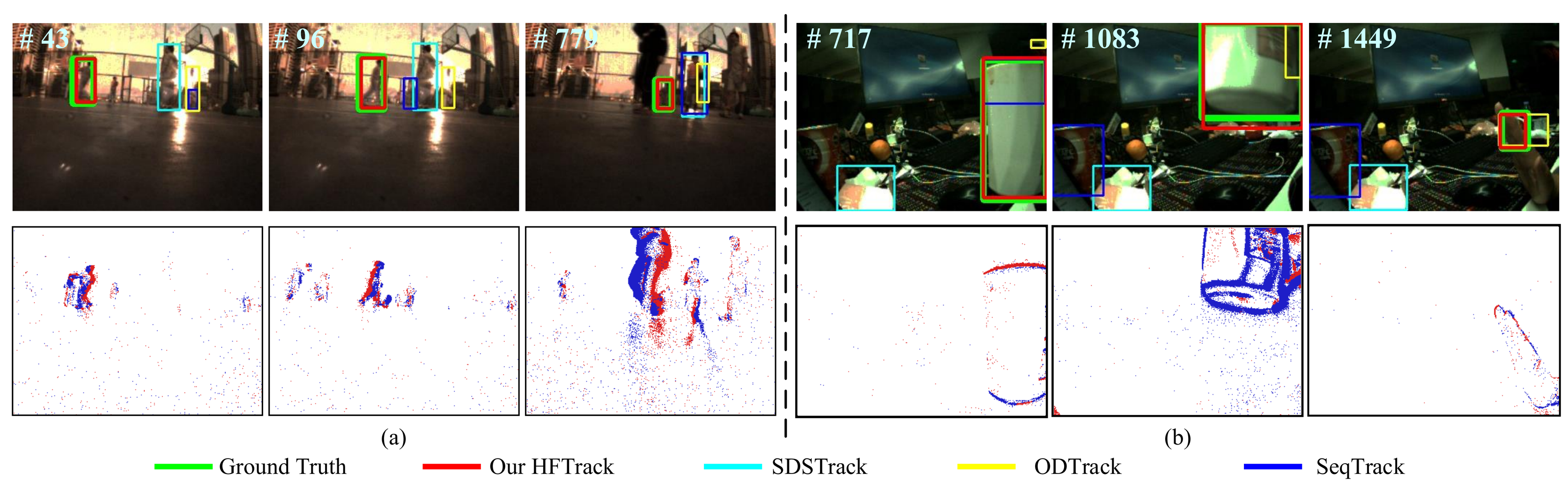


Ablation Studies

FFE	SID	VisEvent SR(%)	VisEvent PR(%)	VisEvent NPR(%)	COESOT SR(%)	COESOT PR(%)	COESOT 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**



Case Analysis



Conclusion

- 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.