Frame-Event Alignment and Fusion Network for High Frame Rate Tracking (Supplementary Material)

Jiqing Zhang[1], Yuanchen Wang[1], Wenxi Liu[2], Meng Li[3], Jinpeng Bai[1], Baocai Yin[1], Xin Yang[1,⋆]
[1]Dalian University of Technology, [2]Fuzhou University, [3]HiSilicon(Shanghai) Technologies Co.,Ltd

In this Supplemental Material, we present additional analyses and qualitative results in support of the findings from the main paper. In section 1, we provide a comparison of single-modality with multi-modality under different challenging conditions. Next, section 2 provides additional visual examples from the VisEvent [5] dataset. Section 3 analyzes whether our fusion module can be replaced by lightweight models. Next, section 4 provides qualitative results from the video interpolation method SuperSloMo [3] to analyze why interpolation on low frame rate sequences cannot produce satisfactory high frame rate tracking results compared to employing event-based cameras. Finally, we provide a Supplemental Video of tracking results on different datasets in section 5 to intuitively demonstrate the effectiveness of the proposed AFNet under various degraded conditions.

1. Impact of Multi-Modality Fusion under Different Conditions

To get additional insight into the influence of multi-modality fusion, we compare single-modality with multi-modality under different challenging conditions. Specifically, we conduct two experiments: (i) Comparison of event-only and multimodal training in High Dynamic Range (HDR), Low Light (LL), and Fast Motion (FM) scenes; (ii) Comparison of grayscale-frame-only and multimodal training in No-Motion (NM) and Severe Background Motion (SBM) scenarios. Table 1 shows that our multi-modality method obtains the best results under all five conditions, demonstrating the significance of multi-modality fusion for robust high frame rate tracking.

Methods	HDR		LL		FM		NM		SBM	
	RSR	RPR	RSR	RPR	RSR	RPR	RSR	RPR	RSR	RPR
STARKs [6]	47.5	73.1	50.6	77.5	39.4	59.5	20.2	40.3	16.9	22.3
TransT [1]	43.4	68.7	48.7	68.5	54.5	82.3	9.9	22.5	18.1	27.2
ToMP [4]	51.3	78.9	46.2	71.5	64.1	94.6	27.8	52.2	17.6	27.5
AFNet(Ours)	**55.5**	**84.9**	**64.7**	**93.8**	**66.3**	**96.4**	**62.0**	**98.8**	**60.1**	**90.3**

Table 1. Single vs. Multi-modality. Blue and green denote methods are trained with only event and frame modality, respectively.

2. Qualitative Results on VisEvent

We provide additional qualitative results of our AFNet compared to state-of-the-art approaches on the VisEvent [5] dataset. Compared with the FE240hz [9] dataset, the VisEvent provides a low annotation frequency, about 25Hz. However, it contains various rigid and non-rigid targets both indoors and outdoors. Therefore, we employ VisEvent to verify that our AFNet still remains effective for low frame rate tracking. Four examples containing rigid and non-rigid targets of the top-5 state-of-the-art approaches (*i.e.*, ToMP [4], DeT [7], HMFT [10], FENet [9] and our AFNet) are shown in Figure 1. The proposed AFNet makes the best estimate in all four examples.

Take note that we adopt a different event representation method for VisEvent to validate the generalization of our AFNet. Specifically, given an event stream $E_{i\rightarrow i+1} = \{[x_k, y_k, t_k, p_k]\}_{k=0}^{N-1}$ contains N events triggered during the interval $[i, i+$

⋆ Xin Yang (xinyang@dlut.edu.cn) is the corresponding author.

1]. Following STNet [8], we record the spatial positions of positive and negative events that have occurred at each pixel, respectively. Which can be defined as,

$$E_p(x, y, t) \doteq \delta\left(x - x_k, y - y_k\right) \delta\left(t - t_k\right), \tag{1}$$
$$E_n(x, y, t) \doteq \delta\left(x - x_k, y - y_k\right) \delta\left(t - t_k\right), \tag{2}$$

where E_p and E_n denote aggregated images from positive and negative events, respectively.

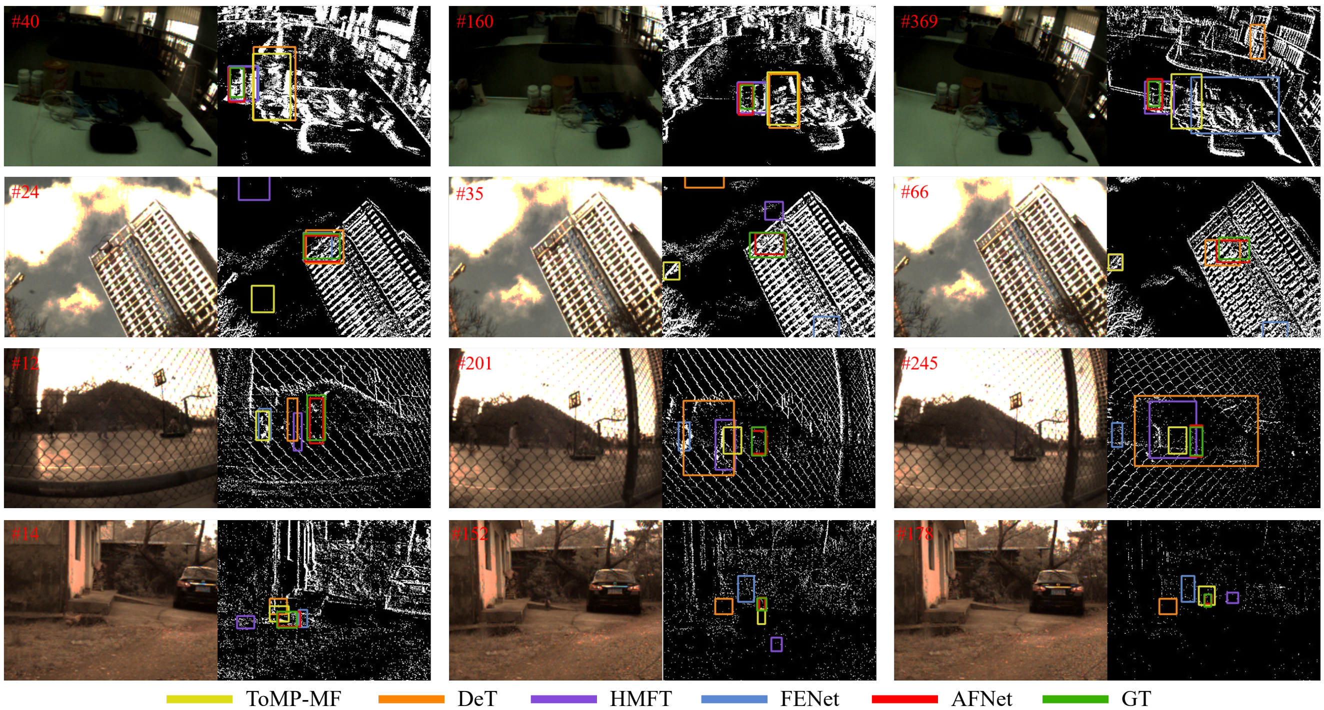

Figure 1. Qualitative comparison of different trackers on the VisEvent dataset. The tracking targets are *bottle*, *UAV*, *person*, and *chicken*, respectively. The first two are rigid targets, while the last two are non-rigid.

3. Influence of Fusion Module

A question in our mind is whether replacing our Cross-correlation Fusion (CF) with lightweight architectures can still achieve similar performance. To answer this question, we replace our CF with the building blocks of two lightweight models: SqueezeNet [2] and ShuffleNet [11], respectively. As shown in Table 2, our CF fares best in all four metrics. Besides, the ablation in our paper verified the effectiveness of our design. Simplifying the model structure while keeping or even improving the performance will be our future work.

	RSR ↑	$OP_{0.50}$ ↑	$OP_{0.75}$ ↑	RPR↑
SqueezeNet [2]	56.1	70.7	31.0	83.2
ShuffleNet [11]	54.7	69.8	28.0	82.3
CF (Ours)	**58.4**	**73.5**	**32.6**	**87.0**

Table 2. Comparison of lightweight models and our CF.

4. Qualitative Results of SuperSloMo

To get insight into why frame interpolation on low frame rate sequences can not facilitate the performance of high frame rate tracking, we offer qualitative examples of video interpolation method SuperSloMo [3] on the FE240hz [9] dataset. As shown in the first two cases in Figure 2, the interpolation results of SuperSloMo in challenging conditions (*i.e.*, HDR and low light) are still insufficient for locating targets. By contrast, an event-based camera does not suffer from these scenarios. Furthermore, due to the irregular motion of the target, the results of interpolation cannot accurately reflect the position of the target as shown in the last two examples of Figure 2, which leads to tracking failure. Conversely, the high temporal resolution of event-based cameras provides auxiliary visual information in the blind-time between frames. These results demonstrate that introducing events for achieving high frame rate tracking is a feasible and significant manner.

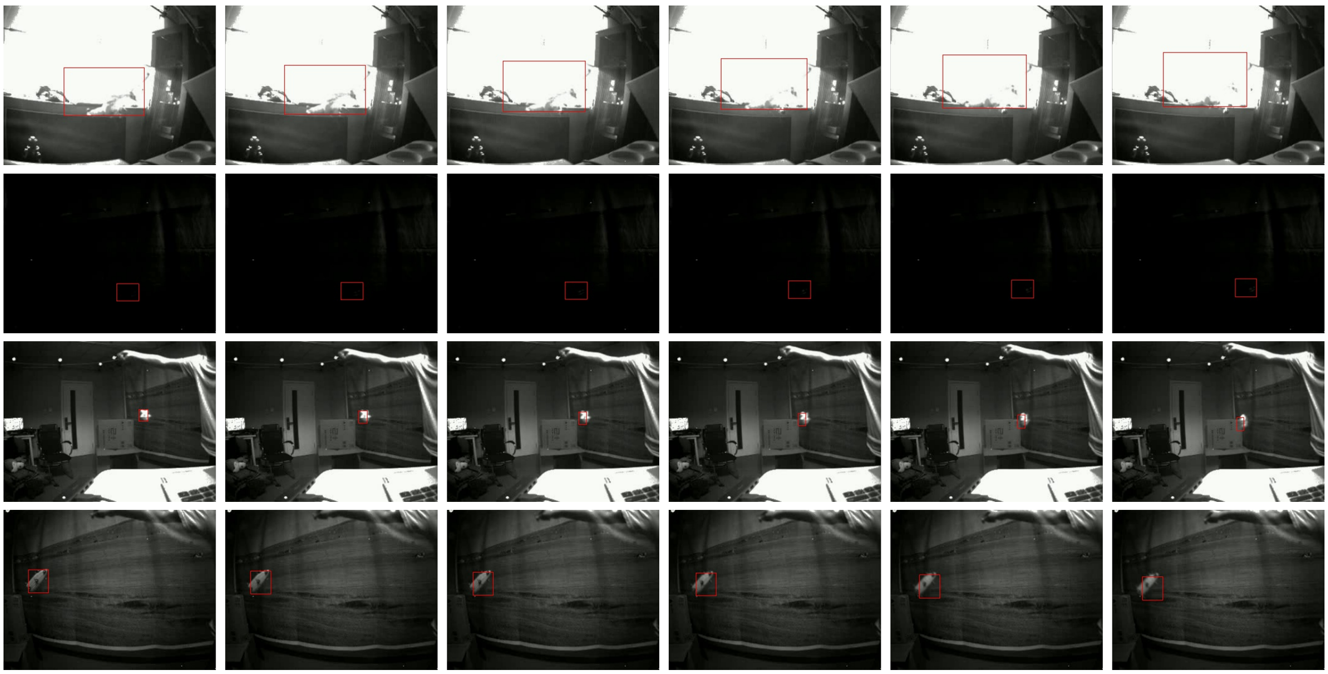

Figure 2. Interpolation results from SuperSloMo [3] on the FE240hz [9] dataset.

5. Supplementary Video

We provide more qualitative results of our method compared to state-of-the-art trackers to further verify the effectiveness of our AFNet under various challenging conditions. This video includes five degraded scenarios (*i.e.*, high dynamic range, low-light, fast motion, no motion, and severe background motion) of the FE240hz dataset and two attributes (*i.e.*, rigid and non-rigid targets) of the VisEvent dataset. We refer to `https://youtu.be/W7EjOiGMiAQ` for more details.

References

[1] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, 2021. 1

[2] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv*, 2016. 2

[3] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *CVPR*, 2018. 1, 2, 3

[4] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. In *CVPR*, 2022. 1

[5] 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*, 2021. 1

[6] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. *ICCV*, 2021. 1

[7] Song Yan, Jinyu Yang, Jani Kapyla, Feng Zheng, Ales Leonardis, and Joni-Kristian Kamarainen. Depthtrack: Unveiling the power of rgbd tracking. In *ICCV*, 2021. 1

[8] Jiqing Zhang, Bo Dong, Haiwei Zhang, Jianchuan Ding, Felix Heide, Baocai Yin, and Xin Yang. Spiking transformers for event-based single object tracking. In *CVPR*, 2022. 2

[9] Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, and Bo Dong. Object tracking by jointly exploiting frame and event domain. In *ICCV*, 2021. 1, 2, 3

[10] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *CVPR*, 2022. 1

[11] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, 2018. 2