MATE: Motion-Augmented Temporal Consistency for Event-based Point Tracking

Supplementary Material

1. Network Structure

Network Encoder. The encoder is based on ResNet [1] and consists of four feature extraction modules, each containing two stacked residual blocks. The first block in each module performs downsampling, while the second maintains the resolution. Features at different resolutions from each module are aligned to a consistent resolution via interpolation and concatenated along the channel to form multi-scale representations. The concatenated features are then compressed along the channel using a 1×1 convolution.

2. Implementation Details

Training details. To accelerate training, events in Ev-PointOdyssey are converted into time surface [2] representations at the original spatial resolution and stored. Sampling is conducted with step sizes of $\{1, 2, 4\}$, where each sampled sequence has a length of 48. Starting points for sampling are selected along the timeline at intervals of 2, ensuring comprehensive coverage of the entire sequence. This process generates approximately $340k$ training samples, each containing 128 tracking points.

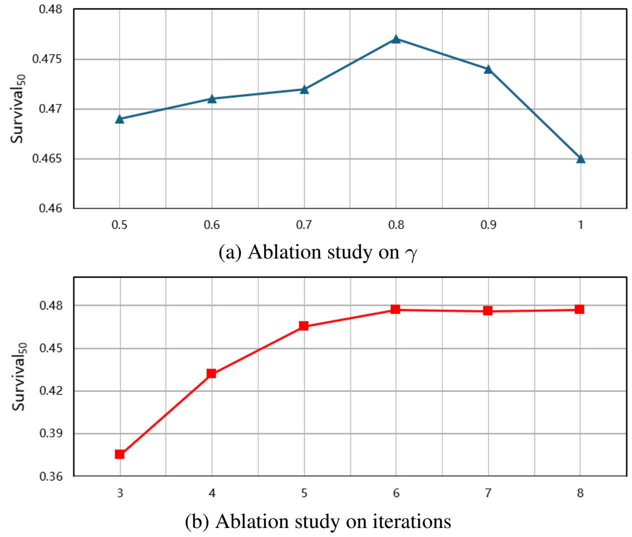
Data augmentations. To improve model robustness, several data augmentation techniques are applied. Random pixel erasure is performed at all time steps, with varying erased regions to enhance the network's ability to handle occlusions. The erased values are replaced by the mean of the original region. Additionally, random scale-up is employed to improve tracking across different object scales, with scale parameters for the next time step initialized from the previous step and adjusted with slight random perturbations. Spatial flipping and temporal reversal are also applied to further diversify the dataset.

3. Additional Ablation Experiments

In addition to the ablation studies reported in the main paper, experiments are conducted on the input event representation and the number of pixels used for plane fitting in the motion-guidance module. Due to time constraints, these experiments are performed on a baseline model, which excludes the variable motion aware module, and are conducted at a lower resolution of 192×256.

Table 1. Ablation studies on different input event representations.

Representations	Ev-PointOdyssey		
	$\sigma_{avg} \uparrow$	MTE $\downarrow$	$Survival_{50} \uparrow$
Event image	0.298	54.43	0.443
Voxel grid	0.312	51.24	0.457
Time surface	0.323	48.41	0.477



(a) Ablation study on γ

(b) Ablation study on iterations

Figure 1. The survival metric is maximized when gamma and iteration are set to 0.8 and 6, respectively.

Input representation. Different input event representations significantly impact event-based networks. To evaluate their effects on this task, various representations are used to train the baseline model, with the results presented in Tab. 1. Event image aggregates the number of events at each pixel over a fixed time window. Voxel grid [3] represents a spatiotemporal histogram, where each voxel corresponds to a specific pixel and time interval. Time surface [2] is a 2D map where each pixel stores the timestamp of the most recent event. As highlighted in the introduction of main paper, the TAP task relies on continuous motion information. Both event image and voxel grid quantize event timestamps, discarding dense temporal details, which limits their ability to capture smooth motion dynamics. In contrast, time surface retains rich temporal information, with each pixel encoding the motion history at its location, making it more suitable for encoding temporal motion cues effectively. As a result, time surface demonstrates superior performance compared with other representations.

Weight factor γ in loss function. As described in Eq. (8) of the main paper, the losses across different time steps are weighted by a factor γ, assigning a higher penalty to losses occurring at later time steps. An ablation study is

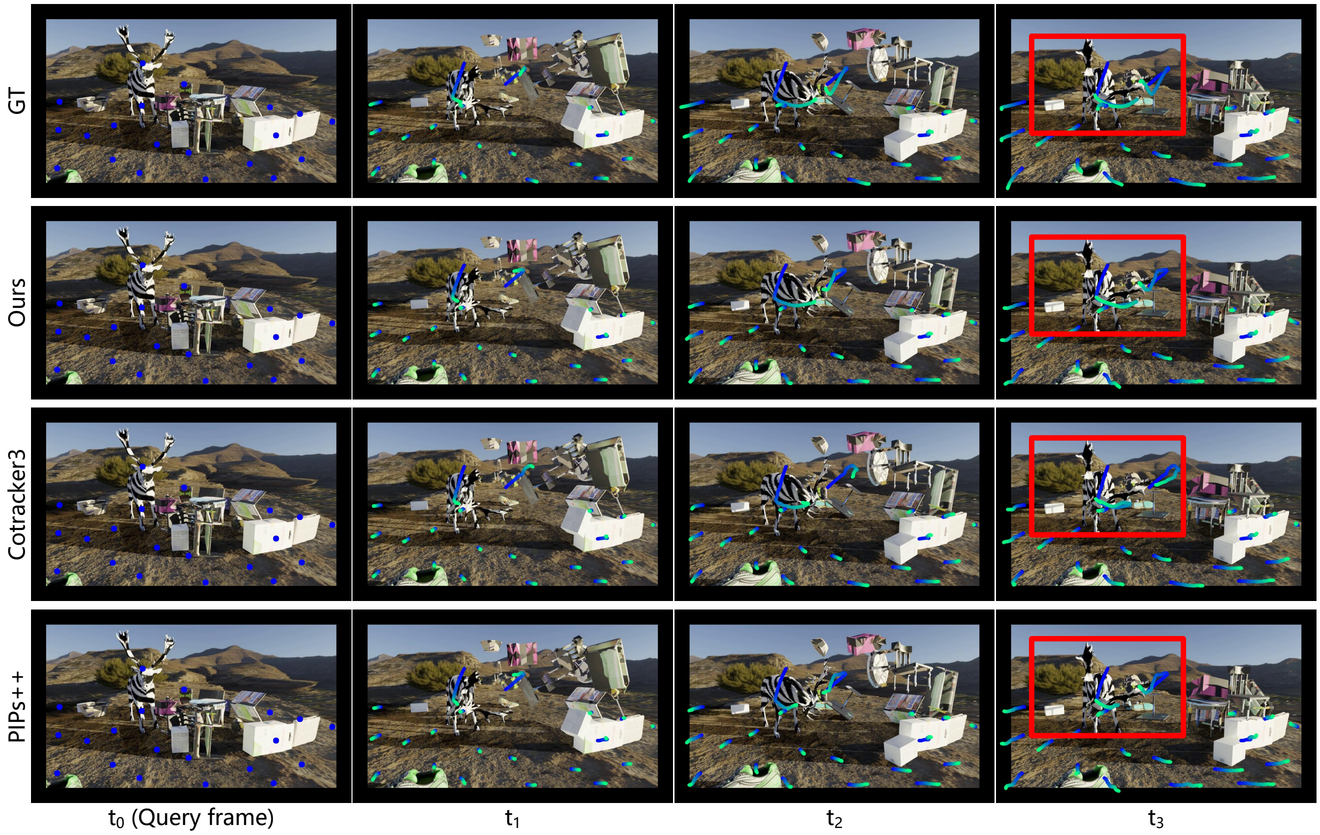


Figure 2. Qualitative results for Ev-PointOdyssey dataset.

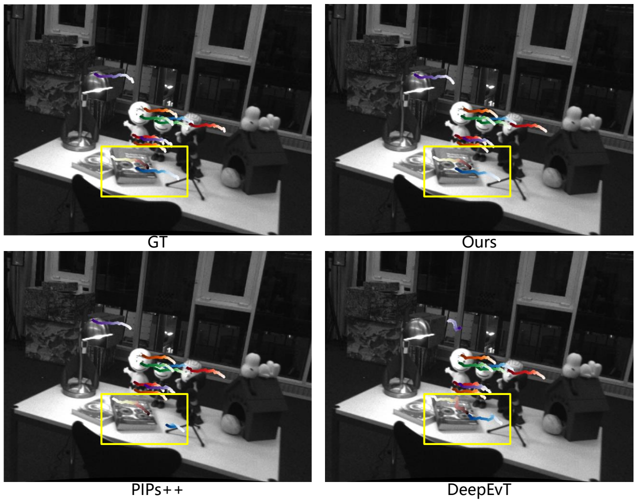


Figure 3. Qualitative results for EDS dataset.

performed to identify the optimal γ. Figure 1a demonstrates that the $Survival_{50}$ is maximized when γ equals 0.8.

The number of iterations. In addition to γ, the number of iterations also significantly affects tracking accuracy. While increasing the iteration count typically enhances accuracy, it also requires more computational resources. To find an optimal trade-off, an ablation study on the iteration count is performed. As illustrated in Fig. 1b, when the iteration count reaches 6, further increases do not yield significant improvements in the metrics.

4. More Qualitative Results

Additional qualitative results are provided in Figs. 2 and 3 with higher-resolution images for clearer visualization.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[2] Elias Mueggler, Chiara Bartolozzi, and Davide Scaramuzza. Fast event-based corner detection. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 33–1, 2017. 1

[3] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 989–997, 2019. 1