Supplementary Material for TrajectoryFormer: 3D Object Tracking Transformer with Predictive Trajectory Hypotheses

Xuesong Chen[1] Shaoshuai Shi[2 *] Chao Zhang[3] Benjin Zhu[1] Qiang Wang[3]
Ka Chun Cheung[4] Simon See[4] Hongsheng Li[1, 5, 6 *]
[1]MMLab, CUHK [2] Max Planck Institute for Informatics [3] Samsung Telecommunication Research
[4] NVIDIA AI Technology Center [5] CPII [6] Shanghai AI Laboratory
{chenxuesong@link, hsli@ee}.cuhk.edu.hk, shaoshuaics@gmail.com

1. More Ablation Studies

Effects of the Global-local interaction. We adopt different designs to investigate the effectiveness of our proposed global-local interaction module in Table 1. In the absence of interaction module, the network cannot incorporate global context information to model relationship among different hypotheses, leading to a 2.2% decrease in performance. In contrast, introducing the self-attention mechanism to enable global interaction among all hypotheses results in a substantial performance improvement of 1.9%. Additionally, the incorporation of local interaction of each tracked object, yields an additional gain of 0.3% in terms of MOTA on Waymo Open dataset.

Method	MOTA↑	FP↓	Miss↓	IDS↓
w/o interaction	57.6	13.2	29.0	0.17
Global interaction	59.5	12.0	28.4	0.16
Global-Local interaction	**59.8**	11.3	28.7	0.23

Table 1. Effects of global-local interaction for context modeling of trajectory hypotheses.

2. Tracking Visualization

As depicted in Figure 1, we visualize a tracking result in the Waymo dataset for a more intuitive comparison with our baseline CenterPoint. Our method exhibits several notable advantages over CenterPoint: (1) For stationary objects, our method generates more stable trajectory results (see (a)). (2) For moving objects, our method has smoother trajectories. Conversely, the center-distance-based strategy does not take into account of moving direction of trajectories, leading to heading inconsistencies (see (b1) for ped. and (b2) for veh.). (3) Our method has less false positives (see (c) and (d)).

*Corresponding authors

Method	Tracking Time	MOTA
CenterPoint [4]	3 ms	55.0
SimpleTrack [1]	374 ms	56.1
ImmotralTrack [3]	>1s	56.4
SpOT [2]	59 ms	55.7
TrajectoryFormer	60 ms	**59.8**

Table 2. Comparison of different method's inference time. All methods are evaluated on NVIDIA 3090 GPU.

3. Runtime Analysis

We provide the inference time of different methods, as depicted in Table 2. CenterPoint utilizes a greedy algorithm for trajectory-box association based on center-distance, which results in the highest efficiency but the lowest performance. Heuristic trackers, such as SimpleTrack and ImmotralTrack, might also be slow. SimpleTrack proposes a two-stage association strategy using 3D GIOU association metrics, which improves performance but also significantly reduces efficiency. This can be attributed to the following reasons: (1) The proposed two-stage association strategy involves a large number of low-quality boxes (score > 0.1) that are not utilized by CenterPoint (score > 0.7), leading to a substantial increase in association cost. (2) The additional components such as the Kalman Filter (for trajectory refining), GIoU, and bipartite matching also contribute to increased tracking time compared to CenterPoint's distance-based greedy algorithm. Based on SimpleTrack, ImmotralTracker never drop generated tracklets, resulting in large amount of redundant computation cost.

Our method strikes a balance between performance and inference time (not including data pre-processing). TrajectoryFormer adopts a long-short feature encoding strategy to minimize the computational overhead of embedding generation. Furthermore, it employs a global-local interaction module to efficiently perform trajectory-box association.

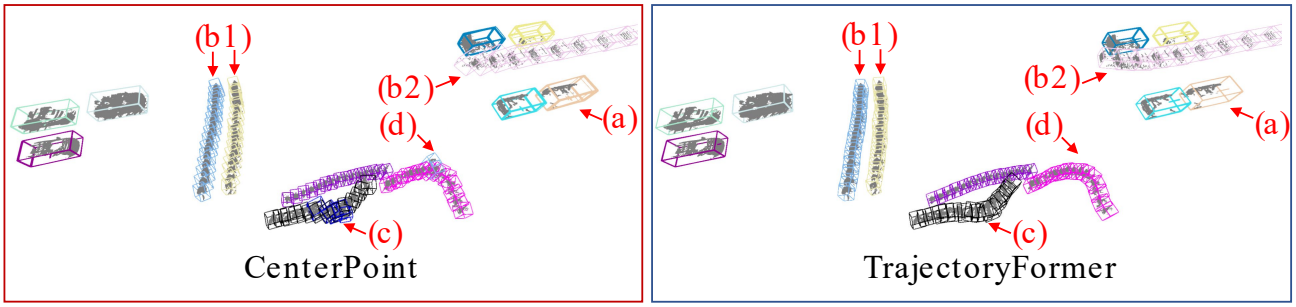

Figure 1. Qualitative comparison of CenterPoint and TrajectoryFormer trajectory quality. Please zoom in for details.

References

[1] Ziqi Pang, Zhichao Li, and Naiyan Wang. Simpletrack: Understanding and rethinking 3d multi-object tracking. *arXiv preprint arXiv:2111.09621*, 2021.

[2] Colton Stearns, Davis Rempe, Jie Li, Rareş Ambruş, Sergey Zakharov, Vitor Guizilini, Yanchao Yang, and Leonidas J Guibas. Spot: Spatiotemporal modeling for 3d object tracking. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pages 639–656. Springer, 2022.

[3] Qitai Wang, Yuntao Chen, Ziqi Pang, Naiyan Wang, and Zhaoxiang Zhang. Immortal tracker: Tracklet never dies. *arXiv preprint arXiv:2111.13672*, 2021.

[4] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.