Temporally Consistent Long-Term Memory for 3D Single Object Tracking

Supplementary Material

Table 1. Performance comparison with different numbers of memory tokens K on the KITTI dataset.

Setting	Car	Pedestrian	Van	Cyclist	Mean
$K = 16$	74.8/86.8	68.6/93.0	62.4/76.7	**78.1/94.8**	71.1/88.7
$K = 32$	**76.0/88.6**	**68.6/94.0**	**64.2/77.7**	77.7/**94.8**	**71.8/90.1**
$K = 64$	74.8/86.7	67.7/92.8	63.5/76.4	76.7/94.4	70.8/88.6

Table 2. Performance comparison with different temperature τ_{cycle} of memory cycle consistency loss on the KITTI dataset.

Setting	Car	Pedestrian	Van	Cyclist	Mean
$\tau_{\text{cycle}} = 0.07$	74.9/86.9	**68.7**/93.2	59.7/74.4	76.2/94.7	70.9/88.7
$\tau_{\text{cycle}} = 0.1$	**76.0/88.6**	68.6/**94.0**	**64.2**/77.7	**77.7/94.8**	**71.8/90.1**
$\tau_{\text{cycle}} = 1.0$	73.6/86.2	68.2/92.4	63.9/**78.2**	76.5/**94.8**	70.4/88.4

Table 3. Performance comparison with different distance thresholds τ_{dist} on the KITTI dataset.

Setting	Car	Pedestrian	Van	Cyclist	Mean
$\tau_{\text{dist}} = 10cm$	74.9/86.7	**69.4**/93.3	64.1/**78.0**	76.6/94.6	71.6/89.0
$\tau_{\text{dist}} = 30cm$	**76.0/88.6**	68.6/**94.0**	**64.2**/77.7	**77.7/94.8**	**71.8/90.1**
$\tau_{\text{dist}} = 70cm$	74.5/86.8	67.9/92.9	63.8/77.6	76.9/**94.8**	70.8/88.8

Table 4. Comparison of model size and inference time with state-of-the-art methods.

Mehod	# Params	Inference Time	FPS	Suc./Prec.
MBPTrack [2]	1.9M	15ms	67	73.4/84.8
M3SOT [1]	4.3M	48ms	21	75.9/87.4
ChronoTrack	2.9M	24ms	42	76.0/88.6

1. Hyperparamter Choices

We conduct ablation studies on different hyperparameter choices: number of memory tokens K, temperature of memory cycle consistency loss τ_{cycle}, and distance threshold of temporal consistency loss τ_{dist}. In Tab. 1, we find that $K = 32$ is suitable for overall performance. Tab. 2 shows performances for the different temperatures of the memory cycle consistency loss. The temperature τ_{cycle} controls the sharpness of the transition probability distribution during the cyclic walk. As shown, $\tau_{\text{cycle}} = 0.1$ yields the best performance. Tab. 3 varies the distance threshold τ_{dist} in the temporal consistency loss. τ_{dist} is used to discard nearest neighbor pairs whose canonical distance exceeds the threshold to suppress unreliable matches. We find $\tau_{\text{dist}} = 30\,\text{cm}$ provides the best overall performance.

2. Model Size and Inference Time

Tab. 4 compares model size and runtime on the KITTI Car category, evaluated on a single RTX 4090 GPU using official implementations of the compared state of the art methods. ChronoTrack processes a frame in 24 ms (42 FPS) with 2.9M parameters, achieving real time performance and the highest Success/Precision among the compared methods (76.0/88.6).

References

[1] Jiaming Liu, Yue Wu, Maoguo Gong, Qiguang Miao, Wenping Ma, Cai Xu, and Can Qin. M3sot: Multi-frame, multi-field, multi-space 3d single object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3630–3638, 2024. 1

[2] Tian-Xing Xu, Yuan-Chen Guo, Yu-Kun Lai, and Song-Hai Zhang. Mbptrack: Improving 3d point cloud tracking with memory networks and box priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9911–9920, 2023. 1