Supplementary Material for “Temporal-aware Siamese Tracker: Integrate Temporal Context for 3D Object Tracking"

Kaihao Lan, Haobo Jiang, and Jin Xie$^{(\boxtimes)}$

Nanjing University of Science and Technology, Nanjing, China
`{lkh, jiang.hao.bo, csjxie}@njust.edu.cn`

1 Overview

In this supplementary material, we provide more results and have some discussion on our method. In Sec.2, we will provide more results on our method. In Sec.3, we will give some discussion of our method.

2 More results

In this section, we will provide richer results, including quantitative results at different point intervals, discuss the running speed of TAT, more category visualizations of sequence tracking on the different datasets, two tracking videos, and more ablation study result.

2.1 Quantitative result.

We provide the performance of our method at different point intervals. Following V2B [5], we set the point interval of large-size categories (Car and Van) into $[0, 150)$, $[150, 1000)$, $[1000, 2500)$ and $[2500, +\infty)$, and set the point interval of small-size categories (Pedestrian and Cyclist) into $[0, 100)$, $[100, 500)$, $[500, 1000)$ and $[1000, +\infty)$. As shown in Tab.2, we compare our method with the previous state-of-art approaches, including SC3D [4], P2B [7], BAT [10] and V2B [5]. It can be seen that our method achieves the best performance on all metrics across all point intervals for all categories (except only the $[2500, +\infty)$ point interval for the Car category).

2.2 Running speed.

Online Tracking Acceleration. In the online tracking process (test phase), to ensure the inference speed of the network, we will save the template scores and related initial feature maps of historical frames. Therefore, we only need to calculate the template score and PointNet++ [6] inference at the current frame each time, which can save a lot of repeated inference time.
Speed comparison. We use FPS (frames per second) to measure the running

speed of the trackers. For a fair comparison, we test the FPS of the tracker runs on the car category of the KITTI dataset with a TITAN RTX GPU. Our method achieve 20 FPS. SC3D [4], P2B [7], and V2B [5] can achieve 3 FPS, 28 FPS and 22 FPS, respectively.

2.3 Sequence visualization.

We provide visualization results of sequence tracking on more categories, which validate the robust tracking ability of our method. As shown in Fig.2, we show the sequence visualization of different categories on the KITTI [3] dataset, including Car, Pedestrian, Van and Cyclist. As shown in Fig.3, we show the sequence visualization of different categories on the nuScenes [1] dataset, including Car, Pedestrian, Truck and Bicycle. Finally, as shown in Fig.4, we show the sequence visualization of Vehicle and Pedestrian categories on the waymo open dataset [8].

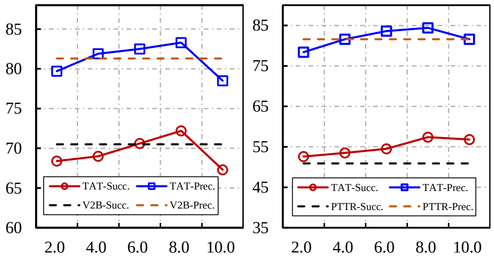
2.4 Tracking video.

In the supplementary material, we have provided two additional visualization videos. Thanks to the temporal context, for car, our method enables more refined tracking. For pedestrian, our method can accurately locate targets in complex scenes.

2.5 Ablation study

Different numbers of templates. As shown in Fig. 1, for the different numbers of templates, an appropriate number can bring better performance. This is because too few templates can not provide enough temporal context, and too many templates will cause internal disturbances and increase time overhead, so the number of templates k selected as 8 is a better choice.

Hyperparameters of attention module. As shown in Tab. 1, we investigate the impact on the performance of some hyperparameters regarding the attention module, including the number of iterations and the number of multi-head attention heads. For the number of iterations, we can see that increasing the number of iterations of the attention module can improve the performance. Taking the pedestrian category as an example, when the number of iterations is 1, 2 and 3 respectively, the *Success* indicators go from 55.8 to 57.4, and to 58.3, which proves that more iterations can extract deeper temporal features. However, more iterations necessarily require more network inference time. For balance, we choose the iteration number m as 2. For the number of multi-head attention heads, a moderate number can effectively improve performance, and too few or too many heads will cause a certain decline in performance.



(a) Car (b) Pedestrian

Fig. 1. Ablation study on different number of templates in template set. Previous SOTA method are marked in the respective graphs.

	Metrics	*Success*/*Precision*	
	Param.	Car	Pedestrian
Iters.	$m = 1$	69.5/80.4	55.8/80.5
	$m = 2$	**72.2**/**83.3**	57.4/**84.4**
	$m = 3$	**72.2**/82.8	**58.3**/83.8
Heads	$n = 1$	70.1/80.9	53.1/81.6
	$n = 2$	70.1/81.0	56.8/**85.5**
	$n = 4$	**72.2**/**83.3**	**57.4**/84.4
	$n = 8$	70.5/82.0	54.8/83.0

Table 1. Ablation study on different hyperparameters of attention module.

3 Discussion

3.1 Differences from 2D temporal-aware Siamese trackers.

In 2D object tracking, a small number of works [2,9] have noted exploiting historical tracking information, so we need to clarify the novelty of our work: 1. We are the first work to use temporal context in 3D object tracking. 2. Unlike the 2D methods TrDiMP [9] and TCTrack [2] which directly use the nearest k frames as templates, we use a template select module to build higher quality template set. 3. In the feature aggregation module, we use an RNN-based fusion module to balance the impact of different temporal, however, TrDiMP and TCTrack treat templates from different temporal as equal.

3.2 Failure cases.

As shown in Fig.5, we give typical failure cases on the Car, Van and Pedestrian categories on the KITTI dataset, respectively. For the large-size target (Car and Van), the failures are mostly due to too sparse point cloud. Note that, there are even without any points inside the GT BBox. For the small-size target (Pedestrian), the failures are mostly due to the fact that it is more difficult to predict a high-quality rotation angle for a Pedestrian target compared to a large target such as Car or Van.

3.3 Limitation and Feature work.

In addition to the limitations mentioned in Sec.3.2, comparing the quantitative results on the KITTI, nuScenes and waymo open dataset provided in the main text, we can see that the advantages of our method on the KITTI and waymo open dataset are comprehensive and huge, while on the nuScenes dataset, although our method also achieves the best performance, it cannot form a comprehensive advantage. The biggest reason is that compared to the 64-line LiDAR used by KITTI and waymo open dataset, the scanner of nuScenes dataset uses 32-line LiDAR, which makes the scene point cloud more sparse. Therefore, when we construct the template set, the overall quality is not high, and it is difficult to provide a stable and rich temporal context. In order to solve such difficulties, in future work, we consider using multimodal data fusion (Point cloud and RGB image) to address the robustness of the algorithm in sparse point cloud scenes, and introduce motion information to improve the accuracy of target rotation angle prediction.

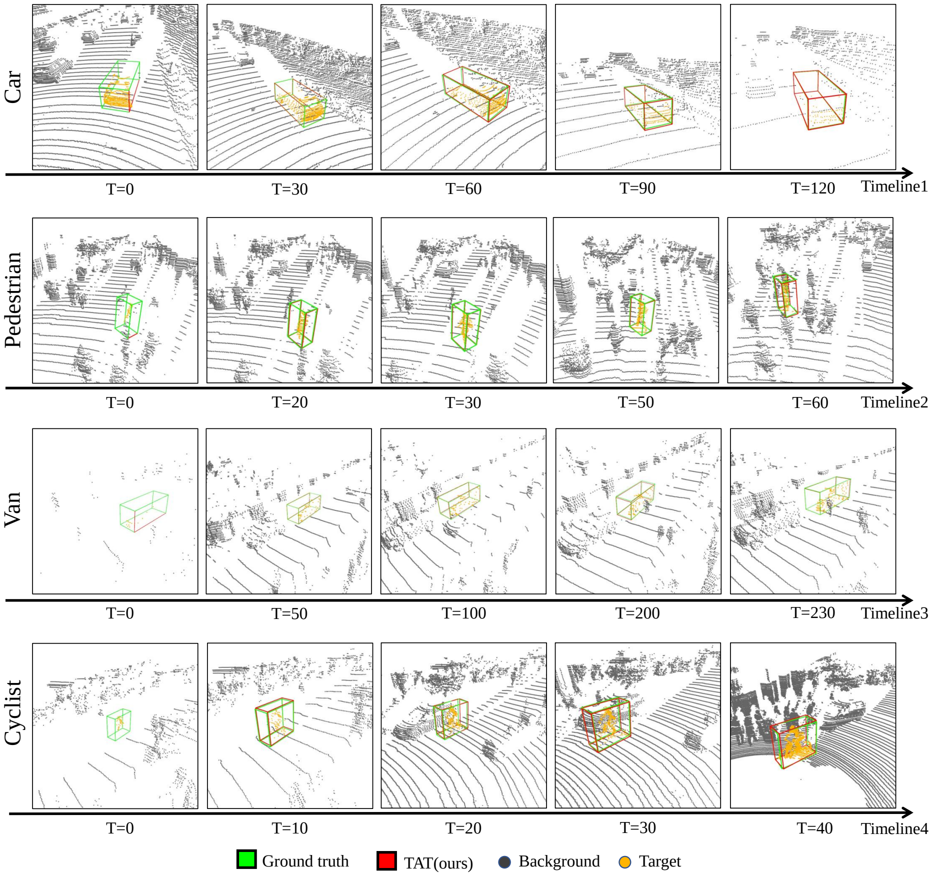


Fig. 2. Sequence visualization results of our method of different categories on the **KITTI** dataset, including Car, Pedestrian, Van, and Cyclist. We mark the points of target object in orange for better indentification from the background.

Table 2. The results of *Success/Precision* of different methods at different point intervals in the KITTI dataset. **Bold** and underline denote the best performance and the second-best performance, respectively. “Mean” denotes the average results of four categories. “*improvement*” denotes the performance improvement between our method and the previous best performance method.

Method Total Frame Number	Car 6424	Pedestrian 6088	Van 1248	Cyclist 308	Mean 14068
Point Interval	[0, 150)	[0, 100)	[0, 150)	[0, 100)	
Frame Number	3293	1654	734	59	5740
SC3D [4]	37.9/53.0	20.1/42.0	36.2/48.7	50.2/69.2	32.7/49.4
P2B [7]	56.0/70.6	33.1/58.2	41.1/46.3	24.1/28.3	47.2/63.5
BAT [10]	60.7/75.5	48.3/77.1	41.5/47.4	25.3/30.5	54.3/71.9
V2B [5]	64.7/77.4	50.8/74.2	46.8/55.1	30.4/37.2	58.0/73.2
TAT (ours)	**66.2/79.1**	**57.1/86.9**	**52.4/63.3**	**65.7/89.8**	**61.8/79.4**
improvment	+1.5/+1.7	+6.3/+9.8	+5.6/+8.2	+15.5/+20.6	+3.8/+6.2
Point Interval	[150, 1000)	[100, 500)	[150, 1000)	[100, 500)	
Frame Number	2156	3112	333	145	5746
SC3D [4]	36.1/53.1	17.7/38.2	38.1/53.3	44.7/76.0	26.5/45.6
P2B [7]	62.3/78.6	25.1/46.0	41.7/50.5	35.4/46.5	40.3/58.5
BAT [10]	71.8/83.9	45.0/71.2	44.0/51.6	41.5/52.2	54.8/74.3
V2B [5]	77.5/87.1	46.8/72.0	51.2/59.6	44.4/53.9	58.5/76.5
TAT (ours)	**78.0/87.2**	**55.2/82.0**	**65.9/76.6**	**77.4/95.1**	**64.9/84.0**
improvment	+0.5/+0.1	+8.4/+10.0	+14.7/+17.0	+32.7/+19.1	+6.4/+7.5
Point Interval	[1000,2500)	[500,1000)	[1000,2500)	[500,1000)	
Frame Number	693	1071	78	42	1884
SC3D [4]	33.8/48.7	15.0/37.1	35.9/50.3	34.9/69.5	23.2/42.6
P2B [7]	51.9/68.1	28.4/49.9	40.7/49.7	25.7/37.7	37.5/56.3
BAT [10]	69.1/81.0	35.2/61.7	50.3/61.3	34.9/48.7	48.3/68.5
V2B [5]	72.3/81.5	47.2/74.3	61.3/67.8	42.3/52.0	56.9/76.2
TAT (ours)	**79.1/88.7**	**57.9/84.1**	**70.1/77.8**	**75.8/94.9**	**66.6/85.8**
improvment	+6.8/+7.2	+10.7/+9.8	+8.8/+10.0	+33.5/+25.4	+9.7/+9.6
Point Interval	[2500,+∞)	[1000,+∞)	[2500,+∞)	[1000,+∞)	
Frame Number	282	251	103	62	698
SC3D [4]	23.7/35.3	14.5/35.3	30.5/42.4	27.7/64.2	21.8/38.9
P2B [7]	43.8/61.8	27.1/49.1	33.8/39.7	24.6/34.2	34.6/51.5
BAT [10]	61.6/72.9	32.6/58.6	48.2/57.9	26.7/37.9	46.1/62.4
V2B [5]	**82.2/90.1**	53.8/82.6	60.9/65.9	41.2/50.4	65.2/80.3
TAT (ours)	81.4/88.9	**62.4/89.6**	**74.2/80.7**	**73.6/94.1**	**72.8/88.4**
improvment	-0.8/-1.2	+8.6/+7.0	+13.3/+14.8	+32.4/+29.9	+7.6/+8.1

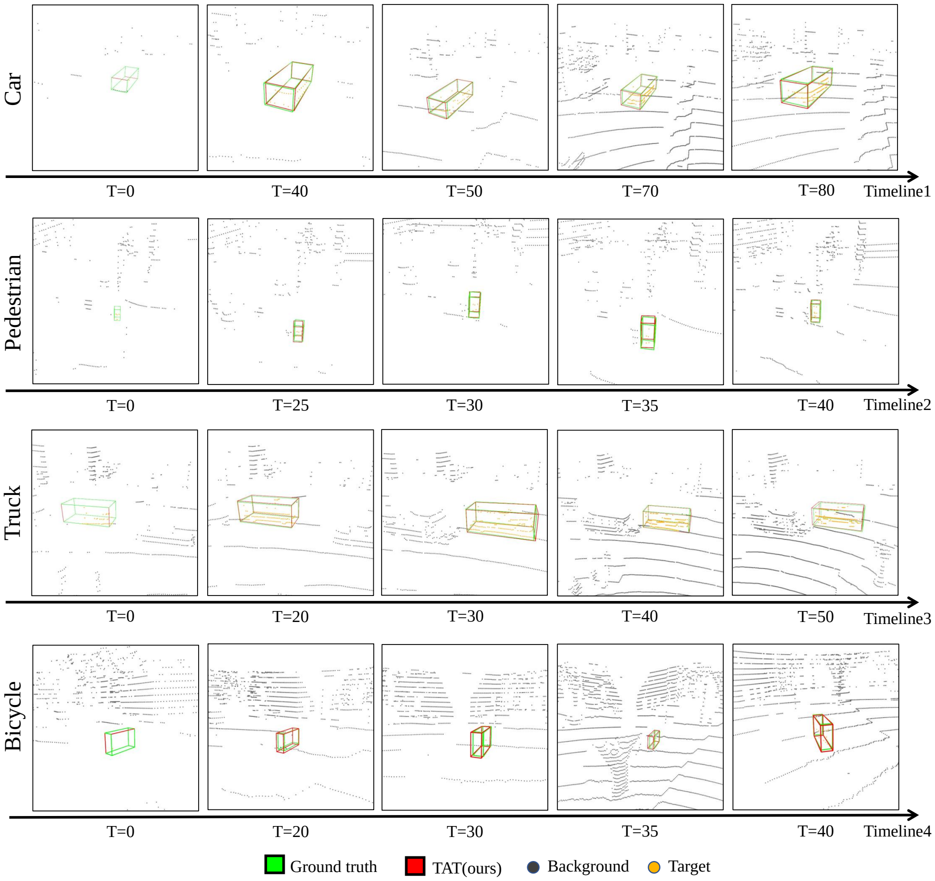


Fig. 3. Sequence visualization results of our method of different categories on the **nuScenes** dataset, including Car, Pedestrian, Truck, and Bicycle. We mark the points of target object in orange for better indentification from the background.

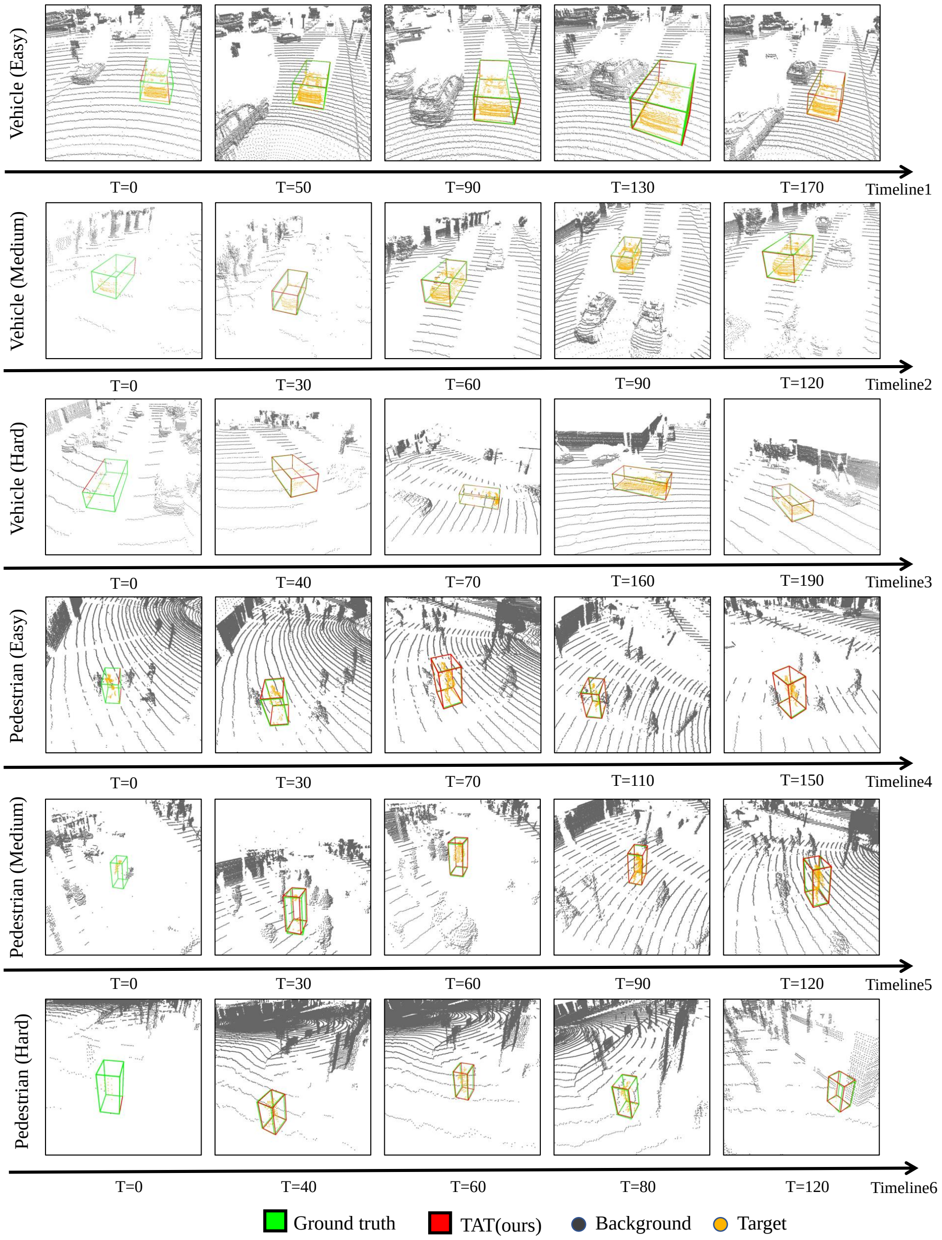


Fig. 4. Sequence visualization results of our method on the **waymo open** dataset, including sequences of varying difficulty on the Vehicle and Pedestrian categories. We mark the points of target object in orange for better indentification from the background.

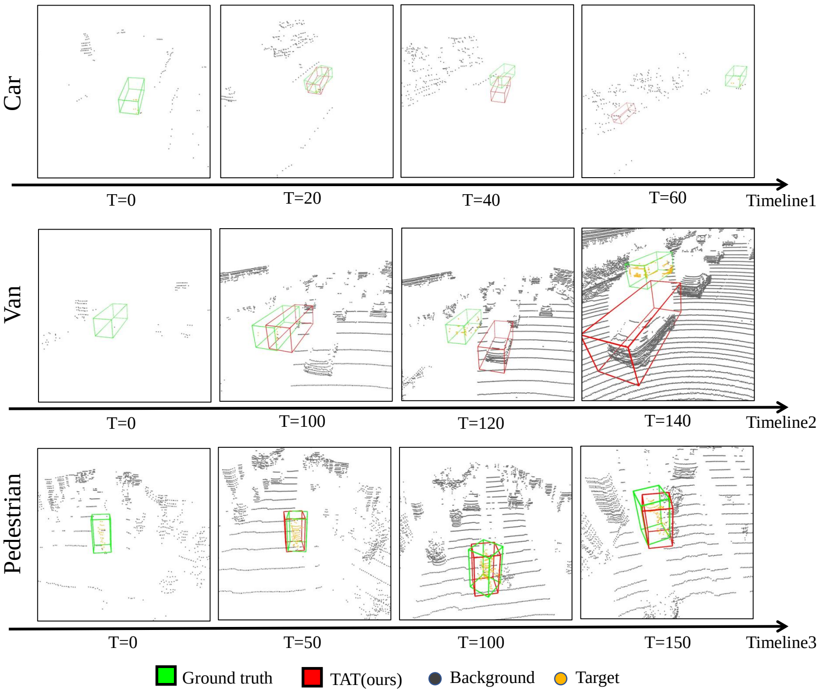


Fig. 5. Three failures in the KITTI dataset. We mark the points of target object in orange for better indentification from the background.

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
2. Cao, Z., Huang, Z., Pan, L., Zhang, S., Liu, Z., Fu, C.: Tctrack: Temporal contexts for aerial tracking. In: CVPR (2022)
3. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012)
4. Giancola, S., Zarzar, J., Ghanem, B.: Leveraging shape completion for 3d siamese tracking. In: CVPR (2019)
5. Hui, L., Wang, L., Cheng, M., Xie, J., Yang, J.: 3d siamese voxel-to-bev tracker for sparse point clouds. In: NeurIPS (2021)
6. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: Deep hierarchical feature learning on point sets in a metric space. In: NeurIPS (2017)
7. Qi, H., Feng, C., Cao, Z., Zhao, F., Xiao, Y.: P2b: Point-to-box network for 3d object tracking in point clouds. In: CVPR (2020)
8. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR (2020)
9. Wang, N., Zhou, W., Wang, J., Li, H.: Transformer meets tracker: Exploiting temporal context for robust visual tracking. In: CVPR (2021)
10. Zheng, C., Yan, X., Gao, J., Zhao, W., Zhang, W., Li, Z., Cui, S.: Box-aware feature enhancement for single object tracking on point clouds. In: ICCV (2021)