Supplementary Material for "3D Siamese Transformer Network for Single Object Tracking on Point Clouds"

Le Hui, Lingpeng Wang, Linghua Tang, Kaihao Lan, Jin Xie*, and Jian Yang*

Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education
Jiangsu Key Lab of Image and Video Understanding for Social Security
PCA Lab, School of Computer Science and Engineering
Nanjing University of Science and Technology, China
{le.hui,cslpwang,tanglinghua,lkh,csjxie,csjyang}@njust.edu.cn

A Overview

In this document, we provide additional quantitative results, visualization, and ablation study for our 3D Siamese Transformer network for single object tracking on point clouds.

B Quantitative Results

Results on the nuScenes dataset. In addition to the generalization ability of our method on the nuScenes dataset. We also train our method on the training set of the nuScenes dataset and test it on its validation set. Since nuScenes only annotates key frames, we follow V2B [2] and use the official toolkit to insert ground truth for unannotated frames. As shown in Tab. 1, we report the results on the nuScenes dataset. Note that the results are computed on the key frames of the validation set. It can be seen that our method achieves the best performance on the mean results of four categories. Please note that the network may not accurately recognize the target due to the low-quality interpolated ground truth of the unannotated frames. Thus, the nuScenes dataset is much hard than the KITTI dataset.

Different template generation schemes. As shown in Tab. 2, we report the evaluation results of different methods on different template generation schemes. It can be seen that our method achieve the best performance in most cases. Note that V2B [2] uses shape completion to enhance the shape information of the target, so it can achieve good performance by using the previous result as the template.

Results of different point intervals. As shown in Tab. 4, we show the results of different methods at different point intervals on the KITTI dataset. It can be seen that our method achieves the best average performance on each interval for four categories.

* Corresponding authors.

Table 1. The performance of different methods on the nuScenes dataset. "Mean" denotes the average results of four categories.

Method	Success					Precision				
Category	Car	Pedestrian	Truck	Bicycle	Mean	Car	Pedestrian	Truck	Bicycle	Mean
Frame Num.	15578	8019	3710	501	27808	15578	8019	3710	501	27808
SC3D [1]	24.5	13.8	32.5	16.6	22.3	25.9	14.7	30.6	18.8	23.2
P2B [3]	32.8	19.2	16.2	19.7	26.4	35.2	26.6	11.1	26.6	29.3
BAT [4]	26.5	19.4	16.5	17.8	23.0	28.8	28.2	10.6	22.8	27.9
V2B [2]	36.2	20.1	28.7	20.3	30.3	38.2	27.4	23.8	27.5	33.0
STNet (ours)	**36.5**	**20.3**	**32.8**	**20.5**	**31.0**	**38.4**	**28.6**	**30.8**	**28.5**	**34.3**

Table 2. The results of different template generation schemes of different methods in the car category on the KITTI dataset.

Scheme	Success					Precision				
	SC3D [1]	P2B [3]	BAT [4]	V2B [2]	STNet (ours)	SC3D [1]	P2B [3]	BAT [4]	V2B [2]	STNet (ours)
The First GT	31.6	46.7	51.8	67.8	**70.8**	44.4	59.7	65.5	79.3	**82.4**
Previous result	25.7	53.1	59.2	**70.0**	66.0	35.1	68.9	75.6	**81.3**	76.6
First & Previous	34.9	56.2	60.5	70.5	**71.2**	49.8	72.8	77.7	81.3	**84.0**
All previous results	41.3	51.4	55.8	69.8	**73.3**	57.9	66.8	71.4	81.2	**85.4**

C Visualization

Attention maps. As shown in Fig. 1, we plot the attention map generated by our method on the KITTI dataset, including car, pedestrian, van, and cyclist. The points marked with the red color can obtain the high attentional weights. It can be found that our method can accurately focus the target in the search area on four categories.

Different categories. As shown in Fig. 2, we visualize the tracking results in the four categories on the KITTI dataset, including car, pedestrian, van, and cyclist, respectively. It can be found that our method can effectively localize the target in the search area with complex background.

Failure cases. As shown in Fig. 3, we show the failure cases of our method in the four categories on the KITTI dataset, including car, pedestrian, van, and cyclist. For the car and truck categories, due to few point clouds on the targets, it is difficult to learn high-quality point features to localize the target from the background, so our method fails in these cases. For the pedestrian and cyclist categories, it could fail in two very close pedestrians or cyclists.

D Ablation Study

Positional embedding. As shown in Tab. 3, we show the results of the ablation study on positional embedding. Specifically, we remove all positional embedding in our network and experiment in the car category on the KITTI dataset. It can be seen that the performance is greatly reduced without positional embedding.

Different correlation learning schemes. In the coarse-to-fine correlation network, we only use the template to update the search area. Here, we use the

Table 3. The ablation study results of different components in the car category on the KITTI dataset.

Method	Success	Precision
w/o positional embedding	69.1	80.5
w/ positional embedding	**72.1**	**84.0**
search area updates template and vice versa	68.6	79.7
only template updates search area	**72.1**	**84.0**

search area to update the template and vice versa. As shown in Tab. 3, we report the results of different correlation learning schemes in the car category on the KITTI dataset. It can be seen that the results of using the search area to update the template are reduced. Due to the sparsity of the point cloud, using the search area to update the template will result in an inaccurate template feature. Conversely, using an inaccurate template to update the search area will cause the search area to be inaccurate. Thus, our coarse-to-fine correlation network achieves better performance.

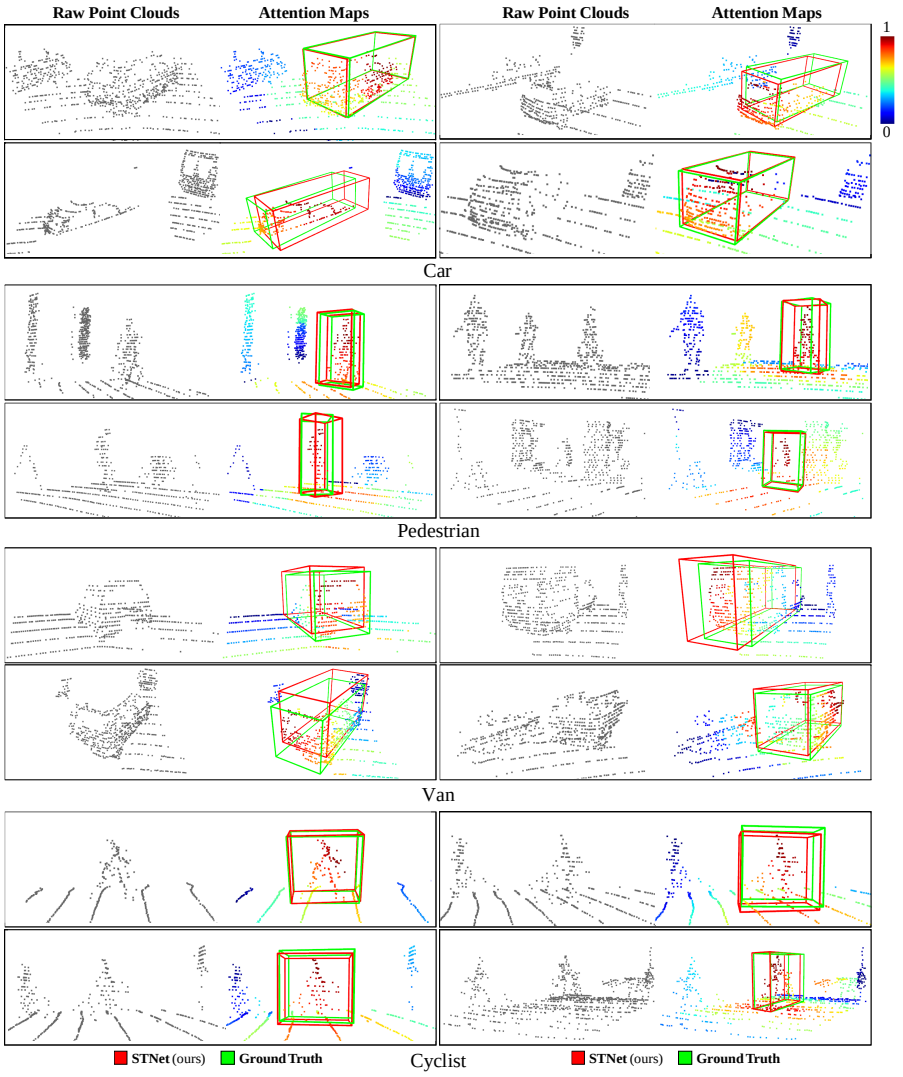

Fig. 1. The attention maps generated by our method on the KITTI dataset, including car, pedestrian, van, and cyclist.

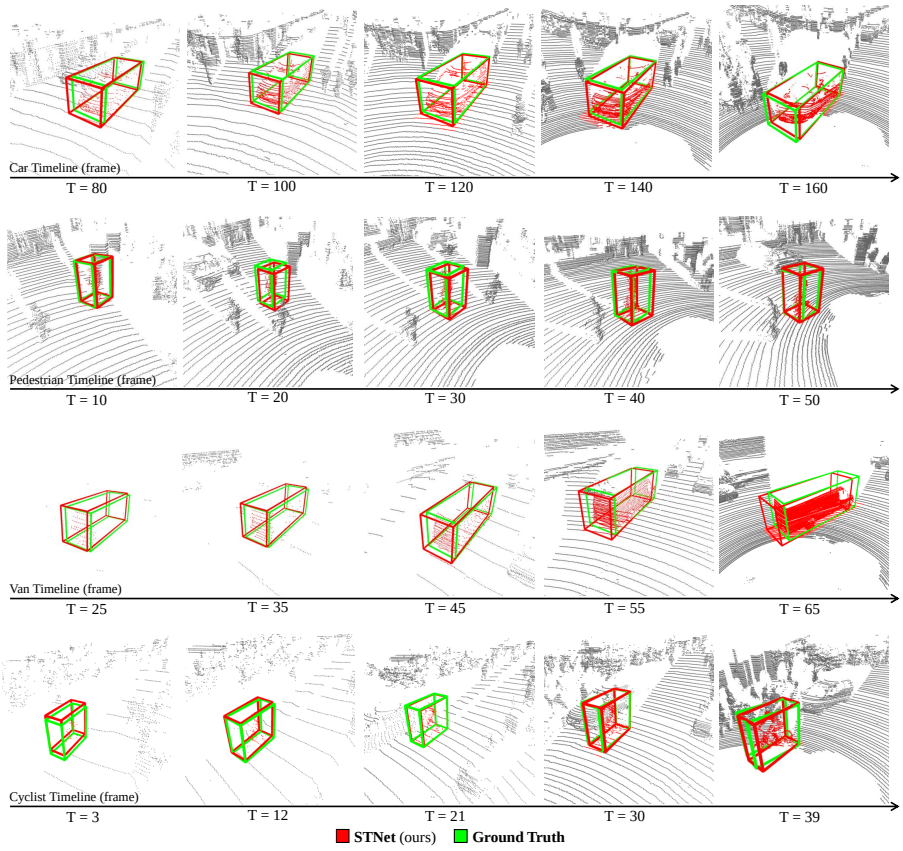

Fig. 2. Visualization results of our method in the four categories on the KITTI dataset, including car, pedestrian, van, and cyclist.

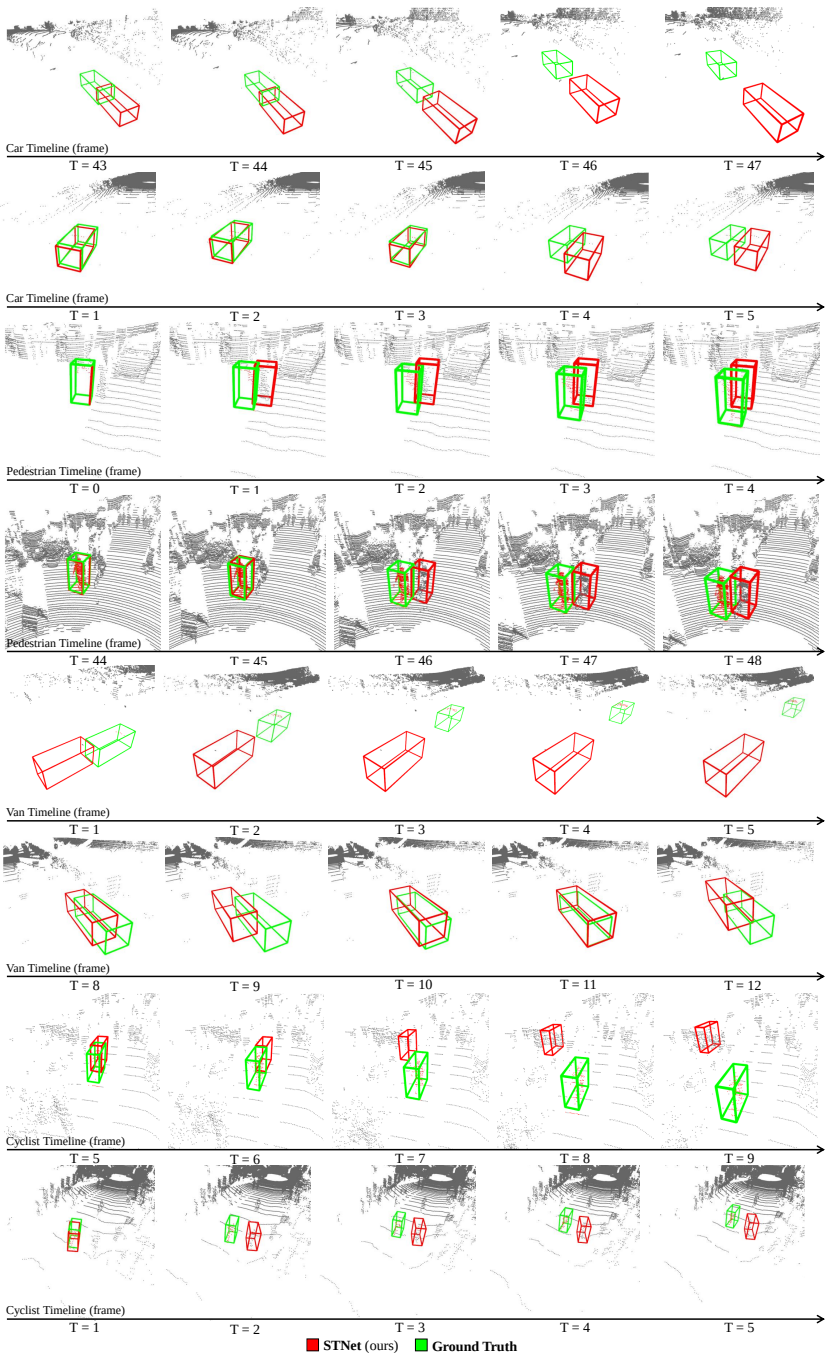

Fig. 3. Failure cases of our method in the four categories on the KITTI dataset, including car, pedestrian, van, and cyclist.

Table 4. The results of different methods at different point intervals on the KITTI dataset. "Mean" denotes the average results of four categories.

Methods	Success					Precision				
	Car	Pedestrian	Van	Cyclist	Mean	Car	Pedestrian	Van	Cyclist	Mean
Frame Num.	6424	6088	1248	308	14068	6424	6088	1248	308	14068
Interval	[0, 150)	[0, 100)	[0, 150)	[0, 100)		[0, 150)	[0, 100)	[0, 150)	[0, 100)	
Frame Num.	3293	1654	734	59	5740	3293	1654	734	59	5740
SC3D [1]	37.9	20.1	36.2	50.2	32.7	53.0	42.0	48.7	69.2	49.4
P2B [3]	56.0	33.1	41.1	24.1	47.2	70.6	58.2	46.3	28.3	63.5
BAT [4]	60.7	48.3	41.5	25.3	54.3	75.5	77.1	47.4	30.5	71.9
V2B [2]	64.7	**50.8**	46.8	30.4	58.0	77.4	74.2	55.1	37.2	73.2
STNet (ours)	**66.3**	50.1	**52.6**	**68.0**	**59.9**	**79.9**	**77.7**	**66.0**	**90.8**	**77.6**
Interval	[150, 1k)	[100, 500)	[150, 1k)	[100, 500)		[150, 1k)	[100, 500)	[150, 1k)	[100, 500)	
Frame Num.	2156	3112	333	145	5746	2156	3112	333	145	5746
SC3D [1]	36.1	17.7	38.1	44.7	26.5	53.1	38.2	53.3	76.0	45.6
P2B [3]	62.3	25.1	41.7	35.4	40.3	78.6	46.0	50.5	46.5	58.5
BAT [4]	71.8	45.0	44.0	41.5	54.8	83.9	71.2	51.6	52.2	74.3
V2B [2]	77.5	46.8	51.2	44.4	58.5	87.1	72.0	59.6	53.9	76.5
STNet (ours)	**77.9**	**48.6**	**63.6**	**75.9**	**61.1**	**87.8**	**75.5**	**77.4**	**94.8**	**80.7**
Interval	[1k,2.5k)	[500,1k)	[1k,2.5k)	[500,1k)		[1k,2.5k)	[500,1k)	[1k,2.5k)	[500,1k)	
Frame Num.	693	1071	78	42	1884	693	1071	78	42	1884
SC3D [1]	33.8	15.0	35.9	34.9	23.2	48.7	37.1	50.3	69.5	42.6
P2B [3]	51.9	28.4	40.7	25.7	37.5	68.1	49.9	49.7	37.7	56.3
BAT [4]	69.1	35.2	50.3	34.9	48.3	81.0	61.7	61.3	48.7	68.5
V2B [2]	72.3	47.2	61.3	42.3	56.9	81.5	74.3	67.8	52.0	76.2
STNet (ours)	**79.3**	**51.3**	**69.6**	**75.0**	**62.8**	**89.6**	**79.5**	**79.0**	**94.2**	**83.5**
Interval	[2.5k,+∞)	[1k,+∞)	[2.5k,+∞)	[1k,+∞)		[2.5k,+∞)	[1k,+∞)	[2.5k,+∞)	[1k,+∞)	
Frame Num.	282	251	103	62	698	282	251	103	62	698
SC3D [1]	23.7	14.5	30.5	27.7	21.8	35.3	35.3	42.4	64.2	38.9
P2B [3]	43.8	27.1	33.8	24.6	34.6	61.8	49.1	39.7	34.2	51.5
BAT [4]	61.6	32.6	48.2	26.7	46.1	72.9	58.6	57.9	37.9	62.4
V2B [2]	82.2	53.8	60.9	41.2	65.2	90.1	82.6	65.9	50.4	80.3
STNet (ours)	**83.1**	**59.3**	**70.2**	**71.2**	**71.5**	**91.0**	**87.7**	**76.1**	**93.4**	**87.8**

References

1. Giancola, S., Zarzar, J., Ghanem, B.: Leveraging shape completion for 3D Siamese tracking. In: CVPR (2019)
2. Hui, L., Wang, L., Cheng, M., Xie, J., Yang, J.: 3D Siamese voxel-to-BEV tracker for sparse point clouds. In: NeurIPS (2021)
3. 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)
4. 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)