3D Single-object Tracking in Point Clouds with High Temporal Variation
Supplementary Material

Qiao Wu[1], Kun Sun[2], Pei An[3], Mathieu Salzmann[4], Yanning Zhang[1], and Jiaqi Yang[1]

[1] Northwestern Polytechnical University
[2] China University of Geosciences, Wuhan
[3] HuaZhong University of Science and Technology
[4] École Polytechnique Fédérale de Lausanne
qiaowu@mail.nwpu.edu.cn, jqyang@nwpu.edu.cn

1 Implementation Details

KITTI-HV. KITTI-HV has the same size as the original KITTI. We can simply construct KITTI-HV with a few lines of code as in Algorithm 1. We set the intervals non-linearly ([2,3,5,10]) instead of the traditional linear setting ([2,4,6,8]). Thus, we have denser tests in point cloud variations close to smooth scenarios (comparing [2,3,5] to [2,4,6]) for a fairer comparison with the existing methods.

Algorithm 1 KITTI-HV Pseudocode, Python-like

```
# HV-tracklets: tracklets in KITTI-HV
for tracklet in KITTI: # read tracklets in KITTI
   for i in range(min(len(tracklet),interval)):
   # starting at different frame
       temp_tracklet = tracklet[i::interval]
       # sampling at frame intervals
       HV-tracklets.append(temp_tracklet)
return HV-tracklets
```

Search areas. Former trackers [10, 14, 16, 17] determine the search area by enlarging the target bounding box in wide and length at the last frame by 2 meters offset. We follow their strategy to generate the search area with enlargement offsets on KITTI [3] as shown in Tab. 1. We first statistically analyze the moving distance in the xy-plane of 'Car' on KITTI with different frame intervals as shown in Tab. 2. We evaluate the performance of BAT [16] and M2-Track [17] with different bounding box enlargement offsets in 5 frame intervals on KITTI-HV. The enlargement offsets are generated by slightly increasing the moving distances under different quantiles in Tab. 2. As illustrated in Tab. 3, BAT and M2-Track reach the peak at the enlargement offset of 4 meters and 6 meters, respectively. Thus, we choose the moving distances between quantiles of 50%

and 75% as the enlargement offset for all the frame intervals and categories. Following [10, 14, 16, 17], we randomly sample 1024 points in the search area as the input of the backbone.

Observation angle. Instead of the original radian $\in \mathbb{R}^1$, we use the sine and cosine values $\in \mathbb{R}^2$ to represent the observation angle.

Ablation details. We construct the vanilla cross-attention and self-attention in the ablation experiment as shown in Fig. 1 (a) and Fig. 2 (a). Compared to the BEA, vanilla cross-attention removes the expansion branch and assigns H heads for the base branch. For the vanilla self-attention, we directly project $\hat{X}_{l-1}$ to K and V.

Table 1: Bounding box enlargement offsets (meter) in different frame intervals and categories on KITTI for generating search areas.

Frame Intervals	Car	Pedestrian	Van	Cyclist
1	2	2	2	2
2	2	2	3	2
3	3	2	3	2
5	4	2	5	3
10	7	3	8	4

Table 2: Quantiles of Car's moving distance in the xy-plane with different frame intervals on the training set of KITTI.

Quantile	1	2	3	5	10
25%	0.32	0.52	0.57	0.44	0.00
50%	0.79	1.55	2.28	3.61	5.51
75%	1.07	2.11	3.12	5.07	9.28
95%	2.26	4.38	6.06	9.17	15.38
99.73%	3.46	6.90	10.30	17.07	32.88
100%	14.56	15.48	16.49	19.21	36.56
Average	0.81	1.57	2.28	3.53	5.78

Table 3: Performance of BAT and M2-Track in different search area sizes on 'Car' of KITTI-HV with 5 frame intervals. We determine the search area size by enlarging the object bounding box in width and length with an offset.

Offset (m)	20		18		10		6		4	
Method	Succ.	Prec.	Succ.	Prec.	Succ.	Prec.	Succ.	Prec.	Succ.	Prec.
BAT [16]	16.62	16.88	17.02	17.18	25.27	27.70	35.05	40.25	**44.13**	**51.11**
M2-Track [17]	16.53	14.59	21.69	22.51	43.12	50.80	**52.64**	**61.58**	50.87	58.56

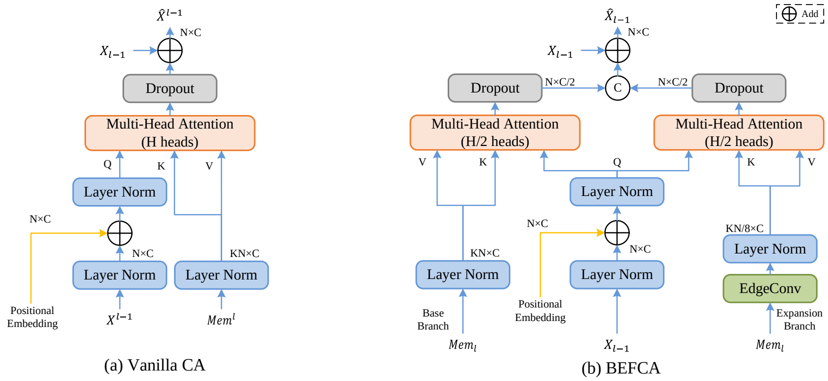

Fig. 1: (a) Vanilla Cross-Attention (CA) and (b) Base-Expansion Feature Cross-Attention (BEA).

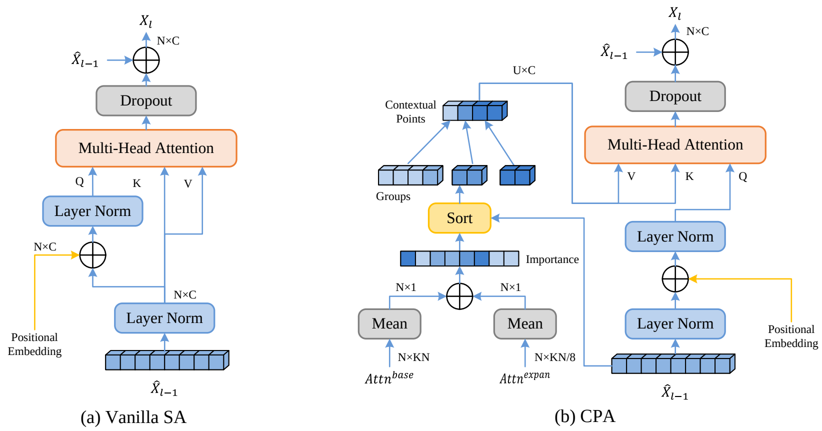

Fig. 2: (a) Vanilla Self-Attention (CA) and (b) Contextual Point Guided Self-Attention (CPA).

2 More Comparisons

Comparison with latest SOTAs. In Tab. 4, we compare HVTrack with the latest SOTAs on KITTI. There still exists a performance gap compared to them. M3SOT [8] extends MBPTrack [15] via the SpaceFormer and achieves better performance. Thus, we report the stronger tracker M3SOT in high temporal variation scenarios in Tab. 5 to validate the effectiveness of HVTrack. HVTrack

Table 4: Comparison with the most recent SOTAs on KITTI.

Category	Car	Pedestrian	Van	Cyclist	Mean	Params (MB)
MBPTrack [15]	73.4/84.8	**68.6**/**93.9**	**61.3**/72.7	**76.7**/**94.3**	70.3/87.9	7.39
M3SOT [8]	**75.9**/**87.4**	66.6/92.5	59.4/**74.7**	70.3/93.4	**70.3**/**88.6**	16.43
HVTrack	68.2/79.2	64.6/90.6	54.8/63.8	72.4/93.7	65.5/83.1	**5.60**

still yields the best results at various intervals, with a notable improvement of 17.2%/21.3% at 5 intervals.

Efficiency. We compare HVTrack with SOTA methods in efficiency on KITTI-HV with 5 frame intervals in Tab. 6. Due to the increased search area, CXTrack shows a 26.5% speed decline compared to the 34 FPS reported in its paper.

Backbone flexibility. As illustrated in Tab. 7, we conduct analysis experiments using different backbones in HVTrack on KITTI-HV with 5 frame intervals. PointNet++ [9] is widely used in former trackers [2, 4–7, 10, 11, 13, 16–18], and GCDNN [12] is employed in [14]. Our HVTrack shows robust performance with different backbones, demonstrating the strong flexibility of our approach. In particular, HVTrack achieves an improvement with 0.7%↑/1.5%↑ on the average in success/precision, confirming the great potential for further improvement.

One pre-trained model. We report the results of KITTI pre-trained models on KITTI-HV in Tab. 8 (top). Our memory module requires rich object pose samples to fit object motion. Thus HVTrack suffers a performance degradation on 'Car'. However, the performance improvement on 'Pedestrian' proves the effectiveness of HVTrack when the object pose distribution changes only slightly. To fully demonstrate the generalizability of HVTrack, we train models in [1,2,3,5,10] intervals together, and test them under different intervals in Tab. 8 (bottom). In contrast to other methods whose performance decreases as the interval grows, HVTrack maintains consistent performance across [1,2,3,5] intervals. This demonstrates the robustness of our method in different temporal variation scenarios.

Waymo-HV. Following the construction of KITTI-HV, we build Waymo-HV for a more comprehensive comparison as illustrated in Tab. 9. Our HVTrack consistently outperforms the state-of-the-art methods [14, 16] across all frame intervals.

NuScenes. Following the setting in M2-Track [17], we evaluate our HVTrack in 4 categories ('Car', 'Truck', 'Trailer' and 'Bus') of the famous nuScenes [1] dataset. The results of SC3D [4], P2B [10], and BAT [16] on NuScenes are provided by M2-Track. CXTrack [14] follows the dataset setting in STNet [7], which is quite different from M2-Track. We train CXTrack on NuScenes using its official code and report the results. As shown in Tab. 10, our method achieves the best performance in success (1.9%↑) and ranks second in precision (0.5%↓). HVTrack surpasses M2-Track in 'Pedestrian' with a great improvement in success (**9.2%↑**) and precision (**6.6%↑**), revealing our excellent ability to handle complex cases. 'Pedestrian' is usually considered to have the largest point cloud variations and proportion of noise, due to the small object sizes and the diversity of body

Table 5: Comparison with the most recent SOTA on KITTI-HV.

Interval	Method	Car	Pedestrian	Van	Cyclist	Mean
2	M3SOT [8]	59.0/67.9	**61.7**/**86.3**	**55.2**/**68.7**	55.1/86.3	59.8/76.3
	HVTrack	**67.1**/**77.5**	60.0/84.0	50.6/61.7	**73.9**/**93.6**	**62.7**/**79.3**
3	M3SOT [8]	46.9/52.6	50.1/**74.0**	**43.3**/**53.7**	32.4/48.1	47.7/61.9
	HVTrack	**66.8**/**76.5**	**51.1**/71.9	38.7/46.9	**66.5**/**89.7**	**57.5**/**72.2**
5	M3SOT [8]	30.5/34.5	31.0/44.0	18.3/21.0	21.6/25.9	29.4/37.2
	HVTrack	**60.3**/**68.9**	**35.1**/**52.1**	**28.7**/**32.4**	**58.2**/**71.7**	**46.6**/**58.5**
10	M3SOT [8]	26.1/26.6	16.2/18.8	17.6/17.1	27.5/26.2	21.1/22.4
	HVTrack	**49.4**/**54.7**	**22.5**/**29.1**	**22.2**/**23.4**	**39.5**/**45.4**	**35.1**/**40.6**

Table 6: Comparison in efficiency with SOTA.

Method	M2-Track [17]	CXTrack [14]	M3SOT [8]	HVTrack
FPS	**42**	25	14	31
Params (MB)	8.54	18.27	16.43	**5.60**

motion. Notably, we achieve **9.1%↑/10.4%↑** improvement in success/precision on average over CXTrack, which has the same backbone and RPN. This gap clearly demonstrates the robustness of our method in regular tracking. However, the performance of HVTrack still drops when dealing with large objects.

NuScenes-HV. As shown in Tab. 11, we compare HVTrack with the state-of-the-art methods on each category of the nuScenes-HV dataset. We construct the high-variation dataset nuScenes-HV for training and testing by setting 2 frame intervals for sampling in the NuScenes dataset. We achieve the best performance in both success (52.4%, 3.8%↑) and precision (62.6%, 2.8%↑) on average. We surpass SOTA trackers in the categories with a large number of samples ('Car', 'Pedestrian', and 'Truck'). However, our performance drops in 'Trailer' and 'Bus', which have a small number of samples. We believe the length of tracklets is another factor that affects the performance of HVTrack on 'Trailer' and 'Bus'. With 2 frame intervals, the average tracklet length of the 'Trailer' is only 11.06 frames on nuScenes-HV, while it is 26.59 frames for the 'Van' on KITTI-HV. With such a short average tracklet length, HVTrack is unable to obtain enough historical information for training and testing, leading to a performance drop. Further, a too short tracklet length is not in line with real-world scenarios. Therefore, we only construct nuScenes-HV with 2 frame intervals.

3 Visualization Results

As illustrated in Fig. 3 and Fig. 4, we visualize our experiment results on KITTI-HV with 5 frame intervals in dense and sparse cases. The 'Car', 'Pedestrian',

Table 7: Analysis experiments of using different backbones in HVTrack on KITTI-HV with 5 frame intervals.

Category Frame Number	Car 6424	Pestrian 6088	Van 1248	Cyclist 308	Mean 14068
DGCNN [12]	**60.3**/**68.9**	35.1/52.1	**28.7**/**32.4**	**58.2**/**71.7**	46.6/58.5
PointNet++ [9]	58.6/66.7	**39.0**/**58.3**	27.5/30.7	57.4/70.9	**47.3**/**60.0**

Table 8: Comparison of different training settings on KITTI-HV.

Training interval(s)	Category	Method	Testing interval 1	2	3	5	10
1	Car	M2-Track	65.5/80.8	**62.8**/**74.4**	**52.5**/**61.0**	**36.1**/**39.8**	**23.5**/**24.5**
		CXTrack	**69.1**/**81.6**	59.4/69.4	51.5/58.4	33.6/36.0	22.5/21.3
		HVTrack	68.2/79.2	59.8/68.2	45.8/51.1	21.2/20.8	18.3/20.2
	Pedestrian	M2-Track	61.5/88.2	58.7/86.5	50.8/74.4	30.7/42.3	16.3/19.5
		CXTrack	**67.0**/**91.5**	**64.9**/**88.0**	56.4/78.7	36.2/48.0	**18.3**/**21.2**
		HVTrack	64.6/90.6	63.6/87.8	**60.5**/**82.6**	**42.7**/**57.6**	16.9/19.6
1,2,3,5,10	Car	M2-Track	57.8/74.2	60.3/**73.7**	57.1/66.7	59.9/68.8	37.5/40.0
		CXTrack	57.8/70.2	51.5/60.3	52.2/58.3	34.9/38.3	25.1/24.6
		HVTrack	**65.6**/**76.5**	**60.3**/69.8	**64.6**/**73.4**	**63.9**/**71.8**	**40.9**/**44.3**
	Pedestrian	M2-Track	53.0/79.2	49.3/70.6	41.9/60.9	37.0/54.7	24.0/30.9
		CXTrack	**60.3**/**84.4**	**60.1**/**84.5**	52.8/73.7	33.2/44.1	17.2/19.6
		HVTrack	56.4/78.9	58.5/81.2	**58.2**/**79.7**	**56.3**/**77.2**	**30.6**/**39.1**

and ‘Cyclist’ in Fig. 3 demonstrate the excellent performance of HVTrack in dealing with the distraction of similar objects and massive noise. Moreover, the success of the sparse cases in Fig. 4 confirms the effective utilization of historical information in our method.

Table 9: Comparison of HVTrack with the state-of-the-art methods on each category of the Waymo-HV dataset.

Frame Interval	Method	Vehicle (185632) Easy	Vehicle Medium	Vehicle Hard	Vehicle Mean	Pedestrian (241168) Easy	Pedestrian Medium	Pedestrian Hard	Pedestrian Mean	Mean (426800)
2	BAT [16]	61.0/68.3	53.3/60.9	48.9/57.8	54.7/62.7	19.3/32.6	17.8/29.8	17.2/28.3	18.2/30.3	34.1/44.4
	CXTrack [14]	63.9/71.1	54.2/62.7	52.1/63.7	57.1/66.1	**35.4**/**55.3**	**29.7**/**47.9**	26.3/44.4	**30.7**/**49.4**	42.2/56.7
	HVTrack(Ours)	**66.2**/**75.2**	**57.0**/**66.0**	**55.3**/**67.1**	**59.8**/**69.7**	34.2/53.5	28.7/**47.9**	**26.7**/**45.2**	30.0/49.1	**43.0**/**58.1**
3	BAT [16]	47.1/52.3	39.8/45.2	35.1/40.6	41.0/46.4	18.2/27.4	15.4/22.8	13.7/19.8	15.9/23.5	26.8/33.5
	CXTrack [14]	59.8/64.7	36.5/40.7	26.7/30.8	42.0/46.5	**28.2**/**41.1**	**21.9**/**33.1**	**16.6**/**25.3**	**22.5**/**33.5**	31.0/39.2
	HVTrack(Ours)	**64.3**/**71.3**	**54.3**/**62.2**	**48.5**/**57.2**	**56.2**/**64.0**	25.7/38.2	18.6/28.2	14.6/22.6	19.9/30.0	**35.7**/**44.8**
5	BAT [16]	47.1/52.4	34.4/38.2	27.9/31.3	37.1/41.3	13.6/18.5	12.4/16.8	10.8/13.8	12.3/16.5	23.1/27.3
	CXTrack [14]	45.9/50.5	27.1/29.2	19.5/21.1	31.7/34.6	**23.0**/32.1	**18.0**/**25.9**	**13.7**/**19.5**	**18.5**/**26.1**	24.2/29.8
	HVTrack(Ours)	**47.1**/**52.3**	**40.1**/**45.4**	**34.3**/**39.4**	**40.9**/**46.1**	22.4/**32.2**	17.5/25.5	13.5/19.3	18.0/26.0	**28.0**/**34.7**
10	BAT	31.7/32.3	23.5/23.7	20.9/21.3	25.7/26.1	10.8/11.9	10.3/11.0	10.3/10.4	10.5/11.1	17.1/17.6
	CXTrack	25.1/23.7	16.3/14.4	14.4/13.1	19.0/17.4	14.1/17.2	12.3/14.2	11.1/11.8	12.6/14.5	15.4/15.8
	HVTrack(Ours)	**36.8**/**39.6**	**26.9**/**28.6**	**22.0**/**23.2**	**29.1**/**31.0**	**16.4**/**20.9**	**14.0**/**17.3**	**12.6**/**14.8**	**14.4**/**17.8**	**20.8**/**23.5**

Table 10: Comparison of HVTrack with the state-of-the-art methods on each category of the NuScenes dataset.

Category Frame Number	Car 64159	Pedestrian 33227	Truck 13587	Trailer 3352	Bus 2953	Mean 117278
SC3D [4]	22.3/21.9	11.3/12.7	30.7/27.7	35.3/28.1	29.4/24.1	20.7/20.2
P2B [10]	38.8/43.2	28.4/52.2	43.0/41.6	49.0/40.1	33.0/27.4	36.5/45.1
BAT [16]	40.7/44.3	28.8/53.3	45.3/42.6	52.6/44.9	35.4/28.0	38.1/45.7
M2-Track [17]	**55.9**/**65.1**	32.1/60.9	**57.4**/**59.5**	57.6/**58.3**	**51.4**/**51.4**	49.2/**62.7**
CXTrack [14]	44.6/50.5	31.5/55.8	51.3/50.7	**59.7**/53.6	42.6/37.3	42.0/51.8
HVTrack	**55.9**/62.9	**41.3**/**67.6**	55.6/55.2	52.0/40.2	36.3/41.6	**51.1**/62.2

Table 11: Comparison of HVTrack with the state-of-the-art methods on each category of the nuScenes-HV dataset. We construct the high-variation dataset nuScenes-HV for training and testing by setting 2 frame intervals for sampling in the NuScenes dataset.

Category Frame Number	Car 64159	Pedestrian 33227	Truck 13587	Trailer 3352	Bus 2953	Mean 117278
P2B [10]	47.5/51.3	23.1/35.0	52.9/51.5	63.6/56.2	40.2/37.2	41.5/46.5
BAT [16]	44.7/48.0	23.1/33.2	52.3/50.9	63.7/57.7	41.6/38.2	39.9/44.2
M2-Track [17]	51.7/60.1	37.8/60.6	55.4/**57.8**	**65.8**/**64.8**	**51.5**/**49.2**	48.6/59.8
CXTrack [14]	50.7/57.6	27.0/43.8	54.3/55.0	62.2/56.5	43.4/40.5	44.5/52.9
HVTrack	**57.0**/**63.4**	**43.1**/**68.2**	**56.0**/56.1	51.7/43.1	31.2/35.2	**52.4**/**62.6**

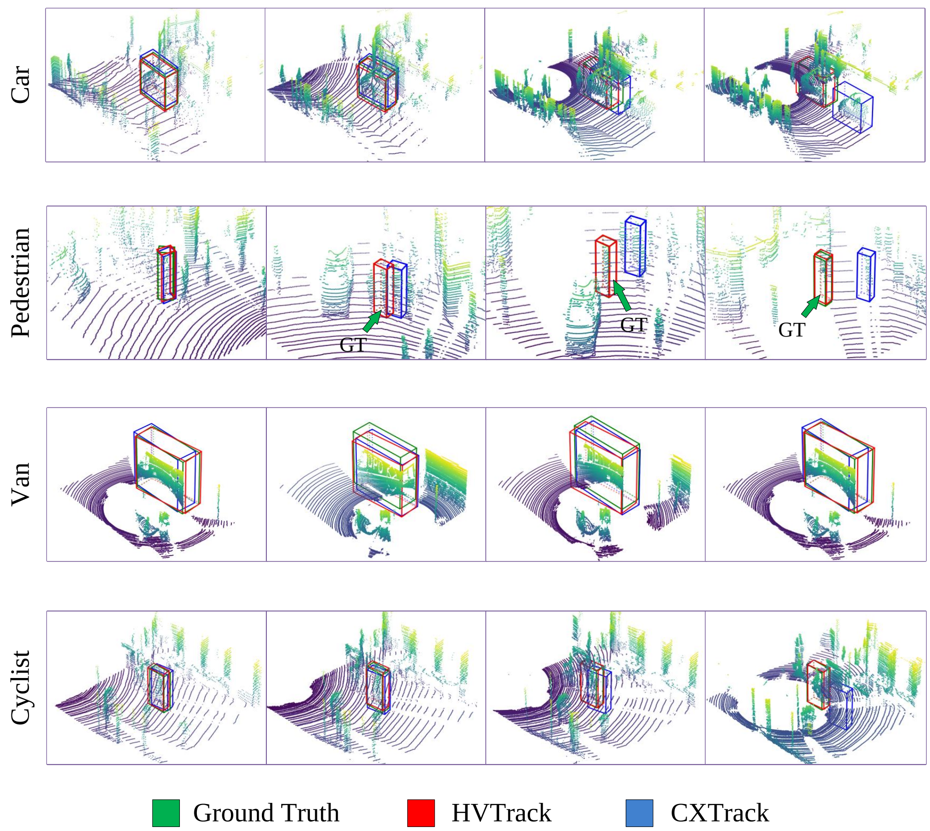

Fig. 3: Visualization results in dense cases on KITTI-HV with 5 frame intervals.

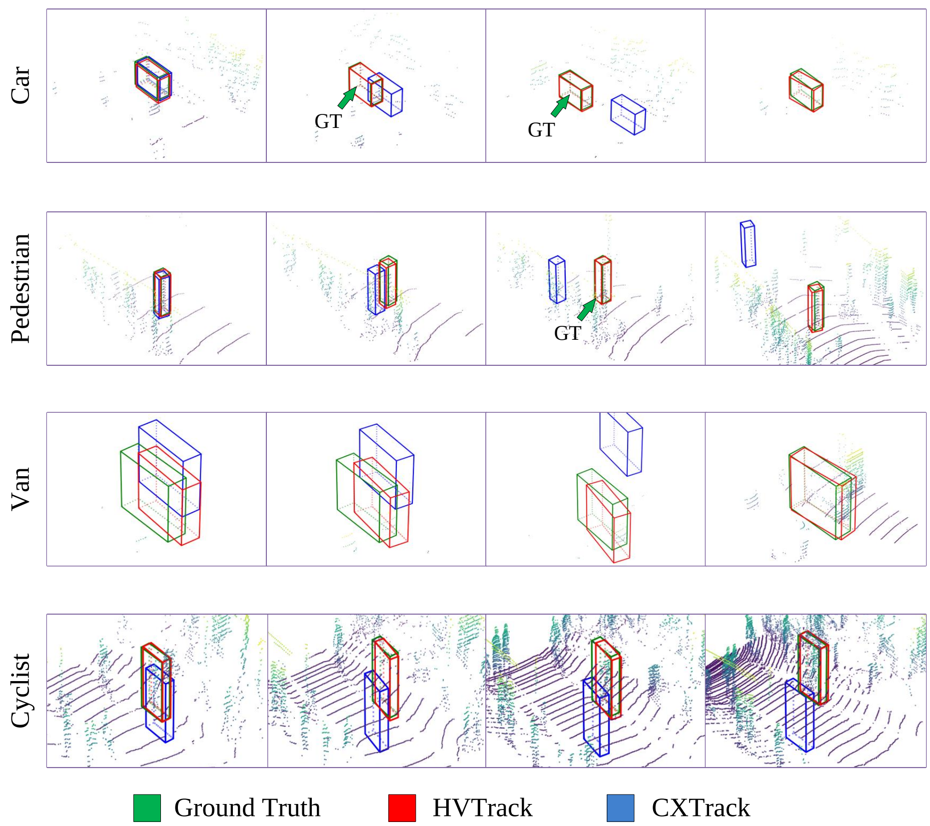

Fig. 4: Visualization results in sparse cases on KITTI-HV with 5 frame intervals.

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. arXiv preprint arXiv:1903.11027 (2019)
2. Cui, Y., Fang, Z., Shan, J., Gu, Z., Zhou, S.: 3d object tracking with transformer. arXiv preprint arXiv:2110.14921 (2021)
3. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361 (2012)
4. Giancola, S., Zarzar, J., Ghanem, B.: Leveraging shape completion for 3d siamese tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1359–1368 (2019)
5. Guo, Z., Mao, Y., Zhou, W., Wang, M., Li, H.: Cmt: Context-matching-guided transformer for 3d tracking in point clouds. In: European Conference on Computer Vision. pp. 95–111. Springer (2022)
6. Hui, L., Wang, L., Cheng, M., Xie, J., Yang, J.: 3d siamese voxel-to-bev tracker for sparse point clouds. Advances in Neural Information Processing Systems **34**, 28714–28727 (2021)
7. Hui, L., Wang, L., Tang, L., Lan, K., Xie, J., Yang, J.: 3d siamese transformer network for single object tracking on point clouds. arXiv preprint arXiv:2207.11995 (2022)
8. Liu, J., Wu, Y., Gong, M., Miao, Q., Ma, W., Xu, C., Qin, C.: M3sot: Multi-frame, multi-field, multi-space 3d single object tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3630–3638 (2024)
9. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
10. Qi, H., Feng, C., Cao, Z., Zhao, F., Xiao, Y.: P2b: Point-to-box network for 3d object tracking in point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6329–6338 (2020)
11. Shan, J., Zhou, S., Fang, Z., Cui, Y.: Ptt: Point-track-transformer module for 3d single object tracking in point clouds. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1310–1316 (2021)
12. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Transactions on Graphics (tog) **38**(5), 1–12 (2019)
13. Wang, Z., Xie, Q., Lai, Y.K., Wu, J., Long, K., Wang, J.: Mlvsnet: Multi-level voting siamese network for 3d visual tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3101–3110 (2021)
14. Xu, T.X., Guo, Y.C., Lai, Y.K., Zhang, S.H.: Cxtrack: Improving 3d point cloud tracking with contextual information. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1084–1093 (2023)
15. Xu, T.X., Guo, Y.C., Lai, Y.K., Zhang, S.H.: Mbptrack: Improving 3d point cloud tracking with memory networks and box priors. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9911–9920 (2023)
16. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13199–13208 (2021)

17. Zheng, C., Yan, X., Zhang, H., Wang, B., Cheng, S., Cui, S., Li, Z.: Beyond 3d siamese tracking: A motion-centric paradigm for 3d single object tracking in point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8111–8120 (2022)
18. Zhou, C., Luo, Z., Luo, Y., Liu, T., Pan, L., Cai, Z., Zhao, H., Lu, S.: Pttr: Relational 3d point cloud object tracking with transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8531–8540 (2022)