

APPENDIX

Anonymous authors

Paper under double-blind review

A DETAILED ANALYSIS OF UNIFIED COMPONENTS

Here, we provide the detailed analysis of the proposed components within our category-unified models, including unified representation network *AdaFormer* (Section A.1), model inputs (Section A.2) and learning objective (Section A.3). All experiments and analysis are conducted using Siamese paradigm (Qi et al., 2020) on the KITTI (Geiger et al., 2012) dataset.

A.1 UNIFIED REPRESENTATION NETWORK: ADAFORMER

The proposed unified representation network *AdaFormer* incorporates a group regression module for learning deformable groups to enable adaptive receptive fields for various object categories, along with a vector-attention mechanism to facilitate feature interaction of points within these deformable groups, ultimately forming a category-unified feature representation.

Tab. 1 presents an ablation study to understand the two sub-components. Benefiting from the adaptive receptive fields achieved by the group regression module, our representation network can learn geometric information of various object categories in a unified manner. Consequently, when this module is removed, average performance drops by 6.9% and 7.0% in terms of *Success* and *Precision*, respectively. It’s noteworthy that the most obvious performance degradation occurs in the Pedestrian category. This is due to the relatively small training samples for the Pedestrian category and the significant differences in shape and size compared to other object categories. To visually understand of how the group regression module works, we provide some visualizations of deformable groups on the Car and Pedestrian categories, as shown in Fig. 1. In addition, when removing the vector-attention mechanism, we employ a feature propagation operator in existing backbone network (Qi et al., 2017a;b) to substitute it. Tab. 1 demonstrates that the vector-attention mechanism plays a crucial role in promoting the learning of a unified feature representation.

Table 1: Ablation study of unified representation network. *Success / Precision* are used for evaluation. **Bold** denote the best performance.

Group Regression Module	Vector-Attention Mechanism	Car [6,424]	Pedestrian [6,088]	Van [1,248]	Cyclist [308]	Mean [14,068]
✗	✗	56.3 / 72.4	33.2 / 60.4	57.0 / 68.6	32.2 / 43.5	45.9 / 63.2
✗	✓	56.5 / 72.7	35.2 / 62.9	59.4 / 69.3	32.3 / 44.0	47.1 / 67.6
✓	✗	57.7 / 73.8	44.9 / 71.2	61.8 / 72.0	35.3 / 46.1	52.1 / 72.0
✓	✓	58.1 / 73.9	48.2 / 76.2	63.1 / 74.9	36.7 / 47.4	54.0 / 74.6

A.2 UNIFIED MODEL INPUT

The scale factor α is an important hyper-parameter in our unified model inputs. Hence, we conduct an ablation experiment using different values to determine the optimal setting for this parameter. As presented in Tab. 2, our method is not sensitive to the scale factor within a reasonable range of values, *i.e.*, when this parameter is set in the range from 0.8 to 1.4. Nevertheless, excessively large value will introduce noise, whereas overly small value will ignore valuable information, both leading to significant performance degradation.

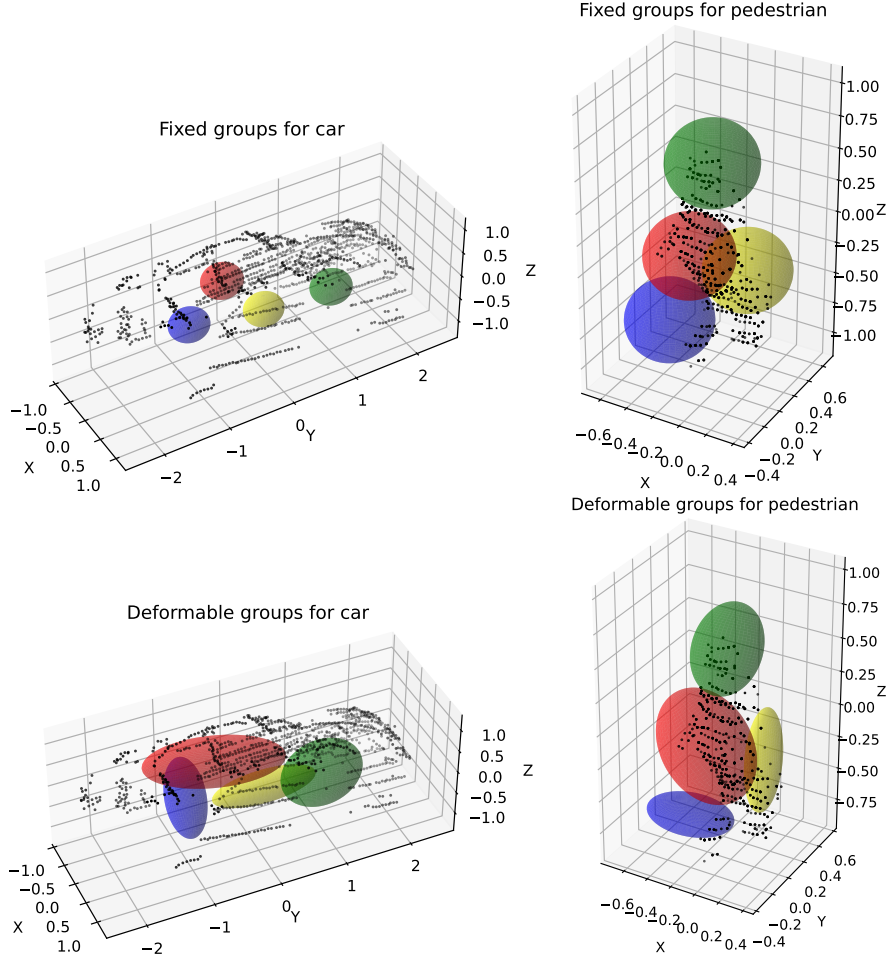


Figure 1: Comparison of groups on the Car and Pedestrian categories. We plot four fixed groups and deformable groups from the first layer in PointNet++ (Qi et al., 2017b) and our AdaFormer network, respectively, using different colors.

Table 2: Performance using different scale factor α . *Success / Precision* are used for evaluation. **Bold** denote the best performance.

Scale Factor α	Car [6,424]	Pedestrian [6,088]	Van [1,248]	Cyclist [308]	Mean [14,068]
0.6	52.4 / 68.1	42.3 / 70.3	48.9 / 57.4	31.2 / 43.0	47.3 / 67.6
0.8	57.5 / 73.2	48.4 / 76.7	63.0 / 74.7	37.2 / 45.0	53.7 / 74.3
1.0	58.1 / 73.9	48.2 / 76.2	63.1 / 74.9	36.7 / 47.4	54.0 / 74.6
1.2	58.3 / 73.7	48.0 / 76.1	62.8 / 74.7	36.1 / 46.6	53.8 / 74.2
1.4	57.8 / 73.1	47.6 / 75.3	62.0 / 73.8	35.0 / 44.8	53.3 / 73.6
1.6	55.8 / 71.5	32.4 / 57.0	56.4 / 66.2	30.2 / 42.5	45.2 / 64.2

30 A.3 UNIFIED LEARNING OBJECTIVE

31 The unified learning objective involves a consistent numerical distribution of predicted targets and
 32 a balanced distribution of positive and negative samples. To investigate their contributions, we
 33 conduct ablation experiments and report the ablation results in Tab. 3. Firstly, as illustrated in
 34 the upper row of Fig. 2, different object categories exhibit significant variations in offset targets,
 35 distracting the model. However, by unifying the offset targets across the three coordinate axes xyz
 36 based on length, width, and height information (as shown in the lower row), the offset targets of

diverse object categories converge within a common numerical space, thereby resulting in improved performance.

In addition, we employ shape-aware labels to define positive and negative samples, which further enhances the tracking performance by 1.2% and 1.4% in average *Success* and *Precision*, as shown in Tab. 3. The scale factor β controls the uniform ratio of positive and negative samples. When the parameter value is set too small, it leads to a scarcity of positive samples, especially at the beginning of training, making it difficult for the model to converge. Conversely, setting this value too large results in an overabundance of positive samples, posing a challenge for the model to distinguish the most accurate ones. Therefore, we further conduct an ablation experiment to determine the optimal value for this parameter. According to Tab. 4, we set the scale factor β to 0.4 in our main experiment.

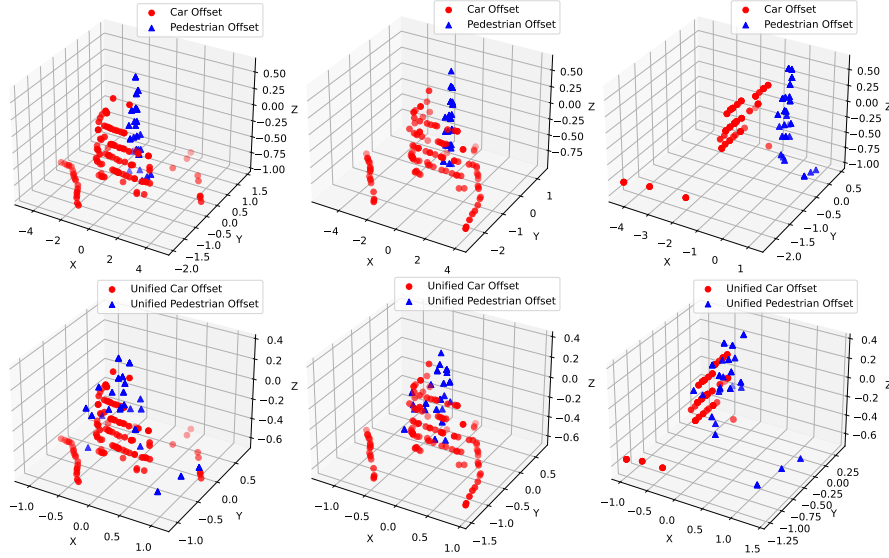


Figure 2: Comparison of offsets between Car and Pedestrian categories. The upper and lower rows represent the cases of without and with unified prediction target design, respectively.

Table 3: Ablation study of unified learning objective. *Success* / *Precision* are used for evaluation. **Bold** denote the best performance.

Unified Prediction Target	Unified Positive -Negative Sample	Car [6,424]	Pedestrian [6,088]	Van [1,248]	Cyclist [308]	Mean [14,068]
✗	✗	57.5 / 72.8	44.5 / 72.1	61.2 / 70.4	35.6 / 44.5	51.8 / 71.7
✗	✓	57.6 / 72.8	45.1 / 72.6	61.4 / 70.8	35.8 / 44.9	52.1 / 72.0
✓	✗	58.1 / 73.7	46.0 / 73.5	62.8 / 74.4	35.9 / 46.6	52.8 / 73.2
✓	✓	58.1 / 73.9	48.2 / 76.2	63.1 / 74.9	36.7 / 47.4	54.0 / 74.6

Table 4: Performance using different scale factor β . *Success* / *Precision* are used for evaluation. **Bold** denote the best performance.

Scale Factor β	Car [6,424]	Pedestrian [6,088]	Van [1,248]	Cyclist [308]	Mean [14,068]
0.2	51.4 / 65.8	26.6 / 46.9	38.4 / 45.0	29.6 / 41.8	39.1 / 55.3
0.3	55.2 / 71.1	34.1 / 61.3	48.2 / 56.7	31.8 / 44.3	45.0 / 65.1
0.4	58.1 / 73.9	48.2 / 76.2	63.1 / 74.9	36.7 / 47.4	54.0 / 74.6
0.5	58.3 / 74.1	47.7 / 75.8	65.7 / 76.0	37.2 / 47.3	54.0 / 74.5
0.6	56.4 / 72.6	44.2 / 73.1	58.8 / 67.7	34.7 / 44.3	50.2 / 71.8
0.7	52.6 / 67.0	39.1 / 66.5	56.3 / 65.8	32.5 / 42.1	46.7 / 66.2

47 B QUALITATIVE RESULTS

48 In order to intuitively demonstrate the effectiveness of our category-unified models, capable of track-
 49 ing objects across all categories using a single network, we conduct a qualitative comparison with
 50 the category-specific counterpart. This comparison is performed on the Car, Pedestrian categories
 51 from the KITTI dataset. As illustrated in Fig. 3, our unified model allow for more accurate and
 52 robust tracking results than the category-specific counterpart across all categories, particularly in
 53 complex scenes marked by numerous distractors and sparse point clouds.

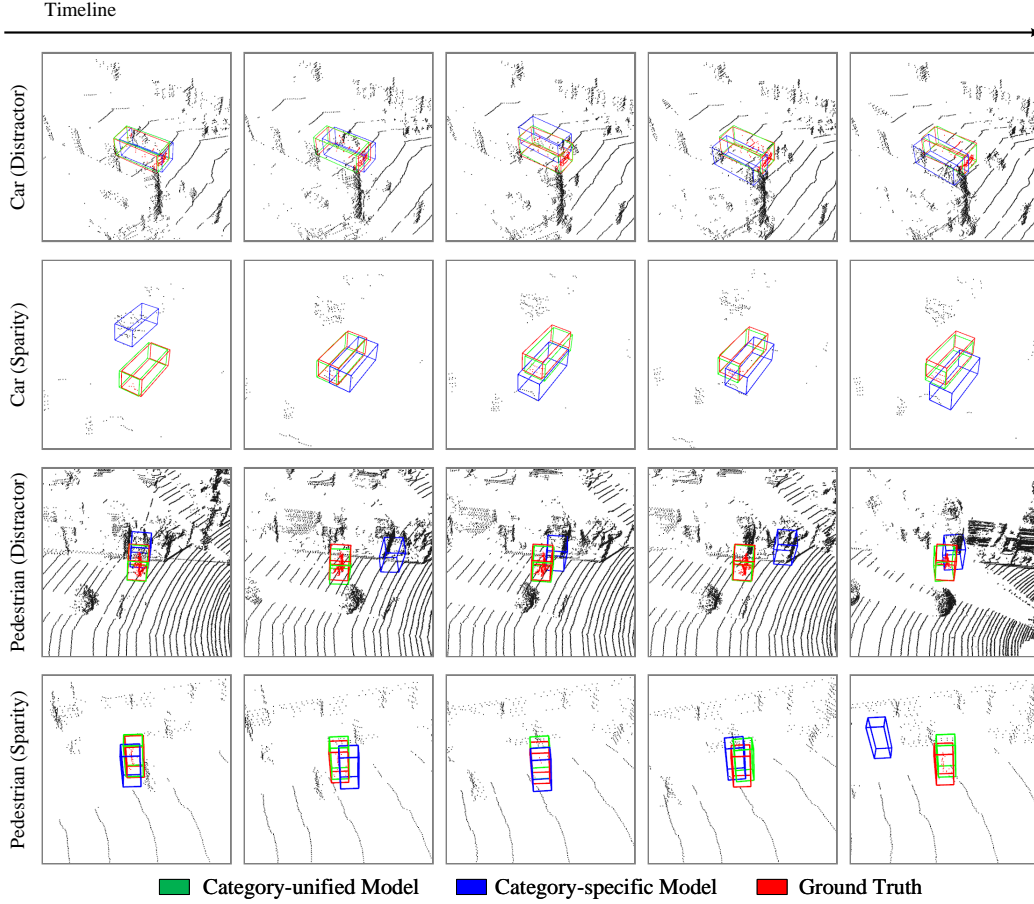






















































Figure 3: Visualization results of Car and Pedestrian categories, including complex scenes marked by numerous distractors and sparse point clouds. The red points are the foreground points of targets. The green and blue boxes denote the prediction results by category-unified model and category-specific counterpart.

54 C COMPARISON WITH CATEGORY-SPECIFIC MODELS.

55 To further demonstrate the potential of our category-unified models, we integrate the proposed uni-
 56 fied components, including unified representation network *AdaFormer*, model inputs and learning
 57 objective into existing tracking methods. We select some classic trackers, such as P2B (Qi et al.,
 58 2020), PTT (Shan et al., 2021), PTTR (Zhou et al., 2022), OSP2B Nie et al. (2023) and (Zheng et al.,
 59 2022) to report the results, as presented in Tab. 5. These unified components not only empower
 60 category-specific trackers to track objects across all categories, but also enhance overall tracking
 61 performance, which proves the effectiveness and promise of our proposed components.

Table 5: Performance comparisons on the KITTI dataset. “Improvement” refers to the performance gain of our category-unified models over the corresponding category-specific counterparts. “” and “” refer to Siamese and motion-centric paradigms, respectively.

Method	Car [6,424]	Pedestrian [6,088]	Van [1,248]	Cyclist [308]	Mean [14,068]
Category-specific P2B (Qi et al., 2020)	56.2 / 72.8	28.7 / 49.6	40.8 / 48.4	32.1 / 44.7	42.4 / 60.0
Category-unified P2B (Ours)	58.1 / 73.9	48.2 / 76.2	63.1 / 74.9	36.7 / 47.4	54.0 / 74.6
Improvement	 1.9 /  1.1	 19.5 /  26.6	 22.3 /  26.5	 4.6 /  3.3	 11.6 /  14.6
Category-specific PTT (Shan et al., 2021)	67.8 / 81.8	44.9 / 72.0	43.6 / 52.5	37.2 / 47.3	55.1 / 74.2
Category-unified PTT (Ours)	67.6 / 82.1	49.2 / 77.4	65.4 / 77.0	37.5 / 46.8	58.8 / 76.4
Improvement	 0.2 /  0.3	 4.3 /  4.6	 21.8 /  24.5	 0.3 /  0.5	 3.7 /  2.2
Category-specific PTTR (Zhou et al., 2022)	65.2 / 77.4	50.9 / 81.6	52.5 / 61.8	65.1 / 90.5	57.9 / 78.2
Category-unified PTTR (Ours)	68.3 / 80.1	53.7 / 84.1	64.2 / 75.6	66.8 / 93.2	61.6 / 81.8
Improvement	 3.1 /  2.7	 2.8 /  2.5	 11.7 /  13.8	 1.7 /  2.7	 3.7 /  3.6
Category-specific OSP2B (Nie et al., 2023)	67.5 / 82.3	53.6 / 85.1	56.3 / 66.2	65.6 / 90.5	60.5 / 82.3
Category-unified OSP2B (Ours)	67.5 / 82.8	55.1 / 86.7	68.7 / 79.3	65.4 / 91.2	62.3 / 84.4
Improvement	 1.0 /  0.5	 1.5 /  1.6	 12.4 /  13.1	 0.2 /  0.7	 1.8 /  2.1
Category-specific M ² Track (Zheng et al., 2022)	65.5 / 80.8	61.5 / 88.2	53.8 / 70.7	73.2 / 93.5	62.9 / 83.4
Category-unified M ² Track (Ours)	67.6 / 80.5	63.3 / 90.0	64.5 / 78.8	76.7 / 94.2	65.8 / 85.0
Improvement	 1.1 /  0.3	 1.8 /  1.8	 9.7 /  8.1	 3.5 /  1.3	 2.9 /  1.6

REFERENCES

- Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.
- Jiahao Nie, Zhiwei He, Yuxiang Yang, Zhengyi Bao, Mingyu Gao, and Jing Zhang. Osp2b: One-stage point-to-box network for 3d siamese tracking. *arXiv preprint*, 2023.
- Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017a.
- Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.
- Haozhe Qi, Chen Feng, Zhiguo Cao, Feng Zhao, and Yang Xiao. 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.
- Jiayao Shan, Sifan Zhou, Zheng Fang, and Yubo Cui. Ptt: Point-track-transformer module for 3d single object tracking in point clouds. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1310–1316. IEEE, 2021.
- Chaoda Zheng, Xu Yan, Haiming Zhang, Baoyuan Wang, Shenghui Cheng, Shuguang Cui, and Zhen Li. 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.
- Changqing Zhou, Zhipeng Luo, Yueru Luo, Tianrui Liu, Liang Pan, Zhongang Cai, Haiyu Zhao, and Shijian Lu. 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.