TACO: Task-Aware Contrastive Learning for Joint LiDAR Localization and 3D Object Detection —Supplementary Material—

In this supplementary, we provide a comprehensive validation of the proposed TACO framework from multiple perspectives. First, we introduce the OxfoLD dataset and analyze annotation reliability to ensure data quality (Sec. A). Then, we present additional quantitative results to further demonstrate the effectiveness of TACO (Sec. B). We also evaluate computational efficiency and model complexity (Sec. C). Next, we provide qualitative analysis to visualize performance improvements (Sec. D). Furthermore, we investigate alternative design choices to validate the necessity of our method (Sec. E). Finally, we clarify evaluation protocols to ensure fair comparisons (Sec. F).

A. Dataset Construction and Annotation Reliability.

A.1. Dataset Details

We evaluate our proposed TACO on the OxfoLD dataset, a large-scale benchmark designed for joint LiDAR localization and 3D object detection. It includes 8 traversals of a fixed route through the city centre of Oxford, covering approximately 10 km across 200 hm^2 of urban area under diverse weather conditions (e.g., clear, overcast). Detailed trajectory statistics for the training and testing splits are summarized in Table 1. The OxfoLD data is derived from the Oxford RobotCar dataset [1], collected using an instrumented Nissan LEAF equipped with dual Velodyne HDL-32E LiDAR. Following prior works [2, 3, 7], we also adopt four sequences for training (11-14-02-26, 14-12-05-52, 14-14-48-55, 18-15-20-12) and four sequences for testing (15-13-06-37, 17-13-26-39, 17-14-03-00, 18-14-14-42).

To ensure consistency with mainstream autonomous driving benchmarks (e.g., KITTI and ONCE), we focus on three core object categories: Car, Pedestrian, and Cyclist. As shown in Fig. 1, the annotated 3D bounding boxes exhibit high geometric consistency and clear category separation across diverse scenarios. These results demonstrate that OxfoLD provides annotation quality comparable to or exceeding widely-used benchmarks, ensuring reliable supervision for both localization and detection tasks.

Sequence	Length	Tag	Training	Test
11-14-02-26	9.37km	sunny	✓	
14-12-05-52	9.22km	overcast	✓	
14-14-48-55	9.04km	overcast	✓	
18-15-20-12	9.04km	overcast	✓	
15-13-06-37	8.85km	overcast		✓
17-13-26-39	9.02km	sunny		✓
17-14-03-00	9.02km	sunny		✓
18-14-14-42	9.04km	overcast		✓

Table 1. Dataset descriptions on the OxfoLD dataset.

A.2. Annotation Consistency and Reliability.

To validate the quality and consistency of annotations in OxfoLD, we conduct an inter-annotator agreement study following established protocols [4]. Specifically, 500 scenes covering diverse categories and environmental conditions are independently annotated by two annotators. We evaluate annotation consistency using 3D **IoU** as well as **NRMSE** of key geometric attributes, including object center, dimensions, and orientation.

As shown in Table 2, the average 3D IoU across the three object categories reaches 0.88, substantially exceeding the standard annotation thresholds used in the mainstream open-source 3D detection benchmarks [6] (e.g., KITTI/Waymo: 0.7; nuScenes: 0.5). This demonstrates strong spatial consistency among the annotators. Moreover, among the three categories of vehicles, pedestrians, and cyclists, the NRMSEs of center coordinates, dimensions, and yaw angles are less than 5%, indicating that the errors of geometric attributes are relatively small. Overall, these results confirm that the annotations in the OxfoLD dataset are highly consistent, reliable, and suitable as a foundation for rigorous experimental evaluation.

B. Additional Quantitative Analysis.

To further validate the effectiveness of TACO, we conduct additional comparisons with task-specific baselines. We

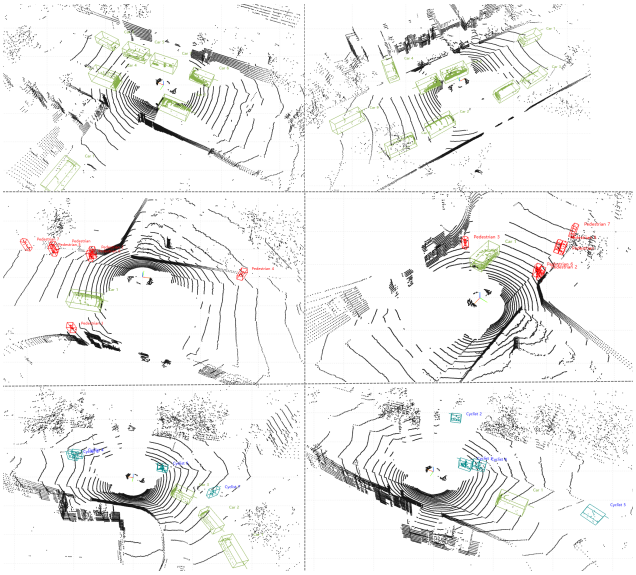

Figure 1. Annotated 3D bounding boxes in OxfoLD: Vehicles (green), Pedestrians (red), Cyclists (blue).

Category	Mean IoU	NRMSE (Center)	NRMSE (Dim.)	NRMSE (Yaw Angle)
Vehicle	0.91	2.3%	3.1%	1.8°
Pedestrian	0.88	3.5%	4.2%	2.5°
Cyclist	0.86	3.8%	4.5%	3.2°
Average	0.88	3.2%	3.9%	2.5°

Table 2. Quantitative results of inter-annotator agreement for the OxfoLD dataset, showing mean 3D IoU and NRMSE of key annotation parameters across vehicle, pedestrian, and cyclist categories. Dim. means dimensions

construct a Naive Cascaded Model (NCM) by combining a localization model [7] and a detection model [8]. While intuitive, this design suffers from error accumulation, limited cross-task interaction, and redundant computation. In contrast, TACO jointly optimizes both tasks within a unified framework, leading to consistent improvements in accuracy. Compared to NCM, TACO consistently improves both localization and detection performance, indicating that the gain is not merely due to increased model capacity. Instead, the improvement stems from effective task-aware feature learning, which enables cross-task interaction and reduces error accumulation. In contrast, the cascaded design of NCM introduces redundant computation and propagates errors between tasks, limiting overall performance.

C. Efficiency and Model Complexity.

As shown in Table 3, we evaluate efficiency in terms of inference time, memory usage, and FLOPs. Compared with NCM, TACO achieves a significantly reduced inference time of only **40 ms**, while reducing memory usage by **1780 MB** and FLOPs by **78.23 G**. The efficiency gain mainly comes from the unified design of TACO, where localization and detection share a common feature backbone. This avoids redundant feature extraction and repeated computation that are inevitable in cascaded pipelines. In contrast, NCM processes the two tasks independently, leading to duplicated computation and increased latency. Notably, the efficiency improvement is achieved *without sacrificing accuracy*. TACO simultaneously improves localization accuracy and detection performance compared to NCM. This confirms that the efficiency gain is structural rather than implementation-dependent.

D. Localization Trajectory Visualization.

To qualitatively evaluate the localization performance of TACO, we visualize trajectory predictions under different methods. As shown in Fig. 2, TACO consistently produces trajectories that are closer to the ground truth compared with existing methods.

In sequence (17-14-03-00), Fig. 2 (3.1)-(3.4) present a detailed comparison, where TACO (3.4) significantly reduces large deviations and suppresses outliers compared to the LiSA baseline Fig. 2 (3.3). Across multiple sequences (Fig. 2 (1.1)-(1.8) and Fig. 2 (2.1)-(2.8)), TACO demonstrates more stable and continuous trajectory predictions. The improvement mainly comes from the integration of object-level priors and task-aware feature learning, which enables the model to effectively suppress interference from dynamic objects while preserving stable structural cues. This is particularly important in complex urban environments where moving objects (e.g., vehicles and pedestrians) introduce significant noise to localization. These improvements are primarily attributed to the integration of object-level priors and task-aware feature learning, which enables the model to suppress dynamic interference while preserving stable structural cues. As a result, TACO achieves more consistent and robust trajectory estimation across diverse urban scenarios.

Methods	Runtime (ms)	Memory (MB)	FLOPs (G)	Loc. Error (m/°)	Det. Acc mGAP(%)
LiSA [7]	38	1577	60.35	0.95/1.14	–
Centerpoint [8]	54	2185	79.89	–	75.20
NCM	66	3748	140.24	0.95/1.14	75.20
TACO (Ours)	40	1968	62.01	0.72/0.85	80.44

Table 3. Efficiency comparison demonstrating the advantage of unified multi-task learning over cascaded pipelines.

E. Analysis of Alternative Design Choices.

As shown in Table 4, we evaluate an alternative baseline that removes dynamic points using GT bounding boxes and applies a single-task localizer (LiSA). However, this strategy

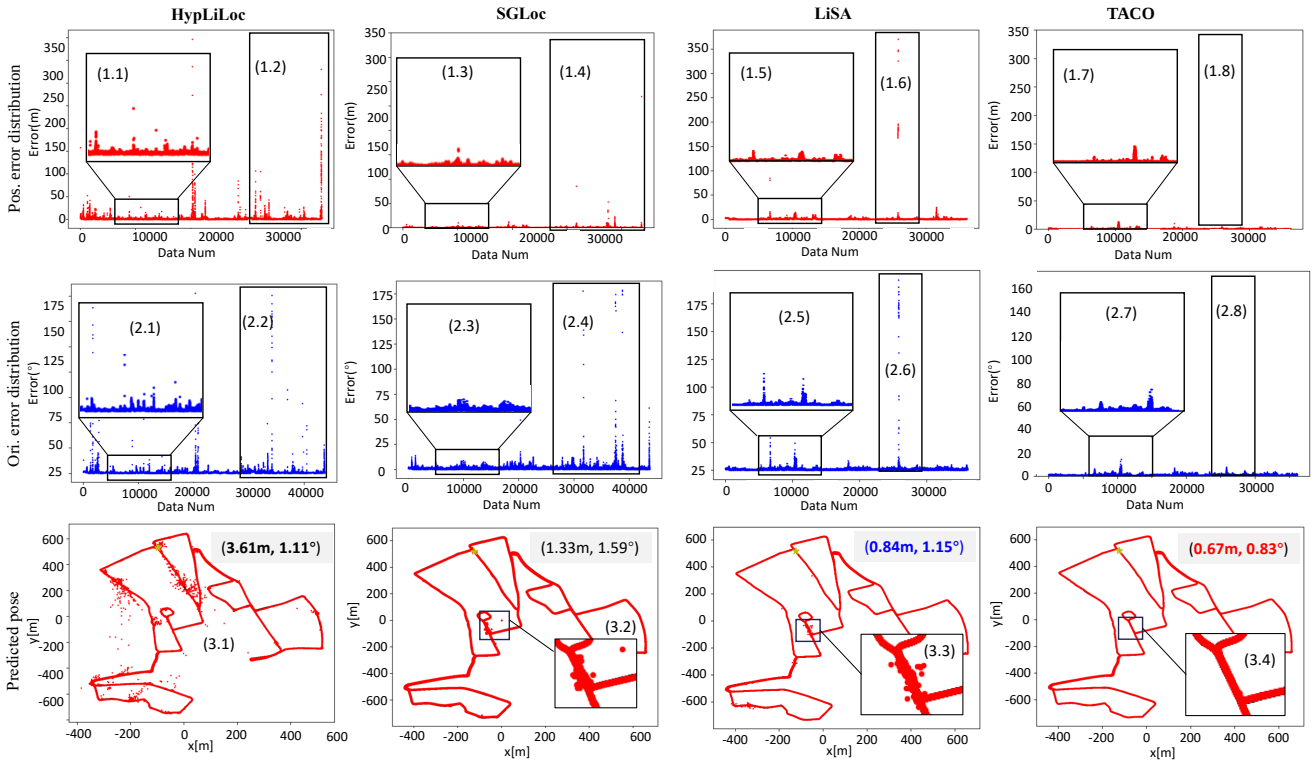

Figure 2. Qualitative results. Visualization of LiDAR localization results for HypLiLoc [5], SGLoc [2], LiSA [7], and TACO including Position Error (m) Distribution, Orientation Error (°) Distribution, and Predicted Pose on OxfoLD dataset.

Methods	Loc. Error (m/°)
GT-boxes + LiSA	1.27m, 1.45°
TACO (Ours)	0.72m, 0.85°

Table 4. Analysis of Alternative Design Choices.

Methods	Loc. Error	NDS	mAP	Veh.	Ped.	Cyc.
TACO (Ours)	0.92m, 0.86°	65.7	78.4	83.4	72.6	45.1

Table 5. Evaluation under official nuScenes metrics (NDS and mAP) for fair comparison.

leads to significantly worse performance, with localization error increasing from **0.72m/0.85°** (TACO) to **1.27m/1.45°**. Although dynamic objects are often considered noise, directly removing them disrupts the geometric completeness of the scene. In LiDAR-based localization, global structure (e.g., road layout, object distribution) plays a critical role, and aggressive filtering inevitably removes useful spatial cues. In contrast, TACO does not discard dynamic information explicitly. Instead, it leverages task-aware contrastive learning to *adaptively suppress* dynamic interference while preserving structurally relevant features. These results demonstrate that learning-based feature modulation is more effective than heuristic dynamic point removal, validating the necessity of our design. Notably, even with oracle-level supervision (GT bounding boxes), the performance still degrades, further highlighting the superiority of learning-based feature modulation over heuristic filtering strategies.

F. Evaluation Protocol and Metrics.

To ensure fair and standardized comparison, we follow widely adopted evaluation protocols. For nuScenes, in addition to IoU-based metrics, we further report official metrics including **NDS** and **mAP**, which comprehensively evaluate detection performance in terms of localization, scale, and orientation. As shown in Table 5, TACO achieves **65.7 NDS** and **78.4 mAP**, demonstrating strong performance under the official evaluation protocol. The consistent improvement across both IoU-based metrics and official nuScenes metrics indicates that the performance gain of TACO is not tied to a specific evaluation criterion. Instead, it reflects genuine improvements in both localization accuracy and detection quality. These results confirm that our evaluation is fully aligned with standard benchmarks and that the superiority of TACO remains consistent under fair and widely accepted evaluation settings.

References

[1] Dan Barnes, Matthew Gadd, Paul Murcutt, Paul Newman, and Ingmar Posner. The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset. In *ICRA*, pages 6433–6438, 2020. 1

[2] Wen Li, Shangshu Yu, Cheng Wang, Guosheng Hu, and Chenglu Wen. Sgloc: Scene geometry encoding for outdoor lidar localization. In *CVPR*, pages 9286–9295, 2023. 1, 3

[3] Wen Li, Chen Liu, Shangshu Yu, Dunqiang Liu, Yin Zhou, Siqi Shen, Chenglu Wen, and Cheng Wang. Lightloc: Learning outdoor lidar localization at light speed. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6680–6689, 2025. 1

[4] Stefanie Nowak and Stefan Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566, 2010. 1

[5] Sijie Wang, Qiyu Kang, Rui She, Wei Wang, Kai Zhao, Yang Song, and Wee Peng Tay. Hypliloc: Towards effective lidar pose regression with hyperbolic fusion. In *CVPR*, pages 5176–5185, 2023. 3

[6] Wenhuan Wu, Innocent Appiah, and Rui Hu. Advancements in 3-d object detection: A comprehensive review. *Journal of King Saud University Computer and Information Sciences*, 37 (9):294, 2025. 1

[7] Bochun Yang, Zijun Li, Wen Li, Zhipeng Cai, Chenglu Wen, Yu Zang, Matthias Muller, and Cheng Wang. Lisa: Lidar localization with semantic awareness. In *CVPR*, pages 15271–15280, 2024. 1, 2, 3

[8] Tianwei Yin, Xingyi Zhou, and Philipp. Center-based 3d object detection and tracking. In *CVPR*, 2021. 2