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

Leyuan Xing[12*] Huanjia Zhang[12*] Dongyu Pan[12] Hai Wu[3]
Qiming Xia[12] Kezheng Xiong[12] Wen Li[4] Chenglu Wen[12†] Cheng Wang[12†]
[1]Fujian Key Laboratory of Urban Intelligent Sensing and Computing, Xiamen University, China
[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, China [3]Pengcheng Laboratory, China
[4]School of Engineering Mathematics and Technology, University of Bristol, UK

Abstract

Reliable navigation and decision-making of autonomous vehicles require both accurate localization and object detection. Traditionally, these two tasks are handled separately, leading to redundant computation and limited cross-task knowledge transfer. This paper proposes TACO, the first Task-Aware COntrastive learning framework, which performs joint LiDAR localization and 3D object detection within a single, unified network. TACO leverages contrastive learning to explicitly decouple and align static geographic features for localization and object-centric features for detection. This bidirectional mutual supervision not only enhances localization robustness in dynamic environments by filtering dynamic noise but also boosts detection accuracy via effective spatial context. Additionally, we propose OxfoLD, the first dataset that provides multi-traversal LiDAR localization ground truth with rich 3D object annotations, thereby supporting task validation across various times and weather conditions. Experimental results demonstrate that TACO achieves state-of-the-art localization accuracy while maintaining competitive detection performance. The code and dataset will be publicly available at: https://github.com/xmuxly/OxfoLD.

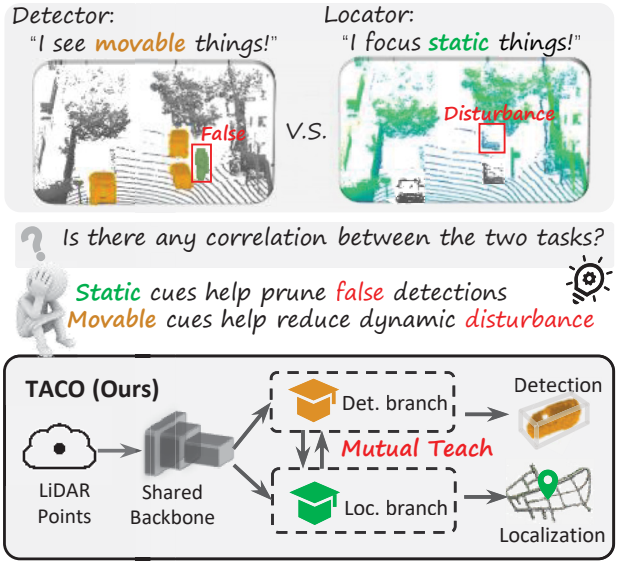

Figure 1. Illustration of the task-aware interaction between detector and locator. (Top) The detector focuses on identifying and classifying movable objects, while the locator concentrates on long-term static structures. We observe that static cues help prune false detections, while movable cues help mitigate location uncertainty. (Bottom) Based on this, we propose TACO, a multi-task learning network for more effective 3D detection and LiDAR localization.

1. Introduction

Self-localization and object detection are crucial tasks for autonomous cars [5, 13, 14] and intelligent robots [17], enabling robust decision-making and safe navigation. Traditional self-localization systems heavily rely on GNSS, which often underperforms in GNSS-denied environments, such as dense urban canyons. LiDAR provides precise 3D spatial information, which enables accurate self-localization in urban scenes via retrieval-based [50], matching-based [32] , and regression-based [51] approaches. Additionally, LiDAR is a cornerstone sensor for detecting traffic participants, which has driven substantial advancements in 3D object detection [44, 47, 56].

However, in traditional design pipelines, the localization

*Equal contribution.
†Corresponding author, clwen@xmu.edu.cn.

and detection models operate within two distinct processing flows, optimizing each task separately through task-specific architectures and metrics. This decoupled paradigm leads to significant computational redundancy and limits the potential for cross-task knowledge transfer. Although methods such as distillation [15] or pruning [10] shrink models, they generally sacrifice performance and fail to fully exploit synergies between tasks at the representation level.

Recently, multi-task learning (MTL) emerges to significantly reduce memory and computation costs by performing tasks via a unified model [6, 16, 54]. However, a key challenge persists for localization and 3D object detection: the two tasks exhibit distinct semantic and geometric priorities, leading to conflicts in feature representation. As shown in Figure 1, the detector identifies and classifies **movable** objects (e.g., vehicles and pedestrians), which requires capturing local, fine-grained dynamic features; In contrast, the locator relies primarily on **static** geographic objects (e.g., roads, buildings), demanding global, stable structural features. This feature-level discrepancy poses a critical issue for conventional MTL frameworks: features useful for one task may introduce noise for another. As a result, naive feature sharing without addressing such task-specific conflicts will lead to poor performance, especially in real-world environments where static and movable elements co-exist densely. Unlike conventional MTL frameworks [22, 36], which use one or more auxiliary tasks to support the main task through feature sharing, we aim to decouple conflicting representations and refine task-aware features to mitigate mutual interference.

To this end, we introduce a novel Task-Aware Contrastive Learning framework (TACO), which is the first unified network for LiDAR localization and 3D object detection tasks. TACO decouples static and movable feature representations through contrastive learning, enabling an intrinsic synergy between the two tasks (Figure 1). As shown in Figure 4, TACO consists of two key stages: **TAFE** (Task-Agnostic Feature Extraction) and **TFCL** (Task-aware Feature Contrastive Learning). (1) In the TAFE stage, a 3D shared backbone extracts point cloud features, minimizing computational overhead through parameter sharing. (2) In the TFCL stage, three complementary modules are designed to refine and decouple task-specific features: the Inter-Task Contrastive Learning module (ITCL), the Intra-Detection Contrastive Learning module (IDCL), and the Intra-Localization Contrastive Learning module (ILCL). ITCL enables bidirectional interaction of task-specific features, allowing the detector's movable cues to guide the locator. Concurrently, the static geographic cues of the locator reinforce the detector. IDCL aggregates object features of the same category across dynamic and static motion states, and separates features of distinct categories via contrastive loss, thereby improving 3D detection performance. ILCL extracts geographic-discriminative features and suppresses pseudo-geographic noise from stationary objects, enhancing LiDAR localization accuracy.

However, existing localization approaches may require multi-traversal data, making existing datasets insufficient for training and evaluating such a joint framework. LiDAR localization benchmarks such as Oxford [1] and NCLT [27] provide multi-traversal LiDAR sequences under varying conditions; however, they lack 3D object annotations. Conversely, detection datasets such as KITTI [12] and Waymo [37] provide rich 3D bounding boxes but do not include repeated traversals necessary for evaluating localization robustness across time and weather conditions. This gap limits its progress in joint detection and localization research. To address this, we introduce the OxfoLD dataset, which is based on the Oxford localization dataset [1] by providing rich detection bounding boxes. Extensive experiments conducted on the OxfoLD dataset evaluate the proposed TACO.

The contributions of this paper are as follows:

- We propose a unified Task-Aware contrastive learning framework (TACO) that jointly enhances the LiDAR localization and 3D object detection tasks through contrastive learning.
- We propose three contrastive modules: inter-task contrastive learning (ITCL) to decouple task-specific features, intra-localization contrastive learning (ILCL) to reinforce localization-discriminative representations, and intra-detection contrastive learning (IDCL) to enhance detection-oriented features.
- We introduce OxfoLD, the first dataset that provides both rich object 3D bounding boxes and multi-traversal localization sequences, supporting the validation of localization and detection tasks across time and weather conditions. Extensive experiments on OxfoLD show that TACO greatly improves over the SOTA localization methods while achieving competitive detection performance.

2. Related work

LiDAR-based localization. LiDAR-based localization aims to estimate the 6-DoF pose from LiDAR point clouds. Traditional LiDAR-based localization methods can be broadly separated into retrieval-based [40, 49, 50] or matching-based [31, 32, 38] approaches. However, these methods require prestored maps, which are time-consuming and labor-intensive [51]. Recently, regression-based methods train CNN models to estimate the 6-DoF pose through Absolute Pose Regression (APR) [3, 20, 33] or Scene Coordinate Regression (SCR) [2, 19, 21, 51]. Both APR and SCR operate under the static scene assumption. However, this assumption is unreliable in dynamic real-world scenarios, where moving objects often corrupt static scene structures and introduce noise. This inspires us to explore the

intrinsic synergy between static scene structures and movable objects to further improve model performance.

LiDAR-based 3D object detection. 3D object detection has garnered significant attention, with mainstream approaches categorized as center-based [26, 29, 46, 53] or anchor-based [8, 34, 35, 45]. The former are predominant in major detection benchmarks and are widely adopted for their efficiency. Our TACO adopts the center-based 3D detection heads, which are attached to its 2D branch. However, the traditional localization and detection tasks are usually executed separately, creating a gap in the task outputs and limiting the potential to improve overall accuracy through bidirectional constraints.

Multi-task Network. Multi-task learning aims to unify multiple tasks into a single network and train them simultaneously in an end-to-end fashion. MultiNet [39] is a seminal work in image-based multi-task learning that unifies street classification, vehicle detection and road segmentation within a single network. LidarMTL [11] proposes a multi-task network for LiDAR-based joint object detection and semantic segmentation. LidarMultiNet [52] unifies object detection, road area classification, and street recognition within a single network. Unlike prior multi-task networks that jointly handle similar tasks, our work focuses on the inherently conflicting tasks of LiDAR localization and 3D object detection.

3. OxfoLD Dataset

Dataset Overview. Oxford RobotCar [1] is a widely-used multi-traversal LiDAR localization dataset. We extend it to OxfoLD with dense, per-frame 3D bounding box annotations for movable objects, enabling the training and evaluation of multitask frameworks such as TACO.Our data splits for OxfoLD adhere to well-established protocols [19, 20, 41, 51, 55] where four traversals are used for training and four routes for testing. We engaged a professional annotation team to label 315k frames with box annotations across eight traversals. As quantified in Table 1, OxfoLD surpasses existing public datasets in annotation quantity (315k vs. WOD's 230k). Figure 2 illustrates the balanced scene distribution and class statistics, confirming its suitability for autonomous driving research. The distribution of the scenes in the training and testing sets is presented in Figure 2(b). The distribution of the objects' numbers for different categories in the training set is presented in Figure 2(a). The distribution of the objects' numbers for different categories in the testing set is presented in Figure 2(c).

Annotation Format. To facilitate 3D object annotation, we employed SUSTechPOINTS, an open-source tool that supports multi-sensor data annotation. We provide 3D annotations for the primary movable objects in urban environments: Vehicles, Pedestrians, and Cyclists. Each object was labeled with a 7-DoF 3D bounding box $(x, y, z, l, w, h, \theta)$, where $[x, y, z]$ represents the center coordinates, $[l, w, h]$ denotes length, width, and height, and θ represents the angle of yaw. Figure 3 illustrates an example of an annotated scene. Note that instances with the same color correspond to the same category.

Labeling Process. To ensure labeling quality, all LiDAR and camera ground truth labels are manually annotated by experienced human annotators. We have then conducted label verification across traversals. The annotation process is organized by dividing each point cloud sequence into segments of 3k scenes, with the most complex scenes requiring an average of 30 hours of annotation per segment. Overall, the dataset consists of 105 segments, totaling over 1,300 hours of annotation time, with an additional 20-40 minutes per segment dedicated to verification and correction. In total, more than 1,600 hours have been invested to ensure the high quality annotations of the dataset. To verify the universality of the OxfoLD dataset, we conduct experiments using three pre-trained models. Table 2 presents the detection results using the model pre-trained on WOD. We also provide the bounding box IoU agreement evaluation in the supplementary material, confirming the high reliability and low inter-annotator variance of the dataset.

4. Method

Problem Statement. For a large-scale outdoor scene, we denote the LiDAR point cloud as $\mathcal{P} = \{p_i\}_{i=1}^{N}$, where $p_i \in \mathbb{R}^3$ denotes a 3D point in the local coordinate frame. The multi-task model TACO extracts deep features $\mathcal{F}$ from $\mathcal{P}$, written as $\mathcal{F} = f(\mathcal{P})$. The objective of the TACO framework is to perform joint 3D object detection and LiDAR-based localization tasks: (1) **Detection**: regressing the 3D bounding boxes $B = \{b_j\}_j$, where $b_j = (x_j, y_j, z_j, l_j, w_j, h_j, \theta_j) \in \mathbb{R}^7$ represents position, width, length, height and azimuth angle, respectively; (2) **Localization**: estimating the 6-DoF ego-pose $\mathbf{p}$ of the vehicle with respect to a global map by predicting world-aligned coordinates and refining the pose through RANSAC for robust estimation.

Overview. As shown in Figure 4, our method consists of two stages: (1) Task-Agnostic Feature Extraction (TAFE) and (2) Task-Aware Feature Contrastive Learning (TFCL). To reduce computational redundancy, TAFE adopts the 3D backbone of LiSA [51] to extract global context features from point clouds, avoiding the redundant voxelization process and minimizing computational overhead through parameter sharing. TFCL further optimizes and refines these features through task-specific contrastive learning.

Inter-task Contrastive Learning Module (ITCL). To mitigate cross-task interference between localization and de-

Dataset	Type	Multi-Traversal	Total Distance	3D Annotations	Label All Traversal	Annotation Frames	Night/Rain
KITTI [12]	City	×	39.2km	✓	×	15k	No/No
nuScenes [4]	City	×	242km	✓	×	40k	Yes/Yes
WOD [37]	City	×	–	✓	×	230k	Yes/Yes
ONCE [24]	City	×	210km	✓	×	16k	Yes/Yes
vReLoc [42]	Indoor	×	272m	×	×	×	No/No
Oxford [1]	City	✓	280km	×	×	×	Yes/Yes
NCLT [27]	Campus	✓	147.4km	×	×	×	Yes/No
Ithaca365 [9]	Outdoor	✓	600km	✓	×	7k	Yes/Yes
OxfoLD (Ours)	City	✓	80km	✓	✓	315k	Yes/Yes

Table 1. Comparison of different datasets. Our proposed OxfoLD not only contains multi-traversal scanning data but also includes 3D annotations for every frame, enabling the large-scale joint training of LiDAR localization and 3D object detection tasks.

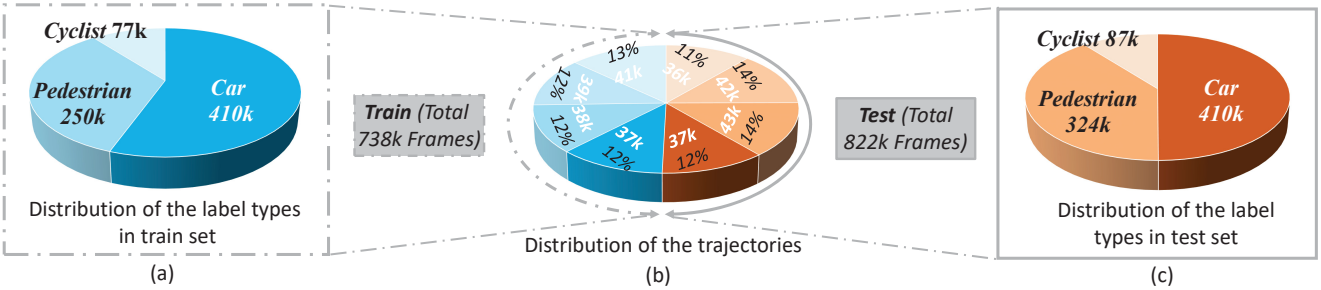

Figure 2. The scene and annotation details. (a) presents the distribution of the objects' numbers for different categories in the training set. (b) presents the distribution of the scenes. (c) presents the distribution of objects' numbers for different categories in the testing set.

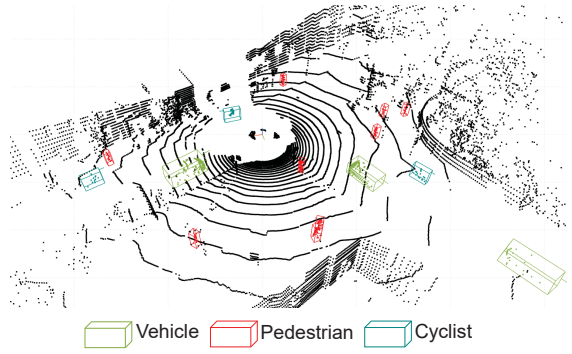

Figure 3. Example of 3D object detection bounding boxes.

Pre-trained	Vehicle		Pedestrian		Cyclist	
	AP	APH	AP	APH	AP	APH
CenterPoint [53]	51.93%	48.32%	20.39%	11.58%	37.32%	32.94%
VoxelRCNN [8]	55.53%	51.53%	22.26%	12.51%	38.44%	33.86%
CasA [43]	57.04%	52.91%	22.15%	12.49%	40.26%	35.21%

Table 2. Detection results of the models pre-trained on WOD [37].

tection, the proposed ITCL module explicitly decouples localization-specific geometric features from detection-specific object features. This decoupling ensures that the localization task focuses on geographic structures while avoiding interference from movable objects, and the detection task avoids learning unwanted correlations with static features (e.g., confusing curbs with vehicles). We define the static geometric features as $\mathcal{G}_g = \{f(i,j)_{\mathrm{g}} | i = 1, ..., w, j = 1, ..., h\}$ and the features of the moving objects as $\mathcal{M}_o = \{f(i,j)_{\mathrm{o}} | i = 1, ..., w, j = 1, ..., h\}$, where $w \times h$ represents the size of the BEV output for the localization and detection heads.

We first rank $\mathcal{G}_g$ according to their L2 norms and select the top-N most prominent features as $\mathcal{F}_g = \{f_g^n\}$. Then, we extract object features $\mathcal{F}_o = \{f_o^m\}$ from $\mathcal{M}_o$ using prior knowledge derived from the ground truth of the objects. Next, we calculate the cosine similarity between each pair of f_g^n and f_o^m. To enforce decoupling, we minimize $\mathcal{L}_{\mathrm{ITCL}}$, which encourages low similarity between static geometric features and movable object features. The ITCL loss function is defined as:

$$\mathcal{L}_{ITCL} = \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_{n=1}^{N} \sum_{m=1}^{M} \mathrm{sim}(f_g^n, f_o^m)}{N \cdot M}, \quad (1)$$

where B is the batch size, $\mathrm{sim}(f_g^n, f_o^m)$ is the cosine similarity between $f_{\mathrm{g}}^n \in F_{\mathrm{g}}$ and $f_{\mathrm{o}}^m \in F_{\mathrm{o}}$. N and M denote the number of static geometric and movable object features in the b-th sample of the batch B, respectively.

Intra-Detection Contrastive Learning Module (IDCL). To address intra-task interference within the detection task, the proposed IDCL module introduces contrastive learning between dynamic (e.g., moving vehicles) and static (e.g.,

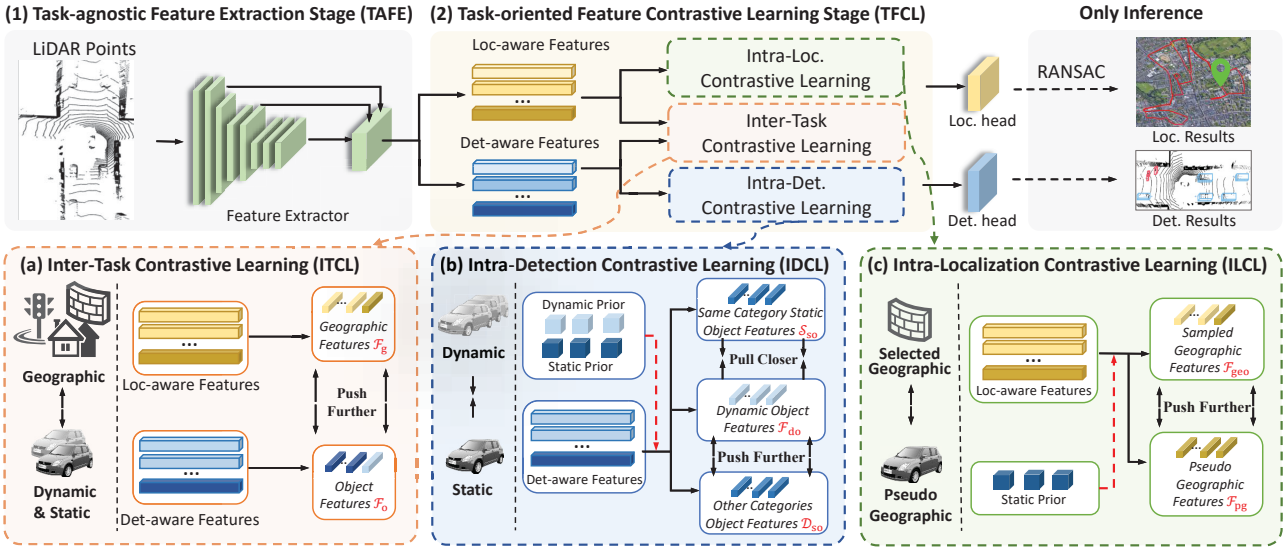

Figure 4. Framework of the proposed TACO method. (1) In the task-agnostic feature extraction stage, we employ a shared 3D backbone to extract context features from LiDAR point clouds, which are simultaneously used for both localization and detection tasks to reduce redundant computation. (2) In the task-aware feature contrastive learning stage (TFCL), we introduce three submodules to explicitly disentangle and align task-specific features.

parked vehicles) object features. Dynamic objects are defined as moving targets with a speed of $v > 0.5m/s$ (calculated based on LiDAR point cloud frame matching), while static objects are defined as targets with a speed of $v \leq 0.5m/s$.

Depending on the motion state of the objects, we further distinguish the object features $\mathcal{F}_o$ into dynamic object features $\mathcal{F}_{do}$ and static features $\mathcal{F}_{so}$. Furthermore, we partition the dynamic object features $\mathcal{F}_{do}$ into same-category dynamic features $\mathcal{S}_{do}$ and different-category dynamic features $\mathcal{D}_{do}$. And we partition the static object features $\mathcal{F}_{so}$ into same-category static features $\mathcal{S}_{so}$ and different-category static features $\mathcal{D}_{so}$. The intra-detection loss effectively pulls the features of same-category dynamic and static objects closer, while separating the features of different-category objects. Then, we define positive and negative sets. For each dynamic object feature $f^i_{\text{do}} \in \mathcal{F}_{do}$, its positive samples are $s^i_{\text{so}} \in \mathcal{S}_{so}$ and its negative samples are $d^i_{\text{so}} \in \mathcal{D}_{so}$. For each static object feature $f^i_{\text{so}} \in \mathcal{F}_{so}$, its positive samples are $s^i_{\text{do}} \in \mathcal{S}_{\text{do}}$ and its negative samples are $d^i_{\text{do}} \in \mathcal{D}_{\text{do}}$.

The intra-detection loss enforces alignment between dynamic and static objects of the same category while separating those of different categories. It consists of two symmetric components:

(1) Dynamic-to-Static Contrastive Loss. For each dynamic object feature f_{do}, the loss pulls its feature toward positive static features of the same class and pushes it away from negative static features:

$$\mathcal{L}_{\text{ds}} = -\log \frac{\sum_{s^i_{so} \in \mathcal{S}_{so}} \exp\left(\text{sim}(f_{\text{do}}, s^i_{so})/\tau\right)}{\sum_{f \in \mathcal{F}_{so}} \exp\left(\text{sim}(f_{\text{do}}, f)/\tau\right)}, \quad (2)$$

where τ is a temperature hyper-parameter that controls the concentration level of the feature distribution.

(2) Static-to-Dynamic Contrastive Loss. For each static object feature f_{so}, the loss pulls its feature toward positive dynamic features of the same class:

$$\mathcal{L}_{\text{sd}} = -\log \frac{\sum_{s^i_{do} \in \mathcal{S}_{do}} \exp\left(\text{sim}(f_{\text{so}}, s^i_{do})/\tau\right)}{\sum_{f \in \mathcal{F}_{do}} \exp\left(\text{sim}(f_{\text{so}}, f)/\tau\right)}, \quad (3)$$

The total loss is the average of the two components over all dynamic and static objects:

$$\mathcal{L}_{\text{IDCL}} = \frac{1}{B}\sum_{b=1}^{B}[\frac{1}{Q}\sum_{q=1}^{Q}\mathcal{L}_{\text{ds}} + \frac{1}{R}\sum_{r=1}^{R}\mathcal{L}_{\text{sd}}], \quad (4)$$

where B is the batch size, Q and R denote the number of dynamic and static objects in the b-th sample, respectively.

Intra-Localization Contrastive Learning Module (ILCL). To further enhance the robustness of the localization features, the proposed ILCL module introduces contrastive learning within the localization task itself. We define the geographic features $\mathcal{F}_g = \{f^n_g\}$, representing the stable long-term geographic features that form the

foundation of robust localization. Then, we define the pseudo-geographic features $\mathcal{F}_{pg} = \{f^n_{pg}\}$, which are erroneously activated by static objects that are visually similar to geographic structures but are inherently movable. ILCL module explicitly disentangles the feature space of the localization head by contrasting the geographic features $\mathcal{F}_g$ and the pseudo-geographic features $\mathcal{F}_{pg}$. By doing so, it ensures that the localization model focuses on stable geographic structures while avoiding interference from temporarily static objects.

Given the localization feature map $\mathcal{F}_{\text{g}} = \{f(i,j)_{\text{g}} | i = 1, 2, ..., w, j = 1, 2, ..., h\}$, where $w \times h$ represents the size of the BEV output for the localization head, we first extract geometric features corresponding to ground-truth bounding boxes of static objects as $\mathcal{F}_{\text{pg}} = \{f^i_{\text{pg}}\}$. Then, we extract pseudo-geometric features corresponding to static environmental elements by ranking $\mathcal{F}_{\text{g}}$ according to their L2 norm and selecting the top-K most prominent features as $\mathcal{F}_{\text{geo}} = \{f^i_{\text{geo}}\}$. The ILCL module minimizes cosine similarity s_{ij} between all pairs of geometric f^i_{geo} and pseudo-geometric object f^j_{pg} features. The total intra-localization contrastive loss is defined as:

$$\mathcal{L}_{\text{ILCL}} = \frac{1}{B}\sum_{b=1}^{B}[\frac{1}{UV}\sum_{i=1}^{U}\sum_{j=1}^{V} s^b_{ij}], \tag{5}$$

where B is the batch size, U and V denote the number of geographic and pseudo-geographic features in the b-th sample of the batch, respectively.

Training Losses. To effectively train the proposed TACO framework, we combine multiple loss functions to optimize both 3D object detection and LiDAR-based localization tasks. The overall training loss $\mathcal{L}$ is defined as a weighted sum of individual losses from each task:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{det}} + \lambda_2\mathcal{L}_{\text{loc}} + \lambda_3\mathcal{L}_{\text{ITCL}} + \lambda_4\mathcal{L}_{\text{ILCL}} + \lambda_5\mathcal{L}_{\text{IDCL}}, \tag{6}$$

where $\mathcal{L}_{\text{det}}, \mathcal{L}_{\text{loc}}, \mathcal{L}_{\text{ITCL}}, \mathcal{L}_{\text{ILCL}}, \mathcal{L}_{\text{IDCL}}$ represent the detection loss, localization loss, inter-task loss, intra-detection loss, and intra-localization loss, respectively. $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ denote the corresponding hyper-parameters that balance the contributions of each loss term.

5. Experiments

5.1. Experimental Settings

Dataset and Metrics. We conducted extensive experiments on the proposed OxfoLD dataset. For localization evaluation, following prior works [21, 51], we report the mean position error (m) and mean orientation error (°). For 3D object detection, we adopt the standard mean Average Precision (mAP) metric with an IoU threshold, following common practice in the literature [48].

To evaluate TACO's generalizability, we further validate it using the nuScenes [4] and KITTI-360 [23] datasets. However, both datasets consist of non-overlapping trajectories, making them unsuitable for evaluating relocalization performance [21, 51]. To overcome this limitation, we construct pseudo-relocalization pairs for nuScenes using 7k a 5-frame sliding window, where the first four frames are used as training references and the last frame serves as the query for testing. For KITTI-360, similarly to [30], we select 7k training frames and choose evaluation frames from the revisited scenarios, requiring each to be within 5 meters of any training frame and at least 30 seconds apart in time.

Implementation Details. All experiments are conducted on a server equipped with an Intel® Xeon® Silver 4314 CPU, 256 GB of RAM, and 4 NVIDIA RTX 3090 GPUs. TACO is implemented with Pytorch [28] and Spconv [7]. For experiments on OxfoLD datasets, we use the Adam optimizer with an initial learning rate of 0.01 and a weight decay of 10^{-4}. The batch size is set to 100, and the model is trained for 100 epochs. The grid range is defined as $[-60\,\text{m}, 60\,\text{m}]$ along the x-axis, $[-60\,\text{m}, 60\,\text{m}]$ along the y-axis, and $[-2\,\text{m}, 6\,\text{m}]$ along the z-axis. The voxel size is set to $[0.2\,\text{m}, 0.2\,\text{m}, 0.2\,\text{m}]$ along the x, y, and z axes, respectively. Further details can be found in the supplementary.

5.2. OxfoLD dataset Results

LiDAR Localization Evaluation. To demonstrate that TACO outperforms single-task models in global localization, we compare it with state-of-the-art methods on the OxfoLD dataset. TACO is jointly trained on LiDAR localization and 3D object detection, allowing it to learn learning more consistent spatial features. As shown in Table 3, our TACO achieves the lowest position error (0.72 m) and orientation error (0.85°), improving over the baseline LiSA [51] by 24.21% and 25.44%, respectively. These results indicate that multi-task feature learning enhances scene representation quality, leading to more accurate global localization.

3D Object Detection Evaluation. We also evaluate the 3D detection performance of TACO on the OxfoLDdataset. The model is jointly trained for 3D object detection and LiDAR localization, which encourages focusing on object foreground points. As shown in Table 4, TACO surpasses the CenterPoint baseline [53] by 28.77% for vehicles, 7.24% for pedestrians, and 8.76% for cyclists in mAP, demonstrating that joint localization constraints further improve detection robustness.

Effect of the Joint Multi-task Training. Table 5 presents an ablation study comparing the results of the jointly-trained model with those of the independently-trained models. We remove the localization head from the jointly-trained model to form the detection-only model, and remove the detection head to form the localization-only model. Compared to single-task models, the jointly-trained model performs better in both detection and localization tasks. In

	Methods	Reference	15-13-06-37	17-13-26-39	17-14-03-00	18-14-14-42	Average
APR	PointLoc	IEEE Sensors 2022	10.75m, 2.36°	11.07m, 2.21°	11.53m, 1.92°	9.82m, 2.07°	10.79m, 2.14°
	PosePN	PR 2022	9.47m, 2.80°	12.98m, 2.35°	8.64m, 2.19°	6.26m, 1.64°	9.34m, 2.25°
	PosePN++	PR 2022	4.54m, 1.83°	6.44m, 1.78°	4.89m, 1.55°	4.64m, 1.61°	5.13m, 1.69°
	PoseMinkLoc	PR 2022	6.77m, 1.84°	8.84m, 1.84°	8.08m, 1.69°	6.56m, 2.06°	7.56m, 1.86°
	PoseSOE	PR 2022	4.17m, 1.76°	6.16m, 1.81°	5.42m, 1.87°	4.16m, 1.70°	4.98m, 1.79°
	STCLoc	TITS2022	5.14m, 1.27°	6.12m, 1.21°	5.32m, 1.08°	4.76m, 1.19°	5.34m, 1.19°
	NIDALoc	TITS2023	3.71m, 1.50°	5.40m, 1.40°	3.94m, 1.30°	4.08m, 1.30°	4.28m, 1.38°
	HypLiLoc	CVPR 2023	5.03m, 1.46°	4.31m, 1.43°	3.61m, 1.11°	2.61m, 1.09°	3.89m, 1.27°
	DiffLoc	CVPR 2024	2.03m, 1.04°	1.78m, 0.79°	2.05m, 0.83°	1.56m, 0.83°	1.86m, 0.87°
SCR	SGLoc	CVPR 2023	1.79m, 1.67°	1.81m, 1.76°	1.33m, 1.59°	1.19m, 1.39°	1.53m, 1.60°
	LiSA	CVPR 2024	0.94m, 1.10°	1.17m, 1.21°	0.84m, 1.15°	0.85m, 1.11°	0.95m, 1.14°
	LightLoc	CVPR 2025	0.82m, 1.12°	0.85m, 1.07°	0.81m, 1.11°	0.82m, 1.16°	0.83m, 1.12°
	TACO (Ours)	–	**0.80m, 0.89°**	**0.75m, 0.83°**	**0.64m, 0.83°**	**0.71m, 0.88°**	**0.72m, 0.85°**

Table 3. Single-model global localization performance comparisons on OxfoLD *test* set. Mean position error (m) and mean orientation error (°) for representative methods are reported. Best performance is highlighted in **bold**, lower is better. TACO outperforms all baseline methods in terms of both position and orientation accuracy.

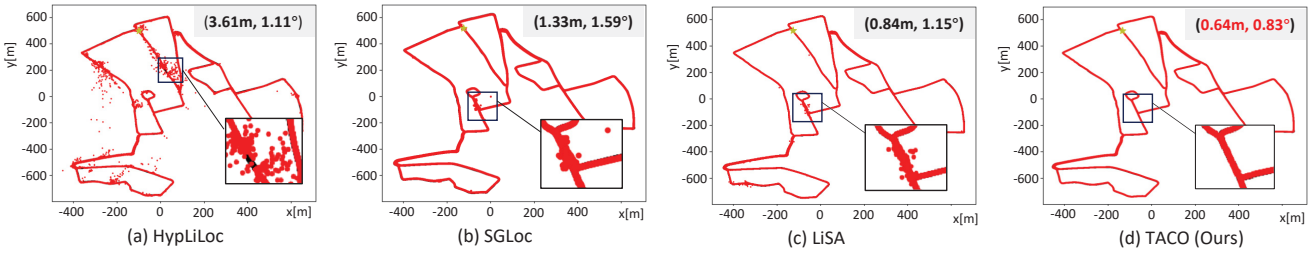

Figure 5. Visualization comparison of different localization results. The ground truth and prediction are black and red lines, respectively. Each subplot displays the predicted discrete points. The star denotes the first frame. Each sub-figure shows noticeable outliers, which renders its results unreliable in these regions. The caption of each subfigure shows the mean position error (m) and orientation error (°).

Methods	Reference	Vehicle AP@0.5	Pedestrian AP@0.3	Cyclist AP@0.3
Pointnet	CVPR 2017	55.63%	44.56%	49.53%
SECOND	Sensor 2018	63.12%	50.66%	54.39%
Centerpoint	CVPR 2021	63.37%	48.79%	52.83%
PillarneXt	CVPR 2023	72.54%	51.83%	55.99%
FSHNet	CVPR 2025	78.31%	50.22%	58.21%
TACO (Ours)	–	**81.60%**	**52.32%**	**57.46%**

Table 4. Comparison with previous detection methods on OxfoLD *test* set.

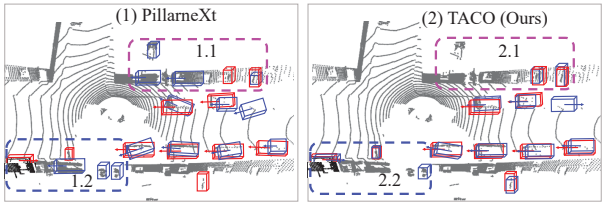

Figure 6. Visualization comparison of different detection results. Blue/red boxes are ground truth/predicted results.

addition, by sharing part of the network among the tasks, the jointly-trained model is more efficient than directly combining the independent single-task models.

Multi-Task Architectures Comparison. To validate the effectiveness of the TACO framework, we conducted comparative experiments between TACO and two previous multi-task learning architectures, Cross-stitch [25] and LiDARFormer [57], for joint detection and localization tasks.

Models	Localization Error	Vehicle AP @0.5	Pedestrian AP @0.3	Cyclist AP @0.3
Detection-only	-	60.01%	43.64%	47.80%
Localization-only	0.95m,1.14°	-	-	-
Joint Training	0.72m,0.85°	81.60%	52.32%	57.46%

Table 5. Comparison of the TACO results of the jointly trained model and independently trained single-task models.

Methods	Localization Error	Vehicle AP@0.5	Pedestrian AP@0.3	Cyclist AP@0.3
Cross-stitch [25]	1.97m, 2.01°	55.23%	40.88%	43.16%
LiDARFormer [57]	0.89m, 1.06°	72.54%	47.21%	51.00%
TACO(Ours)	**0.72m, 0.85°**	**81.60%**	**52.32%**	**57.46%**

Table 6. Comparison of different MLT methods and TACO variants on localization and detection tasks using the OxfoLD dataset.

ID	DET	ITCL	IDCL	ILCL	Localization Error	Vehicle AP@0.5
1					0.95m, 1.14°	–
2	✓				0.89m, 1.06°	72.54%
3	✓	✓			0.82m, 0.99°	75.32%
4	✓		✓		0.85m, 0.93°	77.45%
5	✓			✓	0.76m, 0.89°	76.21%
6	✓	✓	✓	✓	**0.72m, 0.85°**	**81.60%**

Table 7. Ablation studies for the localization and vehicle 3D detection performance of TACO on the OxfoLD *test* set.

Trajectories	LiSA [51]	TACO (Ours)
1	6.42m, 4.21°	4.32m, 3.77°
2	3.97m, 2.87°	2.68m, 2.19°
3	1.46m, 1.66°	1.08m, 1.37°
4	**0.95m, 1.14°**	**0.72m, 0.85°**

Table 8. The impact of training with different number of scanning trajectories on localization performance for LiSA and TACO, tested on sequence 15-13-06-37.

As shown in Table 6, our TACO achieves significant improvements in localization metrics compared to LiDARFormer (0.72m/0.85° vs. 0.89m/1.06°), while also obtaining optimal performance across all three categories in detection metrics. We attribute this improvement to TACO's task-aware contrastive learning mechanism, which enables cross-task feature disentanglement while maintaining parameter efficiency, outperforming LiDARFormer that solely relies on shared backbone and task-specific heads.

Ablation Studies. We conduct ablation experiments on TACO's components to validate their impact on LiDAR localization and 3D object detection. Results on OxfoLD *test* set are presented in Table 7. Specifically, (1) Baseline: LiSA [51], a pure LiDAR localization method via SCR, serves as the baseline with position/orientation errors of 0.95 m and 1.14°. (2) With DET: Integrating the Center-based detector head [53] (DET) reduces the position/orientation error to 0.89 m/1.06° and achieves 72.54% vehicle AP, demonstrating that detection provides implicit spatial cues that benefit localization. (3) DET + ITCL: Adding ITCL further reduces the error to 0.82m/0.99° and improves the vehicle AP to 75.32%, validating its effectiveness in decoupling task-specific features. (4) DET + IDCL: IDCL enhances detection to 77.45% AP by refining motion state-aware features. (5) DET + ILCL: ILCL boosts geographic feature disentanglement, achieving 0.76m/0.89° localization error and 76.21% detection AP. (6) Fully integrated: TACO reduces position/orientation error by 24.21%/25.43%, respectively, over the baseline, achieving 81.60% vehicle average precision.

Impact of Training Trajectory Quantity. To evaluate the effect of training trajectory quantity, we compare LiSA and TACO on the OxfoLD dataset. As shown in Table 8, both methods improve as more trajectories are used. Importantly, TACO consistently outperforms LiSA: with a single trajectory, it reduces position and orientation errors from 5.12m/4.21° (LiSA) to 4.23m/3.77°, and with four trajectories, it further improves to 0.72m/0.85°, compared to LiSA's 0.95m/1.14°. These results demonstrate that TACO provides stable and superior localization performance across different training conditions.

Visualization Analysis. We visualize the predicted trajectories and detection results for sequence 17-14-03-00 in Figure 5. TACO produces smoother, more accurate trajectories with fewer outliers compared to [19, 41, 51], indicating that TACO effectively handles dynamic elements while preserving trajectory accuracy. Figure 6 further compares the detection results, where PillarNeXt [18] frequently misclassifies static structures. Thanks to the task-aware mutual learning design, TACO accurately detects pedestrians and suppresses false positives from background clutter.

Methods	Localization Error	Vehicle AP@0.5	Pedestrian AP@0.3	Cyclist AP@0.3
Cross-stitch [25]	2.11m, 2.67°	60.10%	42.54%	47.03%
LiDARFormer [57]	1.01m, 1.14°	83.81%	86.20%	50.32%
TACO (Ours)	0.92m, 0.86°	85.72%	86.93%	52.30%

Table 9. Comparison of different MLT methods and TACO variants on two tasks using the nuScenes[4] dataset.

Methods	Seq. 00 (7k/647)		Seq. 09 (7k/2534)	
	Loc. Error	Veh. AP@0.5	Loc. Error	Veh. AP@0.5
LiDARFormer [57]	2.03m, 1.89°	50.21%	1.84m, 1.77°	53.21%
TACO (Ours)	1.94m, 1.68°	51.58%	1.79m, 1.67°	53.47%

Table 10. Comparison of different MLT methods and TACO variants on two tasks using the KITTI-360 [23] dataset.

5.3. nuScenes and KITTI-360 Datasets Results

As described in Sec. 5.1, we evaluate TACO's generalization on nuScenes dataset [4] and KITTI-360 dataset [23]. Table 9 reports the nuScenes results: TACO attains 0.92m/0.82° in localization and 85.72%/86.93%/52.3% in detection, outperforming LiDARFormer by 8.91%/24.56%, 2.27%/0.84%/3.93%. Table 10 presents the KITTI-360 results: Following [23], only Vehicle AP is reported. Compared with LiDARFormer, our improvements remain significant. Specifically, we achieve 1.79m/1.67°/53.47% in position accuracy, orientation and vehicle detection accuracy on seq.09. These experiments demonstrate our method's strong generalization across different datasets.

6. Conclusion

In this work, we propose a novel Task-Aware Contrastive Learning framework named TACO. To validate our method, we propose the OxfoLD, the first dataset that supports both LiDAR-based localization and 3D object detection tasks simultaneously. We establish a benchmark on the OxfoLD dataset by integrating the 3D object detection pipeline into classic LiDAR-based localization methods. Moreover, we propose sub-modules named ITCL, ILCL, and IDCL to enhance the performance of both tasks. In summary, OxfoLD and our exploration pave the way for the advancement of LiDAR-based localization and 3D object detection.

Limitations. A limitation of our research is the reliance on multi-traversal LiDAR scan data with rich 3D bounding boxes. Future work could explore unified perception using less labeled information or data from a single traversal scan.

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China (No.42571514).

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. 2, 3, 4

[2] Eric Brachmann and Carsten Rother. Visual camera relocalization from rgb and rgb-d images using dsac. *PAMI*, 44(9):5847–5865, 2021. 2

[3] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *CVPR*, pages 2616–2625, 2018. 2

[4] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 4, 6, 8

[5] Jiaqi Chen, Bingqian Lin, Xinmin Liu, Lin Ma, Xiaodan Liang, and Kwan-Yee K Wong. Affordances-oriented planning using foundation models for continuous vision-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23568–23576, 2025. 1

[6] Xuesong Chen, Shaoshuai Shi, Tao Ma, Jingqiu Zhou, Simon See, Ka Chun Cheung, and Hongsheng Li. M3net: Multimodal multi-task learning for 3d detection, segmentation, and occupancy prediction in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2275–2283, 2025. 2

[7] Spconv Contributors. Spconv: Spatially sparse convolution library. https://github.com/traveller59/spconv, 2022. 6

[8] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wen gang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *AAAI*, 2021. 3, 4

[9] Carlos A. Diaz-Ruiz, Youya Xia, Yurong You, Jose Nino, Junan Chen, Josephine Monica, Xiangyu Chen, Katie Luo, Yan Wang, Marc Emond, Wei-Lun Chao, Bharath Hariharan, Kilian Q. Weinberger, and Mark Campbell. Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21383–21392, 2022. 4

[10] Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. Depgraph: Towards any structural pruning. In *CVPR*, pages 16091–16101, 2023. 2

[11] Di Feng, Yiyang Zhou, Chenfeng Xu, Masayoshi Tomizuka, and Wei Zhan. A simple and efficient multi-task network for 3d object detection and road understanding. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7067–7074, 2021. 3

[12] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 2, 4

[13] Wencheng Han, Dongqian Guo, Cheng-Zhong Xu, and Jianbing Shen. Dme-driver: Integrating human decision logic and 3d scene perception in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3347–3355, 2025. 1

[14] Xiangkun He, Jingda Wu, Zhiyu Huang, Zhongxu Hu, Jun Wang, Alberto Sangiovanni-Vincentelli, and Chen Lv. Fear-neuro-inspired reinforcement learning for safe autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 46(1):267–279, 2023. 1

[15] Yuenan Hou, Xinge Zhu, Yuexin Ma, ChenChange Loy, and Yikang Li. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *CVPR*, 2022. 2

[16] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *CVPR*, 2023. 2

[17] Peter Karkus, Shaojun Cai, and David Hsu. Differentiable slam-net: Learning particle slam for visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2815–2825, 2021. 1

[18] Jinyu Li, Chenxu Luo, Xiaodong Yang, Chenxu Luo, and Xiaodong Yang. Pillarnext: Rethinking network designs for 3d object detection in lidar point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17567–17576, 2023. 8

[19] 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. 2, 3, 8

[20] Wen Li, Yuyang Yang, Shangshu Yu, Guosheng Hu, Chenglu Wen, Ming Cheng, and Cheng Wang. Diffloc: Diffusion model for outdoor lidar localization. In *CVPR*, pages 15045–15054, 2024. 2, 3

[21] 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. 2, 6

[22] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7345–7353, 2019. 2

[23] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022. 6, 8

[24] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, Hang Xu, and Chunjing Xu. One million scenes for autonomous driving: Once dataset. *Cornell University - arXiv,Cornell University - arXiv*, 2021. 4

[25] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3994–4003, 2016. 7, 8

[26] Ramin Nabati and Hairong Qi. Centerfusion: Center-based radar and camera fusion for 3d object detection. In *WACV*, pages 1526–1535, 2021. 3

[27] Carlevaris-Bianco Nicholas, K. Ushani Arash, and M. Eustice Ryan. University of michigan north campus long-term vision and lidar dataset. *IJRR*, 35:545–565, 2015. 2, 4

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[29] Xidong Peng, Xinge Zhu, Tai Wang, and Yuexin Ma. Side: Center-based stereo 3d detector with structure-aware instance depth estimation. In *WACV*, 2022. 3

[30] Georgi Pramatarov, Matthew Gadd, Paul Newman, and Daniele De Martini. That's my point: Compact object-centric lidar pose estimation for large-scale outdoor localisation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12276–12282. IEEE, 2024. 6

[31] Torsten Sattler, Michal Havlena, Konrad Schindler, and Marc Pollefeys. Large-scale location recognition and the geometric burstiness problem. In *CVPR*, pages 1582–1590, 2016. 2

[32] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. Are large-scale 3d models really necessary for accurate visual localization? In *CVPR*, pages 1637–1646, 2017. 1, 2

[33] Yoli Shavit, Ron Ferens, and Yosi Keller. Learning multi-scene absolute pose regression with transformers. In *ICCV*, pages 2733–2742, 2021. 2

[34] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*, pages 770–779, 2019. 3

[35] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *CVPR*, pages 10526 – 10535, 2020. 3

[36] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2647–2664, 2020. 2

[37] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 2, 4

[38] Linus Svärm, Olof Enqvist, Fredrik Kahl, and Magnus Oskarsson. City-scale localization for cameras with known vertical direction. *PAMI*, 39(7):1455–1461, 2016. 2

[39] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *2018 IEEE intelligent vehicles symposium (IV)*, pages 1013–1020. IEEE, 2018. 3

[40] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *CVPR*, pages 1808–1817, 2015. 2

[41] 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, 8

[42] Wei Wang, Bing Wang, Peijun Zhao, Changhao Chen, Ronald Clark, Bo Yang, Andrew Markham, and Niki Trigoni. Pointloc: Deep pose regressor for lidar point cloud localization. *IEEE Sensors Journal*, 22(1):959–968, 2021. 4

[43] Hai Wu, Jinhao Deng, Chenglu Wen, Xin Li, and Cheng Wang. Casa: A cascade attention network for 3d object detection from lidar point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 2022. 4

[44] Hai Wu, Chenglu Wen, Shaoshuai Shi, Xin Li, and Cheng Wang. Virtual sparse convolution for multimodal 3d object detection. In *CVPR*, pages 21653–21662, 2023. 1

[45] Qiming Xia, Yidong Chen, Guorong Cai, Guikun Chen, Daoshun Xie, Jinhe Su, and Zongyue Wang. 3-d hanet: A flexible 3-d heatmap auxiliary network for object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13, 2023. 3

[46] Qiming Xia, Jinhao Deng, Chenglu Wen, Hai Wu, Shaoshuai Shi, Xin Li, and Cheng Wang. Coin: Contrastive instance feature mining for outdoor 3d object detection with very limited annotations. In *ICCV*, pages 6254–6263, 2023. 3

[47] Qiming Xia, Wei Ye, Hai Wu, Shijia Zhao, Leyuan Xing, Xun Huang, Jinhao Deng, Xin Li, Chenglu Wen, and Cheng Wang. Hinted: Hard instance enhanced detector with mixed-density feature fusion for sparsely-supervised 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15321–15330, 2024. 1

[48] Qiming Xia, Wenkai Lin, Haoen Xiang, Xun Huang, Siheng Chen, Zhen Dong, Cheng Wang, and Chenglu Wen. Learning to detect objects from multi-agent lidar scans without manual labels. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1418–1428, 2025. 6

[49] Yan Xia, Mariia Gladkova, Rui Wang, Qianyun Li, Uwe Stilla, Joao F Henriques, and Daniel Cremers. Casspr: Cross attention single scan place recognition. In *ICCV*, pages 8461–8472, 2023. 2

[50] Yan Xia, Mariia Gladkova, Rui Wang, Qianyun Li, Uwe Stilla, Joao F Henriques, and Daniel Cremers. Casspr: Cross attention single scan place recognition. In *ICCV*, pages 8461–8472, 2023. 1, 2

[51] 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, 6, 8

[52] Dongqiangzi Ye, Zixiang Zhou, Weijia Chen, Yufei Xie, Yu Wang, Panqu Wang, and Hassan Foroosh. Lidarmultinet: Towards a unified multi-task network for lidar perception. In *AAAI*, 2022. 3

[53] Tianwei Yin, Xingyi Zhou, and Philipp. Center-based 3d object detection and tracking. In *CVPR*, 2021. 3, 4, 6, 8

[54] Hanxun Yu, Wentong Li, Song Wang, Junbo Chen, and Jianke Zhu. Inst3d-lmm: Instance-aware 3d scene understanding with multi-modal instruction tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14147–14157, 2025. 2

[55] Shangshu Yu, Cheng Wang, Yitai Lin, Chenglu Wen, Ming Cheng, and Guosheng Hu. Stcloc: Deep lidar localization with spatio-temporal constraints. *IEEE TITS*, 24(1):489–500, 2022. 3

[56] Shijia Zhao, Qiming Xia, Xusheng Guo, Pufan Zou, Maoji Zheng, Hai Wu, Chenglu Wen, and Cheng Wang. Sp3d: Boosting sparsely-supervised 3d object detection via accurate cross-modal semantic prompts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29374–29384, 2025. 1

[57] Zixiang Zhou, Dongqiangzi Ye, Weijia Chen, Yufei Xie, Yu Wang, Panqu Wang, and Hassan Foroosh. Lidarformer: A unified transformer-based multi-task network for lidar perception. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14740–14747. IEEE, 2024. 7, 8