Robust 3D Object Detection via Physical-Aware Augmentation and Class-Specific Model Ensembling

Shuangzhi Li, Junlong Shen, and Xingyu Li
Department of Electrical and Computer Engineering, University of Alberta
Edmonton, Alberta, Canada
{shuangzh, junlong6, xingyu}@ualberta.ca

Abstract

3D object detection is a fundamental component of autonomous driving and robotic perception. The RoboSense 2025 Track 5 Challenge focuses on cross-platform LiDAR-based detection, where models trained on vehicle-mounted LiDAR must generalize to drone- and quadruped-mounted platforms. This setting is particularly challenging due to domain gaps introduced by heterogeneous sensors, including differences in viewpoint, height, scene composition, and motion behavior. To tackle these challenges, we propose a framework that combines physics-aware augmentation and class-specific model ensembling. The physics-aware augmentation module introduces scene-aware data resampling, viewpoint-variation augmentation, and height-variation test-time augmentation to explicitly model platform-specific geometric and semantic shifts. Meanwhile, the class-specific ensemble strategy applies tailored optimization for each object category, enhancing cross-domain robustness and detection stability. Extensive experiments on the official Track 5 benchmark demonstrate that our method consistently outperforms strong baselines in both Phase 1 and Phase 2, validating its effectiveness for robust and generalizable cross-platform 3D object detection.

1. Introduction

3D object detection aims to localize and recognize objects in three-dimensional space, serving as a cornerstone for autonomous driving and robotic perception [2, 8, 22, 34, 37, 45]. The RoboSense 2025 Challenge Track 5 focuses on *cross-platform LiDAR-based 3D object detection*, where models trained on vehicle-mounted LiDAR data must generalize effectively to drone- and quadruped-mounted sensing platforms [23, 30, 60]. This setting mirrors a real-world deployment problem: LiDAR sensors are increasingly distributed across heterogeneous robotic systems, and maintaining robust perception across these platforms is essential for safety-critical applications such as intelligent transportation [1, 46, 49, 55], aerial mapping [36], and mobile service robots [9].

However, achieving consistent detection performance across platforms remains highly challenging due to domain shifts introduced by heterogeneous sensor characteristics [10, 17, 18, 25, 30, 52, 60]. First, changes in sensor viewpoint (*e.g.*, aerial drones versus ground quadrupeds) induce large geometric transformations in the captured point clouds, primarily caused by rotation and perspective distortions. Second, variation in scanning height leads to systematic biases in point density and spatial coverage [26]. For instance, drone-mounted LiDARs capture denser observations from object rooftops, whereas quadruped-mounted sensors emphasize lower body structures or ground-contact regions.

Beyond geometric discrepancies, platform-dependent operating environments introduce semantic and behavioral domain gaps. Vehicle-mounted LiDARs typically observe structured road scenes populated by cars and pedestrians at similar ground levels, while quadruped-mounted systems encounter cluttered pedestrian zones with varying elevation and object distributions [7, 30, 35]. Therefore, cross-platform 3D object detection requires models that jointly achieve geometric adaptability and semantic robustness across diverse sensing conditions.

To address these challenges, we propose an effective framework that integrates *physics-aware augmentation* with *class-specific model ensembling*. The proposed approach targets both geometric and semantic disparities through complementary strategies. For object categories such as cars and pedestrians, we design distinct training schemes and scene-resampling strategies that better capture their characteristic geometry and motion patterns. During source-domain training, a viewpoint-variation augmentation module perturbs LiDAR data to simulate multi-angle observations, enhancing robustness to perspective changes. In the unsupervised target adaptation stage, a height-variation test-time augmentation strategy dynamically compensates for biases introduced by different scanning heights, further improving generalization across heterogeneous platforms.

Overall, our framework emphasizes a data-driven yet lightweight solution that requires no architectural redesign, focusing instead on modeling real-world sensor variability through physically grounded augmentations and category-specific learning. This design philosophy enables the detector to achieve strong cross-platform robustness while maintaining high accuracy on standard benchmarks.

2. Related Works

2.1. Point-Cloud-Based 3D Object Detection

Point-cloud-based 3D object detection aims to identify and localize objects directly in three-dimensional point clouds [4, 24, 29, 31, 32, 34, 37]. Existing approaches can be broadly categorized into voxel-based and point-based methods. Voxel-based methods [6, 50, 57] discretize point clouds into regular grids and apply sparse convolutional networks for efficient feature extraction, achieving high computational efficiency but often sacrificing fine geometric details. In contrast, point-based methods [39, 42, 55] operate directly on raw points to preserve precise structural information, albeit at a higher computational cost. To achieve a balance between accuracy and efficiency, hybrid point–voxel approaches [32, 40, 41, 44] combine the strengths of both paradigms by integrating voxel-level efficiency with point-level geometric fidelity. In this work, we primarily focus on point–voxel-based methods as the backbone for robust cross-platform detection.

2.2. Unsupervised Domain Adaptation (UDA) for 3D Object Detection

Unsupervised domain adaptation (UDA) for 3D object detection aims to transfer a model trained on a labeled source domain to an unlabeled target domain while mitigating performance degradation caused by domain shifts [11, 12, 14, 15, 25, 28, 38, 60]. Early UDA approaches, such as SN [51], leverage box-size priors to align global feature distributions between domains. Self-training frameworks, including ST3D [58] and ST3D++ [59], adopt pseudo-label refinement and random object-size augmentation to iteratively adapt detectors to the target domain. More recent methods, such as the density-based UDA extension of [13, 26, 33, 56], explicitly address point-density discrepancies by introducing physics-aware density resampling, further narrowing the geometric domain gap. Despite these advancements, existing UDA methods are not explicitly designed for cross-platform adaptation [30]. Consequently, significant challenges remain when transferring models between vehicle-, drone-, and quadruped-mounted LiDAR sensors, where viewpoint, height, and environmental variations jointly affect feature alignment.

3. Methodology

3.1. Preliminary

Regarding the task of Track 5, a point cloud frame is denoted as $\mathbf{X} = \{p_i\}_{i=1}^{N}$, where each point $p = (x^p, y^p, z^p)$ represents its 3D Cartesian coordinates. The detector $f(\cdot)$ predicts a set of 3D bounding boxes $\{b_j\}_{j=1}^{N_{\text{obj}}}$ from $\mathbf{X}$, where N_{obj} is the number of detected objects. Each bounding box b_j is parameterized by its center location (x_j, y_j, z_j), dimension (w_j, h_j, l_j), heading angle θ_j, classification label c_j, and confidence score s_j. In the UDA task for 3D object detection defined by Track 5, the goal is to adapt a detector trained on a labeled source-domain dataset $\mathcal{D}^s = \{(\mathbf{X}_i^s, \mathbf{B}_i^s)\}_{i=1}^{N^s}$ (on "Vehicle" platform) to an unlabeled target-domain dataset $\mathcal{D}^t = \{\mathbf{X}_j^t\}_{j=1}^{N^t}$ (on "Drone" or "Quadruped" platform), where $\mathbf{B}_i^s$ denotes the ground-truth boxes of $\mathbf{X}_i^s$. Only the source-domain annotations are available for training, while the target domain provides point clouds without labels. The objective is to improve the generalization ability of $f(\cdot)$ on $\mathcal{D}^t$ despite the domain gap between $\mathcal{D}^s$ and $\mathcal{D}^t$.

Our method for cross-platform 3D object detection consists of three stages: source training, target training, and target testing. In source training, we enhance detectors with a set of physically aware augmentations to improve robustness. In target training, we adopt a self-training scheme with pseudo labels to adapt the detector to the unlabeled target domain. During target testing, we apply height-variation test-time augmentation to further mitigate domain gaps. The overall framework is shown in Figure 1. Regarding the target training stage, we directly adopt the ST3D framework [58], which leverages iterative self-training with pseudo labels to progressively adapt the detector to the unlabeled target domain. Also noted that all detection losses L_{det} in our framework follow the original detection loss of detectors as in [40, 41].

3.2. Physical-Aware Data Augmentation

To mitigate cross-platform discrepancies, we design a set of physical-aware augmentation strategies. These strategies explicitly model the scene, viewpoint, and height variations observed across vehicle-, drone-, and quadruped-mounted LiDAR platforms.

Scene-Aware Data Resampling. Certain classes exhibit domain-specific distribution biases depending on the deployment platform. For example, cars frequently appear in parking-lot environments in drone and quadruped scenarios. To better adapt the car detector, we enhance the training set by resampling parking-lot samples:

$$\mathcal{D}^{s'} = \mathcal{D}^s \cup \mathcal{R}_{\text{scene}}(\mathcal{D}^s \in \text{parking-lot}), \tag{1}$$

where $\mathcal{D}^s$ is the original source dataset and $\mathcal{R}_{\text{scene}}(\cdot)$ denotes the resampling operation conditioned on a specific

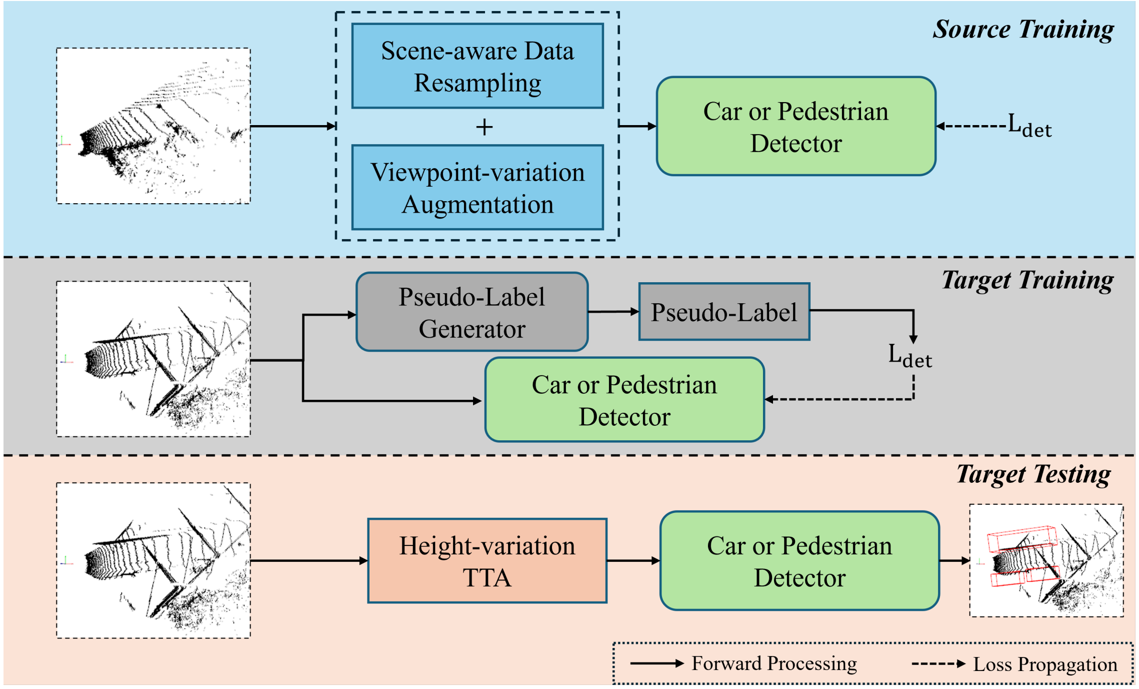

Figure 1. Overall framework of our proposed cross-platform 3D object detection method. In the **source training** stage, we employ scene-aware data resampling and viewpoint-variation augmentation to enhance the robustness of category-specific detectors (car or pedestrian). In the **target training** stage, pseudo labels are generated on the unlabeled target data and iteratively refined to adapt the detector through self-training. During **target testing**, we apply height-variation test-time augmentation (TTA) to further mitigate the domain gap caused by different sensor mounting heights.

scene type. This augmentation improves the robustness of detectors to scene-level distribution shifts.

Viewpoint-Variation Augmentation. Viewpoint differences across platforms lead to global geometric shifts in the captured point clouds. To simulate such variations, we augment the training data with random in-plane rotations along the x- and y-axes:

$$\mathbf{X}^s_{\text{rot}} = \mathcal{R}_{xy}(\mathbf{X}^s, \theta_x, \theta_y), \tag{2}$$

where $\mathcal{R}_{xy}(\cdot)$ applies rotations of (θ_x, θ_y) around the x and y axes. This augmentation exposes the detector to diverse viewpoints, enhancing its invariance to cross-platform rotations.

Height-variation test-time augmentation (TTA). Test-time augmentation (TTA) [3, 48] has been widely recognized as an effective strategy to mitigate domain shifts. Variations in mounting heights (*e.g.*, high-altitude drones versus low-mounted quadrupeds) and scene-dependent object behaviors (*e.g.*, pedestrians walking on streets versus standing on elevated stairs in plazas) introduce systematic observation biases. To address this, we adopt a height-aware TTA. Specifically, multiple height-shifted versions of the point cloud are generated and individually passed through the detector, and the results are fused via Non-Maximum Suppression (NMS):

$$\mathbf{B}^t = \text{NMS}\left(\bigcup_{\Delta h \in \mathcal{H}} \left(f(\mathbf{X}^t + \Delta h) - \Delta h \right) \right), \tag{3}$$

where $\mathcal{H}$ denotes the set of sampled height offsets and $f(\cdot)$ is the detector. This strategy reduces sensitivity to platform-specific mounting heights and improves robustness at inference.

3.3. Class-specific Model-ensembling

To enhance cross-platform robustness, we introduce a class-specific model ensembling strategy. Instead of training a single detector for all categories, we decouple the training process and construct individual detectors for different classes (*e.g.*, car and pedestrian). This design allows us to apply targeted augmentation strategies that explicitly account for the geometric and semantic characteristics of each class.

Source Training. At the stage of source training, for different categories, we design tailored augmentation strategies to match their characteristics. For instance, the car detector is enhanced by resampling data from parking-lot scenes to improve adaptation to drone and quadruped platforms, while the pedestrian detector benefits more from augmentations that emphasize viewpoint variation and spatial diversity.

Such class-specific designs allow each detector to better specialize in its own domain and improve robustness under cross-platform shifts.

Target Testing. At inference time, we further incorporate a class-specific TTA strategy. Leveraging the fact that cars and pedestrians exhibit distinct height distributions in point clouds, the TTA is applied in a category-aware manner, selectively enhancing predictions according to class-specific geometric priors.

Result Ensembling. Finally, the outputs of different detectors are seamlessly integrated through simple concatenation, producing the final set of detection results:

$$\mathbf{B}^t_{\text{final}} = \text{Concate}\big(f_{\text{car}}(\mathbf{X}^t),\ f_{\text{ped}}(\mathbf{X}^t)\big), \tag{4}$$

where $f_{\text{car}}(\cdot)$ and $f_{\text{ped}}(\cdot)$ denote the car and pedestrian detectors, respectively, and $\mathcal{B}_{\text{final}}$ represents the final set of detection results. This design enables flexible adaptation to class-specific distributions while maintaining a unified detection output without requiring additional post-processing overhead.

4. Experiments

4.1. Dataset

We use the official data provided by the *RoboSense Challenge 2025* [21] held at IROS 2025. This competition builds upon the legacy of the *RoboDepth Challenge 2023* [16, 19] at ICRA 2023 and the *RoboDrive Challenge 2024* [20, 54] at ICRA 2024, continuing the collective effort to advance robust and scalable robot perception. Each track in this competition is grounded on an established benchmark designed for evaluating real-world robustness and generalization [5, 9, 27, 30, 53]. Specifically, this task is built upon the **Pi3DET** benchmark [30] in **Track 5**, which studies cross-platform LiDAR-based 3D object detection across vehicle, drone, and quadruped platforms through viewpoint normalization and unified pre-training.

4.2. Experimental Setups

We follow the official settings of the RoboSense 2025 Track 5 Challenge, which evaluates cross-platform 3D object detection across vehicle-, drone-, and quadruped-mounted LiDARs. The challenge is divided into two phases.

In **Phase 1**: The detector is trained on labeled vehicle LiDAR data and adapted to unlabeled drone LiDAR data. The evaluation is conducted on the target drone domain using average precision (AP) of 40 recall points [43] with an IoU threshold of 0.5 for the Car class.

In **Phase 2**, the source domain remains vehicle LiDAR with annotations, while the target domain is quadruped LiDAR without labels. The final score is computed by averaging the AP of 40 recall points with an IoU threshold of 0.5 for the Pedestrian and Car classes.

Table 1. Performance Comparison in AP (%) on the RoboSense 2025 Track 5 Challenge benchmark for Phase 1 and Phase 2. Baseline denotes the baseline PV-RCNN detector [40] as in the official codebase released by the RoboSense 2025 Track 5 Challenge. The best performance in each setting is highlighted in **bold**.

	Methods	**Car**	**Pedestrian**	**Average**
Phase1	Baseline	46.64	-	-
	Ours	**62.25**	-	-
Phase2	Baseline	28.97	43.51	36.24
	Ours	**56.30**	**55.02**	**55.66**

4.3. Implementation Details

For both phases, our implementation is based on OpenPCDet [47], with ST3D [58] employed as the adaptation framework. We adopt the Adam_onecycle optimizer, using a learning rate of 0.01 for source training and 0.0005 for target training. The detailed settings for each phase are as follows:

Phase 1: We use PV-RCNN [40] as the baseline detector for the Car class. Source training is conducted for 36 epochs, where we apply *scene-aware data resampling* to enrich the dataset with parking-lot samples and *viewpoint-variation augmentation* with rotation angles (θ_x, θ_y) randomly sampled from $[-0.3925, 0.3925]$. Target training is performed for 6 epochs.

Phase 2: We adopt PV-RCNN++ [41] as the baseline detector for both Car and Pedestrian classes. Source training runs for 36 epochs. For the Car detector, we apply *scene-aware data resampling* with parking-lot samples and *viewpoint-variation augmentation* with $(\theta_x, \theta_y) \in [-0.2182, 0.1745]$. For the Pedestrian detector, we employ only *viewpoint-variation augmentation* with the same sampling range. Target training is carried out for 4 epochs. During target testing, we introduce *height-variation TTA*, where the Car detector uses $\mathcal{H} = \{-0.4, 0.0, 0.4\}$ and the Pedestrian detector uses $\mathcal{H} = \{-0.6, 0.0, 0.6\}$ for multi-height inference, followed by NMS fusion.

4.4. Comparative Study

Table 1 reports the quantitative results on the RoboSense 2025 Track 5 Challenge. In Phase 1, our method achieves 62.25% AP for the Car class, outperforming the baseline (46.64%) by a large margin. In Phase 2, our approach obtains 56.3% AP on Car and 55.02% AP on Pedestrian, leading to an overall average of 55.66%, which is substantially higher than the baseline average of 36.24%. These results demonstrate the effectiveness of our framework in improving cross-platform 3D object detection across both phases.

Table 2. Ablation study in AP (%) on the Car class in Phase 1. Baseline+ denotes the baseline method with height adjustment [60]. SA-DR denotes the proposed scene-aware data resampling, and VV-DA denotes the proposed viewpoint-variation data augmentation. The best result is highlighted in **bold**.

	Baseline+	**SA-DR**	**VV-DA**	**ST3D**	**Car**
(a)	✓				48.09
(b)	✓	✓			55.44
(c)	✓	✓	✓		59.06
(d)	✓	✓	✓	✓	**62.25**

Table 3. Ablation study in AP (%) on the Car class in Phase 2. Baseline+ denotes the baseline method with height adjustment [60]. SA-DR denotes the proposed scene-aware data resampling, and VV-DA denotes the proposed viewpoint-variation data augmentation. HV-TTA represents the proposed height-variation test-time augmentation. The best result is highlighted in **bold**.

	Baseline+	**VV-DA**	**SA-DA**	**ST3D**	**HV-TTA**	**Car**
(a)	✓					39.33
(b)	✓	✓				47.47
(c)	✓	✓	✓			51.95
(d)	✓	✓	✓	✓		54.32
(e)	✓	✓	✓	✓	✓	**56.30**

4.5. Ablation Study

Table 2 presents the ablation results on the Car class in Phase 1. Starting from the baseline with height adjustment, our proposed scene-aware data resampling brings a clear improvement by better modeling parking-lot scenarios. Incorporating viewpoint-variation augmentation further enhances robustness to cross-platform viewpoint shifts, yielding a notable gain. Finally, combining these strategies with ST3D achieves the best performance of 62.25% AP, demonstrating the complementary benefits of scene- and viewpoint-aware augmentations together with self-training adaptation.

Table 3 reports the ablation study on the Car class in Phase 2. Compared with the baseline, applying viewpoint-variation augmentation and scene-aware data augmentation significantly improves performance by addressing viewpoint and scene biases. Integrating ST3D further boosts the results through iterative self-training on unlabeled target data. Finally, the addition of height-variation test-time augmentation achieves the best performance of 56.30% AP, showing the effectiveness of explicitly modeling sensor height discrepancies during inference.

Table 4 shows the ablation results on the Pedestrian class in Phase 2. Starting from the baseline, applying viewpoint-variation augmentation already yields clear gains by improving robustness to viewpoint shifts. Incorporating ST3D further enhances adaptation through pseudo-label self-training. Finally, adding height-variation TTA achieves the best performance of 55.02% AP, demonstrating the effectiveness of leveraging class-specific height priors during inference.

Table 4. Ablation study in AP (%) on the Pedestrian class in Phase 2. Baseline+ denotes the baseline method with height adjustment [60]. VV-DA denotes the viewpoint-variation data augmentation, and HV-TTA denotes the proposed height-variation test-time augmentation. The best result is highlighted in **bold**.

	Baseline+	**VV-DA**	**ST3D**	**HV-TTA**	**Pedestrian**
(a)	✓				39.74
(b)	✓	✓			50.76
(c)	✓	✓	✓		52.87
(d)	✓	✓	✓	✓	**55.02**

5. Conclusion

In this work, we targeted the challenging problem of cross-platform 3D object detection in the RoboSense 2025 Track 5 Challenge. To mitigate the domain gaps introduced by different sensing platforms (vehicle, drone, and quadruped), we proposed a framework that integrates physical-aware augmentation and class-specific model ensembling to enhance geometric adaptability and semantic robustness across heterogeneous domains. Extensive experiments on both Phase 1 and Phase 2 benchmarks demonstrate that our method consistently outperforms strong baselines, achieving significant improvements across various categories. These results validate the effectiveness of our framework in addressing domain shifts in cross-platform LiDAR perception. In future work, we plan to extend our approach to multi-class detection with broader category coverage and to explore more unified adaptation strategies that can simultaneously capture platform-specific and class-specific variations.

References

[1] Hengwei Bian et al. Dynamiccity: Large-scale 4D occupancy generation from dynamic scenes. In *International Conference on Learning Representations*, 2025.

[2] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.

[3] Liang Chen, Yong Zhang, Yibing Song, Ying Shan, and Lingqiao Liu. Improved test-time adaptation for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24172–24182, 2023.

[4] Runnan Chen et al. Clip2scene: Towards label-efficient 3D scene understanding by CLIP. In *IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pages 7020–7030, 2023.

[5] Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. Towards natural language-guided drones: GeoText-1652 benchmark with spatial relation matching. In *European Conference on Computer Vision*, pages 213–231, 2024.

[6] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1201–1209, 2021.

[7] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuScenes: A large-scale benchmark for LiDAR panoptic segmentation and tracking. *IEEE Robotics and Automation Letters*, 7(2):3795–3802, 2022.

[8] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[9] Zeying Gong, Tianshuai Hu, Ronghe Qiu, and Junwei Liang. From cognition to precognition: A future-aware framework for social navigation. In *IEEE International Conference on Robotics and Automation*, pages 9122–9129, 2025.

[10] Xiaoshuai Hao, Mengchuan Wei, Yifan Yang, et al. Is your HD map constructor reliable under sensor corruptions? In *Advances in Neural Information Processing Systems*, pages 22441–22482, 2024.

[11] Xiaoshuai Hao, Guanqun Liu, Yuting Zhao, et al. Msc-bench: Benchmarking and analyzing multi-sensor corruption for driving perception. *arXiv preprint arXiv:2501.01037*, 2025.

[12] Xiaoshuai Hao et al. Safemap: Robust HD map construction from incomplete observations. In *International Conference on Machine Learning*, pages 22091–22102. PMLR, 2025.

[13] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12605–12614, 2020.

[14] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for LiDAR segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 228–240, 2023.

[15] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3D: Towards robust and reliable 3D perception against corruptions. In *IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023.

[16] Lingdong Kong, Yaru Niu, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, Liangjun Zhang, Hesheng Wang, Wei Tsang Ooi, Ruijie Zhu, Ziyang Song, Li Liu, Tianzhu Zhang, Jun Yu, Mohan Jing, Pengwei Li, Xiaohua Qi, Cheng Jin, Yingfeng Chen, Jie Hou, Jie Zhang, Zhen Kan, Qiang Lin, Liang Peng, Minglei Li, Di Xu, Changpeng Yang, Yuanqi Yao, Gang Wu, Jian Kuai, Xianming Liu, Junjun Jiang, Jiamian Huang, Baojun Li, Jiale Chen, Shuang Zhang, Sun Ao, Zhenyu Li, Runze Chen, Haiyong Luo, Fang Zhao, and Jingze Yu. The RoboDepth challenge: Methods and advancements towards robust depth estimation. *arXiv preprint arXiv:2307.15061*, 2023.

[17] Lingdong Kong, Niamul Quader, and Venice Erin Liong. Conda: Unsupervised domain adaptation for LiDAR segmentation via regularized domain concatenation. In *IEEE International Conference on Robotics and Automation*, pages 9338–9345, 2023.

[18] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised LiDAR semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21705–21715, 2023.

[19] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit R. Cottereau, and Wei Tsang Ooi. RoboDepth: Robust out-of-distribution depth estimation under corruptions. In *Advances in Neural Information Processing Systems*, pages 21298–21342, 2023.

[20] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Yaru Niu, Wei Tsang Ooi, Benoit R. Cottereau, Lai Xing Ng, Yuexin Ma, Wenwei Zhang, Liang Pan, Kai Chen, Ziwei Liu, Weichao Qiu, Wei Zhang, Xu Cao, Hao Lu, Ying-Cong Chen, Caixin Kang, Xinning Zhou, Chengyang Ying, Wentao Shang, Xingxing Wei, Yinpeng Dong, Bo Yang, Shengyin Jiang, Zeliang Ma, Dengyi Ji, Haiwen Li, Xingliang Huang, Yu Tian, Genghua Kou, Fan Jia, Yingfei Liu, Tiancai Wang, Ying Li, Xiaoshuai Hao, Yifan Yang, Hui Zhang, Mengchuan Wei, Yi Zhou, Haimei Zhao, Jing Zhang, Jinke Li, Xiao He, Xiaoqiang Cheng, Bingyang Zhang, Lirong Zhao, Dianlei Ding, Fangsheng Liu, Yixiang Yan, Hongming Wang, Nanfei Ye, Lun Luo, Yubo Tian, Yiwei Zuo, Zhe Cao, Yi Ren, Yunfan Li, Wenjie Liu, Xun Wu, Yifan Mao, Ming Li, Jian Liu, Jiayang Liu, Zihan Qin, Cunxi Chu, Jialei Xu, Wenbo Zhao, Junjun Jiang, Xianming Liu, Ziyan Wang, Chiwei Li, Shilong Li, Chendong Yuan, Songyue Yang, Wentao Liu, Peng Chen, Bin Zhou, Yubo Wang, Chi Zhang, Jianhang Sun, Hai Chen, Xiao Yang, Lizhong Wang, Dongyi Fu, Yongchun Lin, Huitong Yang, Haoang Li, Yadan Luo, Xianjing Cheng, and Yong Xu. The RoboDrive challenge: Drive anytime anywhere in any condition. *arXiv preprint arXiv:2405.08816*, 2024.

[21] Lingdong Kong, Shaoyuan Xie, Zeying Gong, Ye Li, Meng Chu, Ao Liang, Yuhao Dong, Tianshuai Hu, Ronghe Qiu, Rong Li, Hanjiang Hu, Dongyue Lu, Wei Yin, Wenhao Ding, Linfeng Li, Hang Song, Wenwei Zhang, Yuexin Ma, Junwei Liang, Zhedong Zheng, Lai Xing Ng, Benoit R. Cottereau, Wei Tsang Ooi, Ziwei Liu, Zhanpeng Zhang, Weichao Qiu, Wei Zhang, Ji Ao, Jiangpeng Zheng, Siyu Wang, Guang Yang, Zihao Zhang, Yu Zhong, Enzhu Gao, Xinhan Zheng, Xueting Wang, Shouming Li, Yunkai Gao, Siming Lan, Mingfei Han, Xing Hu, Dusan Malic, Christian Fruhwirth-Reisinger, Alexander Prutsch, Wei Lin, Samuel Schulter, Horst Possegger, Linfeng Li, Jian Zhao, Zepeng Yang, Yuhang Song, Bojun Lin, Tianle Zhang, Yuchen Yuan, Chi Zhang, Xuelong Li, Youngseok Kim, Sihwan Hwang, Hyeonjun Jeong, Aodi Wu, Xubo Luo, Erjia Xiao, Lingfeng Zhang, Yingbo Tang, Hao Cheng, Renjing Xu, Wenbo Ding, Lei Zhou, Long Chen, Hangjun Ye, Xiaoshuai Hao, Shuangzhi Li, Junlong Shen, Xingyu Li, Hao Ruan, Jinliang Lin, Zhiming Luo, Yu

Zang, Cheng Wang, Hanshi Wang, Xijie Gong, Yixiang Yang, Qianli Ma, Zhipeng Zhang, Wenxiang Shi, Jingmeng Zhou, Weijun Zeng, Kexin Xu, Yuchen Zhang, Haoxiang Fu, Ruibin Hu, Yanbiao Ma, Xiyan Feng, Wenbo Zhang, Lu Zhang, Yunzhi Zhuge, Huchuan Lu, You He, Seungjun Yu, Junsung Park, Youngsun Lim, Hyunjung Shim, Faduo Liang, Zihang Wang, Yiming Peng, Guanyu Zong, Xu Li, Binghao Wang, Hao Wei, Yongxin Ma, Yunke Shi, Shuaipeng Liu, Dong Kong, Yongchun Lin, Huitong Yang, Liang Lei, Haoang Li, Xinliang Zhang, Zhiyong Wang, Xiaofeng Wang, Yuxia Fu, Yadan Luo, Djamahl Etchegaray, Yang Li, Congfei Li, Yuxiang Sun, Wenkai Zhu, Wang Xu, Linru Li, Longjie Liao, Jun Yan, Benwu Wang, Xueliang Ren, Xiaoyu Yue, Jixian Zheng, Jinfeng Wu, Shurui Qin, Wei Cong, and Yao He. The RoboSense challenge: Sense anything, navigate anywhere, adapt across platforms. https://robosense2025.github.io, 2025.

[22] Lingdong Kong, Wesley Yang, Jianbiao Mei, Youquan Liu, Ao Liang, Dekai Zhu, Dongyue Lu, Wei Yin, Xiaotao Hu, Mingkai Jia, Junyuan Deng, Kaiwen Zhang, Yang Wu, Tianyi Yan, Shenyuan Gao, Song Wang, Linfeng Li, Liang Pan, Yong Liu, Jianke Zhu, Wei Tsang Ooi, Steven C. H. Hoi, and Ziwei Liu. 3D and 4D world modeling: A survey. *arXiv preprint arXiv:2509.07996*, 2025.

[23] Rong Li, Yuhao Dong, Tianshuai Hu, Ao Liang, et al. 3eed: Ground everything everywhere in 3D. *arXiv preprint arXiv:2511.01755*, 2025.

[24] Rong Li et al. Seeground: See and ground for zero-shot open-vocabulary 3D visual grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3707–3717, 2025.

[25] Shuangzhi Li, Zhijie Wang, Felix Juefei-Xu, Qing Guo, Xingyu Li, and Lei Ma. Common corruption robustness of point cloud detectors: Benchmark and enhancement. *IEEE Transactions on Multimedia*, 2023.

[26] Shuangzhi Li, Lei Ma, and Xingyu Li. Domain generalization of 3d object detection by density-resampling. In *European Conference on Computer Vision*, pages 456–473. Springer, 2024.

[27] Ye Li, Lingdong Kong, Hanjiang Hu, Xiaohao Xu, and Xiaonan Huang. Is your LiDAR placement optimized for 3D scene understanding? In *Advances in Neural Information Processing Systems*, pages 34980–35017, 2024.

[28] Ye Li et al. Optimizing LiDAR placements for robust driving perception in adverse conditions. *arXiv preprint arXiv:2403.17009*, 2024.

[29] Ao Liang et al. LiDARCrafter: Dynamic 4D world modeling from LiDAR sequences. *arXiv preprint arXiv:2508.03692*, 2025.

[30] Ao Liang et al. Perspective-invariant 3D object detection. In *IEEE/CVF International Conference on Computer Vision*, pages 27725–27738, 2025.

[31] Youquan Liu et al. Segment any point cloud sequences by distilling vision foundation models. In *Advances in Neural Information Processing Systems*, pages 37193–37229, 2023.

[32] Youquan Liu et al. Uniseg: A unified multi-modal LiDAR segmentation network and the openpcseg codebase. In *IEEE/CVF International Conference on Computer Vision*, pages 21662–21673, 2023.

[33] Youquan Liu et al. Multi-space alignments towards universal lidar segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14648–14661, 2024.

[34] Jiageng Mao, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 3d object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 131(8):1909–1963, 2023.

[35] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate LiDAR semantic segmentation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4213–4220, 2019.

[36] Uthman Olawoye and Jason N Gross. Uav position estimation using a lidar-based 3d object detection method. In *2023 IEEE/ION Position, Location and Navigation Symposium (PLANS)*, pages 46–51. IEEE, 2023.

[37] Rui Qian, Xin Lai, and Xirong Li. 3d object detection for autonomous driving: A survey. *Pattern Recognition*, 130: 108796, 2022.

[38] David Schinagl, Georg Krispel, Horst Possegger, Peter M. Roth, and Horst Bischof. OccAM's Laser: Occlusion-Based Attribution Maps for 3D Object Detectors on LiDAR Data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1141–1150, 2022.

[39] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019.

[40] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10529–10538, 2020.

[41] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. *International Journal of Computer Vision*, 131(2):531–551, 2023.

[42] Weijing Shi and Raj Rajkumar. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1711–1719, 2020.

[43] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1991–1999, 2019.

[44] Jiahao Sun, Chunmei Qing, Xiang Xu, et al. An empirical study of training state-of-the-art LiDAR segmentation models. *arXiv preprint arXiv:2405.14870*, 2024.

[45] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020.

[46] Muhammad Hassan Tanveer, Zainab Fatima, Hira Mariam, Tanazzah Rehman, and Razvan Cristian Voicu. Three-dimensional outdoor object detection in quadrupedal robots for surveillance navigations. In *Actuators*, page 422. MDPI, 2024.

[47] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. https://github.com/open-mmlab/OpenPCDet, 2020.

[48] Devavrat Tomar, Guillaume Vray, Jean-Philippe Thiran, and Behzad Bozorgtabar. Opttta: Learnable test-time augmentation for source-free medical image segmentation under domain shift. *Proceedings of Machine Learning Research*, 172: 1–26, 2022.

[49] Xuzhi Wang, Xinran Wu, Song Wang, et al. Monocular semantic scene completion via masked recurrent networks. In *IEEE/CVF International Conference on Computer Vision*, pages 24811–24822, 2025.

[50] Xuzhi Wang et al. Nuc-net: Non-uniform cylindrical partition network for efficient LiDAR semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(9):9090–9104, 2025.

[51] Yan Wang, Xiangyu Chen, Yurong You, Li Erran Li, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. Train in germany, test in the usa: Making 3d object detectors generalize. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11713–11723, 2020.

[52] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. RoboBEV: Towards robust bird's eye view perception under corruptions. *arXiv preprint arXiv:2304.06719*, 2023.

[53] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are VLMs ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. In *IEEE/CVF International Conference on Computer Vision*, pages 6585–6597, 2025.

[54] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and improving bird's eye view perception robustness in autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3878–3894, 2025.

[55] Xiang Xu et al. Frnet: Frustum-range networks for scalable LiDAR segmentation. *IEEE Transactions on Image Processing*, 34:2173–2186, 2025.

[56] Xiang Xu et al. Beyond one shot, beyond one perspective: Cross-view and long-horizon distillation for better lidar representations. In *IEEE/CVF International Conference on Computer Vision*, pages 25506–25518, 2025.

[57] Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[58] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10368–10378, 2021.

[59] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. St3d++: denoised self-training for unsupervised domain adaptation on 3d object detection. *arXiv preprint arXiv:2108.06682*, 2021.

[60] Bo Zhang, Jiakang Yuan, Botian Shi, Tao Chen, Yikang Li, and Yu Qiao. Uni3d: A unified baseline for multi-dataset 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9253–9262, 2023.