Towards Generalizable 3D Object Detection Across Sensor Placements

KeXin Xu
University of Alberta
Edmonton, Alberta, Canada
kxu10@ualberta.ca

Abstract

Reliable 3D object detection is essential for autonomous driving, yet the performance of existing LiDAR-based detectors can degrade significantly when sensors are mounted at different positions across vehicle platforms. This sensitivity to sensor placement presents a major challenge for scalable and widely deployable perception systems. To address this issue, we introduce a refined LiDAR-only detection framework built upon the BEVFusion architecture. Our approach incorporates two complementary strategies. First, we adopt multi-sweep aggregation to fuse several consecutive LiDAR frames, producing a denser and more temporally consistent spatial representation. Second, we apply placement-mixed training, which exposes the model to a diverse set of sensor configurations during training and promotes the learning of placement-agnostic geometric features. These components are seamlessly integrated into the BEVFusion LiDAR pipeline without requiring complex architectural modifications. Experimental results demonstrate that this dual-strategy framework substantially improves both detection accuracy and robustness under varying sensor placements, highlighting its potential for practical deployment in real-world autonomous driving scenarios.

1. Introduction

Accurate and reliable 3D object detection lies at the core of autonomous driving perception systems [1, 3–5, 7, 8, 16, 47, 53]. LiDAR-based detectors, in particular, have demonstrated strong performance due to their ability to capture high-fidelity geometric structures from point clouds [2, 10, 14, 33, 34, 36, 37, 39, 42, 44, 45, 48, 49, 54]. Despite this progress, a key limitation persists: most existing detectors implicitly assume a fixed and consistent LiDAR sensor placement. In real-world deployments, however, such assumptions rarely hold [31, 32].

Differences in vehicle design, manufacturing tolerances, and aftermarket modifications often result in diverse sensor mounting positions across platforms [13, 30, 31, 41, 43, 55]. Even minor deviations in orientation or height can introduce substantial geometric discrepancies in the captured point clouds, leading to notable degradation in model performance and hindering the scalability of perception systems across heterogeneous fleets.

This challenge exposes an important research gap. While prior work has devoted considerable effort to improving accuracy under a static sensor setup, robustness to sensor placement variability has received relatively limited attention [15, 17, 19, 22, 27, 50, 57]. Yet, such robustness is crucial for real-world autonomous driving, where a perception model must be deployable on multiple vehicle platforms without expensive re-training or manual calibration adjustments [24, 30, 46]. Ensuring consistent detection quality across varying LiDAR placements is a foundational requirement for building scalable and reliable autonomous systems [11, 12, 20, 25, 26, 28, 35, 38].

In this challenge, we develop a LiDAR-only enhancement of the BEVFusion framework [40] aimed at improving generalization under diverse sensor configurations. Our approach centers on two complementary strategies designed to address both geometric sparsity and placement-induced domain shifts. First, we employ multi-sweep aggregation, which fuses several consecutive LiDAR frames to create a denser and more temporally coherent spatial representation. This temporal enrichment helps mitigate occlusions and reduces sensitivity to sparse or incomplete single-frame observations. Second, we introduce placement-mixed training, where the model is trained using LiDAR data collected from multiple sensor mounting positions [29]. This exposure encourages the network to learn viewpoint-invariant representations that generalize naturally across different sensor layouts.

By integrating these strategies directly into the LiDAR branch of BEVFusion [40] without modifying its core architecture [56], our method significantly enhances detection robustness under sensor placement variability. The resulting system offers a practical, lightweight, and scalable solution for improving cross-platform consistency in autonomous driving perception.

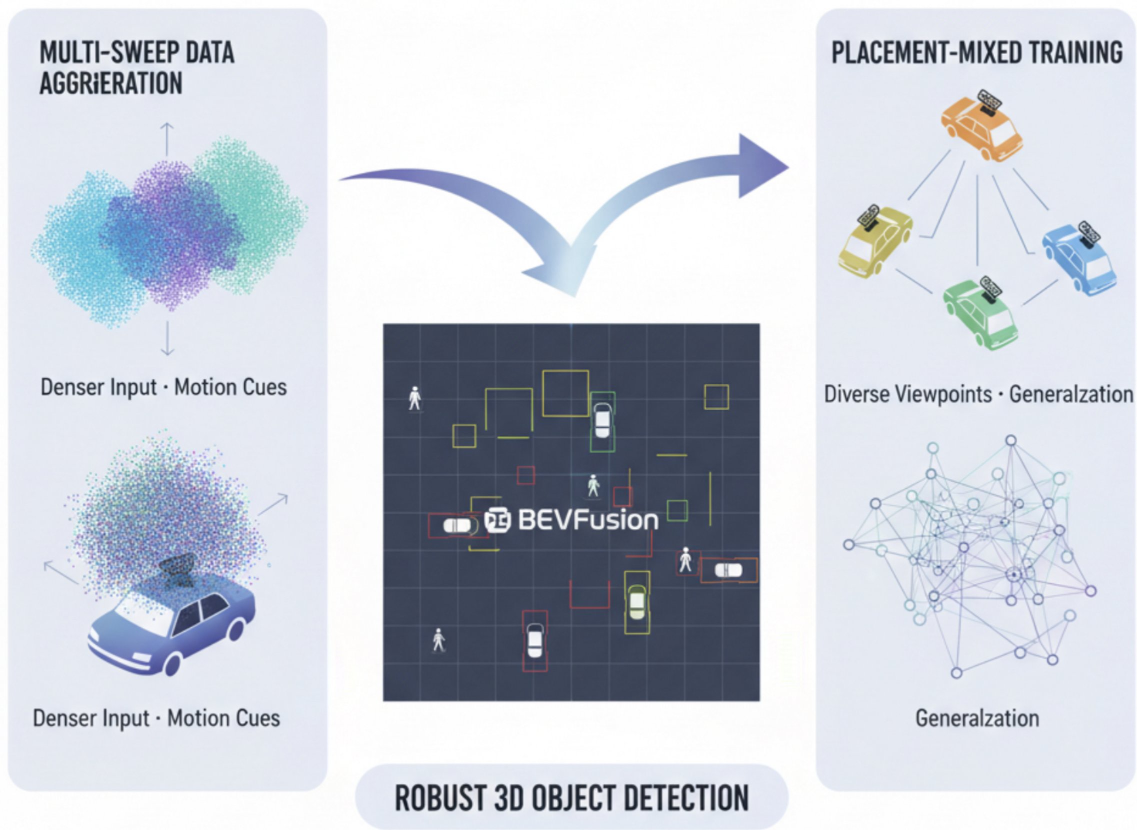

Figure 1. Overview of the Multi-Sweep Data Aggregation and Placement-Mixed Training framework.

2. Methodology

Our methodology enhances the BEVFusion framework through data-centric strategies designed to improve robustness against sensor placement variations. The approach builds upon the LiDAR branch of BEVFusion, a widely adopted architecture for 3D perception in the bird's-eye-view (BEV) space. The standard BEVFusion pipeline first converts raw 3D point clouds into a voxelized grid representation, after which a 3D backbone network extracts spatial and semantic features from the voxel tensor. These features are subsequently projected onto a unified BEV map and passed through a detection head to output 3D bounding boxes for objects in the scene. This design effectively leverages the geometric precision of LiDAR data and has demonstrated strong performance across various perception benchmarks (see Fig. 1).

In this work, we retain the core BEVFusion architecture and introduce two complementary enhancements that improve its resilience under diverse LiDAR mounting configurations: (1) multi-sweep data aggregation, which addresses sparsity and occlusion challenges in single-frame point clouds, and (2) placement-mixed training, which encourages viewpoint-invariant feature learning by exposing the model to multiple sensor configurations during training. Both strategies are applied without modifying the underlying network design, ensuring compatibility and simplicity.

2.1. Multi-Sweep Data Aggregation

Single-frame LiDAR scans often exhibit sparsity, missing surfaces, and occlusions caused by dynamic agents, all of which hinder reliable object detection. To mitigate these limitations, we incorporate *multi-sweep data aggregation* prior to voxelization. Specifically, multiple consecutive LiDAR sweeps are temporally aligned to the current frame's coordinate system using ego-motion compensation. These aligned sweeps are then merged to form a denser and more complete point cloud.

This temporally enriched representation provides two major benefits. First, it significantly increases point density on distant or partially observed objects, which improves feature quality in both the voxel backbone and BEV head. Second, it captures short-term motion cues from dynamic objects, enabling more stable and consistent representation learning across frames. Importantly, this aggregation process occurs entirely at the data level and does not require architectural modifications or additional network parameters. By supplying denser and more informative geometry to the BEVFusion backbone, the detector achieves improved robustness even under challenging scenarios where single-frame observations are incomplete.

2.2. Placement-Mixed Training

To explicitly improve generalization across different sensor mounting positions, we introduce *placement-mixed training*. Instead of training solely on data collected from a single LiDAR placement, we construct a unified training split that combines point clouds obtained from multiple mounting configurations. This exposes the model to diverse viewpoints, scanning heights, and geometric distributions during training.

By simulating placement variability as part of the training distribution, the model is encouraged to learn viewpoint-invariant geometric representations rather than overfitting to a specific sensor configuration. This strategy effectively serves as a real-world data augmentation mechanism, mirroring the type of domain shifts encountered when deploying perception models on heterogeneous fleets of vehicles. As a result, the detector becomes more resilient to differences in LiDAR extrinsics, enabling robust inference on previously unseen platforms without requiring additional fine-tuning or calibration.

Taken together, multi-sweep data aggregation enriches the geometric input to the model, while placement-mixed training enhances its ability to generalize across sensor layouts. These enhancements operate entirely within the LiDAR branch of BEVFusion and significantly improve robustness without modifying the model's architecture.

3. Experiments

3.1. Dataset

We use the official data provided by the *RoboSense Challenge 2025* [23] held at IROS 2025. This competition builds upon the legacy of the *RoboDepth Challenge 2023* [18, 21] at ICRA 2023 and the *RoboDrive Challenge 2024* [22, 52] 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 [6, 9, 31, 34, 51]. Specifically, this task is built upon the **Place3D** dataset [31] in **Track 3**, which provides a standardized foundation for benchmarking performance under challenging conditions such as cross-domain shifts, sensor variability, and multi-modal alignment.

3.2. Implementation Details

Our method was implemented based on the Track3 open source github repo. We used the standard BEVFusion as our baseline. All training was conducted using the standard settings provided by RoboSense, which include optimizers, learning rate schedules, and data processing pipelines. Multi-sweep aggregation was achieved by merging consecutive frames during data preprocessing, and placement-mixed training was implemented by combining data from different sensor configurations into the training set.

Method	**mAP**
BEVFusion	0.605
+ Multi Sweep	0.729
+ Multi Sweep + Mix Placements	0.743

Table 1. Comparative Results.

3.3. Evaluation Protocol

Performance was assessed using the official RoboSense Track3 evaluation metrics, with a primary focus on 3D mean Average Precision (mAP).

3.4. Comparative Results

Our enhanced model demonstrated superior performance compared to the baseline BEVFusion configuration, showing clear gains in variable placement scenarios. The results indicate that our strategies lead to significant improvements in robustness and generalization.

4. Conclusion

In this work, we introduced an enhanced LiDAR-only 3D object detection framework aimed at improving robustness against variations in sensor placement. By integrating multi-sweep data aggregation and placement-mixed training into the BEVFusion model, our approach effectively addresses challenges related to data sparsity and sensor variability. The experimental results clearly show that each strategy contributes positively, with their combination delivering the most accurate and stable performance across different sensor configurations. As these enhancements do not require fundamental changes to the model architecture, they offer a practical and readily deployable solution for real-world autonomous driving systems.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In *IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019. 1

[2] Hengwei Bian et al. DynamicCity: Large-scale 4D occupancy

generation from dynamic scenes. In *International Conference on Learning Representations*, 2025. 1

[3] 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. 1

[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] Runnan Chen et al. Towards label-free scene understanding by vision foundation models. In *Advances in Neural Information Processing Systems*, pages 75896–75910, 2023. 1

[6] 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. 3

[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. 1

[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. 1

[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. 3

[10] Xiaoshuai Hao, Mengchuan Wei, Yifan Yang, Haimei Zhao, Hui Zhang, Yi Zhou, Qiang Wang, Weiming Li, Lingdong Kong, and Jing Zhang. Is your HD map constructor reliable under sensor corruptions? *Advances in Neural Information Processing Systems*, 37:22441–22482, 2024. 1

[11] Xiaoshuai Hao, Hui Zhang, Yifan Yang, Yi Zhou, Sangil Jung, Seung-In Park, and ByungIn Yoo. MBFusion: A new multi-modal bev feature fusion method for HD map construction. In *IEEE International Conference on Robotics and Automation*, pages 15922–15928, 2024. 1

[12] 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. 1

[13] Xiaoshuai Hao et al. SafeMap: Robust HD map construction from incomplete observations. In *International Conference on Machine Learning*, pages 22091–22102. PMLR, 2025. 1

[14] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11108–11117, 2020. 1

[15] 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. 1

[16] 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. 1

[17] 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. 1

[18] 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. 3

[19] 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. 1

[20] 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. 1

[21] 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. 3

[22] 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. 1, 3

[23] 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. 3

[24] Lingdong Kong, Xiang Xu, Jun Cen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Calib3D: Calibrating model preferences for reliable 3D scene understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1965–1978, 2025. 1

[25] Lingdong Kong, Xiang Xu, Youquan Liu, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. LargeAD: Large-scale cross-sensor data pretraining for autonomous driving. *arXiv preprint arXiv:2501.04005*, 2025. 1

[26] Lingdong Kong, Xiang Xu, Jiawei Ren, et al. Multi-modal data-efficient 3D scene understanding for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3748–3765, 2025. 1

[27] 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. 1

[28] Rong Li, Yuhao Dong, Tianshuai Hu, Ao Liang, et al. 3EED: Ground everything everywhere in 3D. *arXiv preprint arXiv:2511.01755*, 2025. 1

[29] 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. 1

[30] 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. 1

[31] 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. 1, 3

[32] Ye Li et al. Optimizing LiDAR placements for robust driving perception in adverse conditions. *arXiv preprint arXiv:2403.17009*, 2024. 1

[33] Ao Liang, Youquan Liu, Yu Yang, et al. LiDARCrafter: Dynamic 4D world modeling from LiDAR sequences. *arXiv preprint arXiv:2508.03692*, 2025. 1

[34] Ao Liang et al. Perspective-invariant 3D object detection. In *IEEE/CVF International Conference on Computer Vision*, pages 27725–27738, 2025. 1, 3

[35] Venice Erin Liong, Thi Ngoc Tho Nguyen, Sergi Widjaja, Dhananjai Sharma, and Zhuang Jie Chong. AMVNet: Assertion-based multi-view fusion network for LiDAR semantic segmentation. *arXiv preprint arXiv:2012.04934*, 2020. 1

[36] 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. 1

[37] 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. 1

[38] 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. 1

[39] Youquan Liu et al. La La LiDAR: Large-scale layout generation from LiDAR data. *arXiv preprint arXiv:2508.03691*, 2025. 1

[40] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 1

[41] 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. 1

[42] 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. 1

[43] Rui Qian, Xin Lai, and Xirong Li. 3d object detection for autonomous driving: A survey. *Pattern Recognition*, 130: 108796, 2022. 1

[44] 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. 1

[45] 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. 1

[46] 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. 1

[47] 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. 1

[48] 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. 1

[49] 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. 1

[50] Aoran Xiao, Jiaxing Huang, Dayan Guan, Kaiwen Cui, Shijian Lu, and Ling Shao. PolarMix: A general data augmentation technique for LiDAR point clouds. In *Advances in Neural Information Processing Systems*, pages 11035–11048, 2022. 1

[51] 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. 3

[52] 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. 3

[53] Jingyi Xu, Weidong Yang, Lingdong Kong, Youquan Liu, Qingyuan Zhou, Rui Zhang, Zhijun Li, Wen-Ming Chen, and Ben Fei. Visual foundation models boost cross-modal unsupervised domain adaptation for 3d semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 26 (11):20287–20301, 2025. 1

[54] Xiang Xu et al. 4D contrastive superflows are dense 3D representation learners. In *European Conference on Computer Vision*, pages 58–80, 2024. 1

[55] Xiang Xu et al. FRNet: Frustum-range networks for scalable LiDAR segmentation. *IEEE Transactions on Image Processing*, 34:2173–2186, 2025. 1

[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. 1

[57] Yan Yan, Yuxing Mao, and Bo Li. SECOND: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 1