RobusTor3D: Robust Multimodal 3D Object Detector for Autonomous Driving by Vision-Language Knowledge Blending

Ying Yang[1,2,3], **Hui Yin**[1,2,4*], **Aixin Chong**[5], **Hui Wang**[6] **Zhengyin Liang**[3,7]

[1]State Key Laboratory of Advanced Rail Autonomous Operation, Beijing Jiaotong University
[2]Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, Beijing Jiaotong University
[3]School of Computer Science & Technology, Beijing Jiaotong University
[4]Frontiers Science Center for Smart High-speed Railway System, Beijing Jiaotong University
[5] School of Computer Science & Technology, Shandong University of Technology
[6]School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast
[7]Key Laboratory of Beijing for Railway Engineering, Beijing Jiaotong University
{22110126, hyin, lzyinxx}@bjtu.edu.cn, aixinchong@sdut.edu.cn, h.wang@qub.ac.uk

Abstract

Multimodal 3D object detection for autonomous driving, a task for real-world applications, poses substantial challenges in maintaining robust performance under various perturbations and complex environmental conditions. However, most existing approaches primarily focus on performance optimization under relatively ideal scenarios or focus on one or few disturbing conditions (or adverse conditions), lacking systematic exploration of robustness against real-world factors, including high class imbalance, adverse weather conditions, sensor jitter and failures, and significant scene variations. To address this issue, we propose a robust multimodal 3D detector, termed RobusTor3D, which integrates robustness at both the structural and supervisory levels by blending the knowledge from Vision-Language Models (VLMs). Structurally, textual descriptions are incorporated to enhance the semantic richness and diversity of rare classes. This novel semantic injection operation compensates for the inherent class imbalance and modality weakness in conventional visual features. Furthermore, semantic alignment capability and robust representation by Vision-Language Knowledge Extraction (V-LKE) serve as semantic priors to complement modality-specific representations, significantly improving model adaptability. At the supervisory level, we propose a Scene-level Multimodal Consistency Learning (SMCL) strategy, which jointly enforces global semantic constraints across modalities, encouraging the learning of stable and abundant semantic representations. This special design specifically reduces the impact of spatial alignment, while notably enabling semantic compensation under modality-loss conditions. Extensive robustness experiments conducted on KITTI, KITTI-C, and CADC benchmarks evaluate five robustness aspects, including long-tail problem, adverse weather (rain, snow, fog, strong sunlight), sensor spatial misalignment and motion blur, modality loss, and cross-domain scenarios. The results show that RobusTor3D demonstrates superior robustness across all five evaluated aspects. It consistently outperforms the state-of-the-art methods under various challenging conditions.

Code — https://github.com/yingyang-1/RobusTor3D

*Corresponding Author

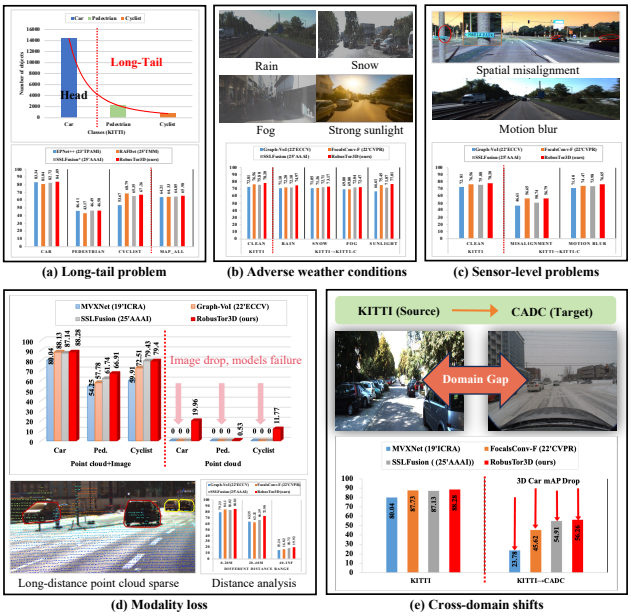


Figure 1: Five key aspects for evaluating model robustness. (a) Long-tail problem, (b) Diverse adverse weather conditions, (c) Sensor-level problems, (d) Modality loss, and (e) Cross-domain shifts. It shows that our RobusTor3D has optimal robustness under various challenging conditions.

Introduction

Multimodal 3D object detection is a core task of autonomous driving perception systems and requires accuracy and robustness. Numerous multimodal 3D object detection algorithms have been developed to improve 3D perception performance (Peng et al. 2022) in recent years. However, most existing methods focus on performance optimization under relatively ideal conditions, only pursuing the performance of a specific indicator while ignoring the overall performance (Sindagi, Zhou, and Tuzel 2019; Vora et al. 2020; Liu et al. 2022; Yang et al. 2022; Chen et al. 2022; Li et al. 2025). In the real world, maintaining robust perception un-

der various disturbances and complex environmental conditions faces great challenges (Chen et al. 2024). Figure 1 illustrates five key aspects for evaluating model robustness: (a) **Long-tail problem**. In the dynamic real-world environment of autonomous driving, multimodal data presents a serious long-tail distribution (Zhang et al. 2023) (such as multimodal dataset KITTI (Geiger, Lenz, and Urtasun 2012)). The long-tail distribution leads to significant performance differences between the head and tail categories. (b) **Various adverse weather conditions**. Such conditions can cause sensor data degradation, which poses a huge challenge to the perception model. (c) **Sensor-level problems**. Vehicle motion or sensor jitter can cause spatial misalignment between different sensors (Song et al. 2024) and motion blur problems. (d) **Modality loss**. Due to factors such as hardware, environment, or synchronization, sensor modality may be lost (including the long-distance sparsity of point cloud), affecting the perception system's perception capabilities in normal environments. As shown in Figure 1(d), we simulate the loss of image modality, and some multimodal detection models fail completely. (e) **Cross-domain shifts**. As shown in Figure 1(e), due to the obvious domain gap between the source domain and the target domain in terms of modal features, semantic structure, environmental conditions, etc., the performance of most models drops significantly when tested across datasets. Common adverse conditions are inevitable in a dynamic real-world environment of autonomous driving and can significantly impact performance differences.

In autonomous driving, 3D perception model needs comprehensive robustness. Although some methods consider single (Pan et al. 2024; Ding et al. 2025; Song et al. 2023; Chang et al. 2024; Hegde et al. 2025a) or several challenging conditions (Yoon, Jung, and Yoo 2025), almost all models fail to meet comprehensive robustness requirements simultaneously. Moreover, although some methods (Song et al. 2024; Dong et al. 2023; Yu et al. 2023) consider robustness, they only evaluate and analyze robustness, without proposing systematic solutions or model designs for the comprehensive challenges. The above five challenges arise from different perspectives, but effectively utilizing the knowledge of VLMs can alleviate these challenges. In recent years, the rich knowledge of VLMs, aided by textual modalities, provides insights into solving long-tail problems (Zhang et al. 2024b), generalization (Huang et al. 2024), domain adaptation (Wu et al. 2024), open-set problems (Greer et al. 2025), et.al. However, how to introduce VLMs to improve the overall robustness of 3D detector remains a major challenge.

In this paper, we comprehensively explore and alleviate the above robustness issue by exerting the advantages of VLMs to improve model robustness. Considering the characteristics of VLMs with cross-modal alignment and universal representation learning, as well as the multimodal alignment problem, we propose a robust multimodal 3D detection detector, RobusTor3D, which improves the robustness of multimodal 3D detection from both structural and supervisory perspectives. Structurally, we employ Multimodal Large Language Models (MLLMs) to generate textual descriptions from images offline to balance class distribution and enhance tail-class semantics. Then, the designed Vision Language Knowledge Extraction (V-LKE) provides robust and extensive semantic priors to complement modality-specific representations, significantly improving model adaptability under environmental variations and modality degradation. In the supervisory level, to enhance cross-modal alignment robustness and semantic consistency, we propose a scene-level multimodal consistency learning (SMCL) strategy. SMCL jointly enforces global semantic constraints across modalities by two consistency contrastive objectives, encouraging the model to learn stable and complementary semantic representations. In addition, we conduct extensive robustness experiments to demonstrate the comprehensive robustness of our RobusTor3D.

The main contribution is summarized as follows:

- A robust multimodal 3D detector, termed RobusTor3D, is proposed by Vision-Language Knowledge Blending, implemented by introducing VLMs from structure and supervision design.
- Structurally, textual descriptions are incorporated to enhance the semantic richness and diversity of rare classes, robust and extensive prior knowledge by Vision-Language Knowledge Extraction (V-LKE) to complement modality-specific representations.
- In supervision, a Scene-level Multi-modal Consistency Learning (SMCL) strategy is designed to learn stable and complementary semantic representations by jointly enforcing global semantic constraints across modalities.
- The proposed RobusTor3D showcases superior robustness under various complex interference conditions, surpassing the state-of-the-art detectors.

Related Works

Multimodal 3D Object Detection

In autonomous driving, multimodal 3D object detection has gained more attention and witnessed advancements. Although the latest methods can achieve high accuracy, these methods ignore the model robustness issue (Song et al. 2024). Some approaches take into account specific or few challenging conditions, such as long-tail problem (Pan et al. 2024), adverse weather conditions (Yoon, Jung, and Yoo 2025), spatial misalignment (Ding et al. 2025; Song et al. 2023; Yoon, Jung, and Yoo 2025), or cross-domain problem (Yoon, Jung, and Yoo 2025; Chang et al. 2024; Hegde et al. 2025a), However, current works still lack systematic robustness exploration of 3D detectors, which are difficult to adapt to complex, varying, and challenging scenes.

VLMs for 3D Object Detection

Recently, VLMs, especially CLIP (Radford et al. 2021), have been applied to a variety of tasks in autonomous driving, including reasoning and planning, prediction, simulation, and testing (Gao et al. 2024), as well as 3D object detection (Sapkota et al. 2025). Most existing methods introduce text/category prompts and use VLMs to extract text embeddings to enhance detection performance in a distillation manner (Zhang et al. 2024a; Hegde et al. 2025b; Jiao et al. 2024). Different from these methods, this paper utilizes

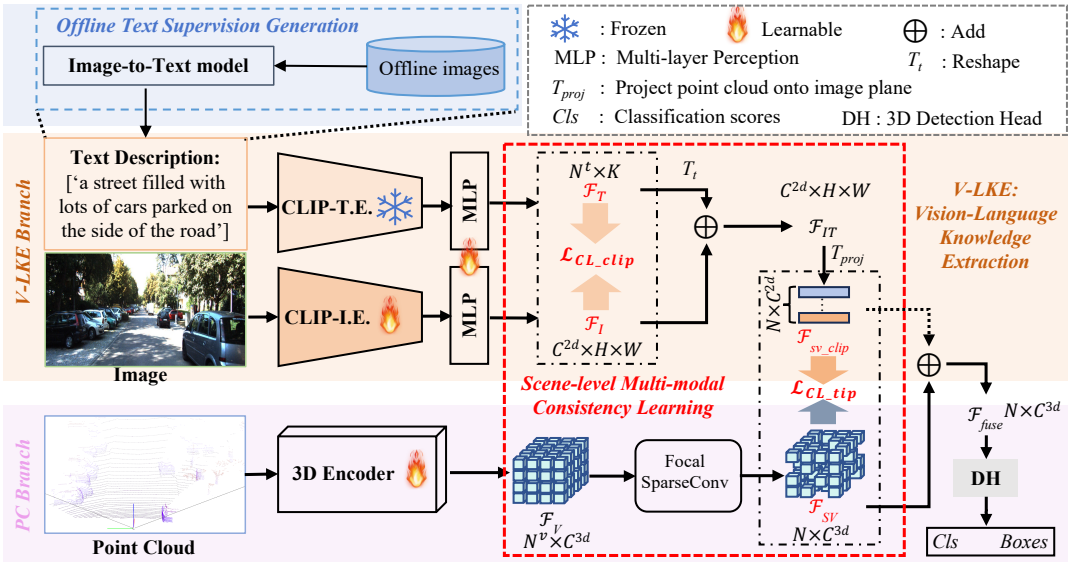


Figure 2: The overall architecture of RobusTor3D. Structurally, it consists of three modules, (1) Offline Text Supervision Generation from images by image-to-text model; (2) Vision-Language Knowledge Extraction branch to extract semantic prior knowledge in image-text space features; (3) Point Cloud (PC) branch to extract sparse voxel features, and the 3D detection head generates 3D detection results based on the fused features. In supervision, the designed Scene-level Multimodal Consistent Learning strategy with two contrastive objectives jointly aligns text and vision features in a scene-level aligned manner.

the rich knowledge of VLMs to enhance the overall robustness of 3D detectors under diverse challenging conditions through structure and supervision design.

Methods

Framework Overview

The overall architecture of RobusTor3D with both structural and supervisory level design is shown in Figure 2. Structurally, three modules: (1) an Offline Text Supervision Generation (OTSG) module: text descriptions are produced from images offline by an Image-to-Text model, (2) a Vision-Language Knowledge Extraction (V-LKE) branch utilizes pre-trained CLIP to extract prior representation $\mathcal{F}_{sv\_clip}$, (3) Point Cloud (PC) branch containing 3D sparse convolutions to extract sparse voxel features $\mathcal{F}_{SV}$(N denotes the number of sparse voxels), and a 3D detection head generates 3D locations and classification scores based on the fused features $\mathcal{F}_{fuse}$. In supervision, a Scene-level Multimodal Consistent Learning strategy (SMCL) is embedded with two contrastive objectives to jointly align text, image, and voxel features.

Offline Text Supervision Generation

Based on existing multimodal datasets (KITTI and CADC) in autonomous driving, we utilize ViT-GPT2 to generate corresponding text descriptions from images, providing auxiliary semantic supervision. The KITTI dataset includes LiDAR point clouds and the corresponding front-view images. We employ ViT-GPT2 with an empty prompt to produce text descriptions of images, which summarize the entire image scene, spatial location information, and context information(samples in Fig.1 of the **Supp. material**). The generated description of different classes appears only once but contains quantitative descriptions, which is fair to all classes. We use different models to generate text descriptions and conduct experimental comparisons, showing that ViT-GPT2 is more effective(Fig.2 of the **Supp. material**)). The CADC dataset includes LiDAR point clouds and corresponding eight-view images. We employ ViT-GPT2 to produce text descriptions of the front-view image. Consequently, we construct the triplet input, consisting of point clouds, images, and text descriptions.

Vision-Language Knowledge Extraction

V-LKE is designed to extract robust semantic representation $\mathcal{F}_{sv\_clip}$ in the image-text space. $\mathcal{F}_{sv\_clip}$ serves as prior knowledge to compensate for modality-specific representations, improve the model's adaptability to modality degradation and cross-domain shifts.

Specifically, the rich and extensive semantics captured by the pre-trained CLIP (Radford et al. 2021) compensate for the more serious loss of long-tail category features caused by sparse point clouds, alleviating the quantitative gap between the head category and the long-tail category information. Different from the original CLIP image encoder, we extract the first-stage features to avoid the semantic loss due to the multiple downsamplings, and remove the final attention pool operation, which might harm the downstream fine-grained visual understanding, to preserve more spatial structures of the image. We also employ the CLIP text encoder to extract the text features. To make the representation extracted by CLIP suitable for our task, we chose to freeze the text encoder but not the image encoder because the struc-

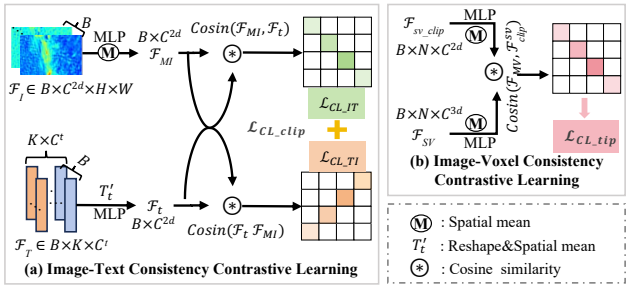


Figure 3: Structure of Scene-level Multimodal Consistent Learning with joint contrastive learning (SMCL).

ture of the text encoder is relatively complex, and as confirmed by performance comparison, frozen T.E.(78.2) is a wiser choice than frozen I.E.(76.3). We added a two-layer learnable multi-layer perceptions(MLP) after both to obtain learnable image-text features $\mathcal{F}_I$ and $\mathcal{F}_T$.

$$\mathcal{F}_T, \mathcal{F}_I = MLP_{T/I}(CLIP(T, I)), \tag{1}$$

Since we use contrastive learning to ensure consistency between $\mathcal{F}_I$ and $\mathcal{F}_T$, after projecting the text features into the image feature space by T_t including two linear layers, we obtain the joint prior knowledge $\mathcal{F}_{IT}$ by addition. Then $\mathcal{F}_{IT}$ are transformed into voxel-wise features $\mathcal{F}_{sv\_clip}$ through the 3D-2D projection matrix T_{proj}, which are in the same feature space as $\mathcal{F}_{SV}$. The final fused features $\mathcal{F}_{fuse}$ are obtained by addition.

$$\begin{gathered}\mathcal{F}_{IT} = \mathcal{F}_T \bigoplus T_t(\mathcal{F}_T), \mathcal{F}_{sv\_clip} = T_{proj}(\mathcal{F}_{IT}), \\ \mathcal{F}_{fuse} = \mathcal{F}_{sv\_clip} \bigoplus \mathcal{F}_{SV},\end{gathered} \tag{2}$$

where $\bigoplus$ denotes the addition operation.

Scene-level Multimodal Consistent Learning

SMCL promotes feature alignment and learns stable and complementary semantic representations by constraining the global semantic consistency of multiple modalities at the scene level. Unlike the projected relationships between point clouds and images, there is no direct correspondence between point clouds and texts. Therefore, as shown in Figure 3, we design Image-Text consistency contrastive learning (ITCCL) and Image-Voxel consistency contrastive learning (IVCCL), which fulfill consistency learning between $\mathcal{F}_I$ and $\mathcal{F}_T$, $\mathcal{F}_{sv\_clip}$ with $\mathcal{F}_{SV}$. These contrastive mechanisms pull together heterogeneous modality features from the same scene in a shared semantic space, reducing the impact of cross-modal misalignment and modality loss.

Image-Text Consistency Contrastive Learning. ITCCL is designed to shorten the distance between an image feature and its corresponding text feature in a batch and to increase the distance with text features in other scenes, thereby enhancing the consistency and encouraging the learning of stable and rich semantic representations in image-text space. Specifically, first, the image and text features are projected to the same representation space through global spatial mean (SM) and T_t, respectively, and then they are passed through a two-layer MLP to obtain the representation vectors with the same dimension.

$$\mathcal{F}_{MI} = MLP(SM(\mathcal{F}_I)), \mathcal{F}_t = MLP(T'_t(\mathcal{F}_T)), \tag{3}$$

where T'_t includes reshape and global spatial mean operations. The cosine similarity and cross-entropy loss of both are calculated and averaged to obtain the final loss $\mathcal{L}_{CL\_clip}$.

$$\begin{aligned}\mathcal{L}_{CL\_clip} = -(\sum_j^B log\frac{\exp(Sim(\mathcal{F}^j_{MI},\mathcal{F}^j_t)/\tau)}{\sum_k^B \exp(Sim(\mathcal{F}^j_{MI},\mathcal{F}^k_t)/\tau)} + \\ \sum_j^B log\frac{\exp(Sim(\mathcal{F}^j_t,\mathcal{F}^j_{MI})/\tau)}{\sum_k^B \exp(Sim(\mathcal{F}^j_t,\mathcal{F}^k_{MI})/\tau)})/2,\end{aligned} \tag{4}$$

where B, *Sim*, denote batch size, cosine similarity, respectively, and τ is the temperature for the cross-modal matching to modulate the loss.

Image-Voxel Consistency Contrastive Learning. The projected features F_{sv_clip} are already in the same feature space with F_{SV}. Similar to ITCCL, we use contrastive learning to bring F_{SV} and F_{sv_clip} of the same scene closer together. F_{sv_clip} and F_{SV} are respectively passed through SM and a two-layer MLP. But only the cosine similarity and cross-entropy loss of the voxel features and clip features are calculated to obtain the loss $\mathcal{L}_{CL\_tip}$.

$$\mathcal{L}_{CL\_tip} = -\sum_j^B log\frac{\exp(Sim(\mathcal{F}^j_{SV},\mathcal{F}^j_{sv\_clip})/\tau)}{\sum_k^B \exp(Sim(\mathcal{F}^j_{SV},\mathcal{F}^k_{sv\_clip})/\tau)}. \tag{5}$$

Experiments

Datasets and Metrics

KITTI dataset. The KITTI dataset (Geiger, Lenz, and Urtasun 2012) contains three categories (car, pedestrian, cyclist), provides 7481 front-view train images, 7518 test images, and LiDAR point clouds. There are three difficulty levels: Easy, Moderate (Mod.), and Hard. We compare RobusTor3D with other state-of-the-art methods for three categories on the KITTI dataset. The evaluation metric is the standard AP_{40} calculated from 40 recall positions for each class, model parameter(#params) and inference time.

KITTI-C dataset. The KITTI-C benchmark (Yu et al. 2023) is used to test the robustness of 3D detectors to corruptions present in real environments in out-of-distribution scenarios, including adverse weather conditions such as rain, snow, fog, and strong sunlight, external disturbances caused by motion blur, and cross-sensor scenarios such as spatial misalignment. We evaluate the 3D mAP performance of different models across all three classes on KITTI-C under the above challenging conditions with severity 1.

CADC dataset. The CADC dataset (Pitropov et al. 2021) serves as an object detection benchmark specifically designed to evaluate algorithms under the real snow scenarios. It has been converted to the KITTI format, comprising 6901 LiDAR scans and images, with 6315 for training and 586 for testing. We further divide the training set into a train split set (5052 samples) and a val split set (1263 samples) with a 4:1 ratio. We chose the 3D AP_{40} to evaluate models.

Method	Car(IoU=0.7)				Pedestrian (IoU=0.5)				Cyclist(IoU=0.5)				mAP_{All}
	Easy	Mod.	Hard	mAP	Easy	Mod.	Hard	mAP	Easy	Mod.	Hard	mAP	
MVXNet* (Sindagi, Zhou, and Tuzel 2019)	81.6	71.2	64.8	72.5	48.1	38.6	36.1	40.9	64.7	48.8	43.7	52.4	53.33
PointPainting (Vora et al. 2020)	82.1	71.7	67.0	73.6	50.3	40.9	37.8	43.0	77.6	63.7	55.8	65.7	60.83
MoCa (Zhang, Wang, and Loy 2020)	86.0	75.9	70.7	77.5	50.9	43.7	40.0	44.8	76.1	61.0	53.4	63.5	61.97
FocalsConv-F* (Chen et al. 2022)	90.4	81.4	77.1	83.0	50.4	42.5	40.2	44.4	75.4	61.6	57.1	64.7	64.06
CL3D (Lin et al. 2022b)	87.4	80.2	76.2	81.3	47.3	39.4	36.9	41.2	77.3	62.0	55.5	64.9	62.49
Graph-VoI* (Yang et al. 2022)	**91.5**	**82.4**	75.4	83.1	52.4	43.0	40.6	45.4	76.2	58.8	52.2	62.4	64.09
3D-DFM (Lin et al. 2022a)	87.7	80.9	76.1	81.6	46.9	39.6	37.6	41.4	79.6	63.3	56.6	66.5	63.19
EPNet++ (Liu et al. 2022)	91.3	81.9	76.7	83.3	**52.7**	44.3	41.2	46.1	76.1	59.7	53.6	63.1	64.21
PCDR-DFF (Xia et al. 2024)	81.3	73.2	68.9	74.4	49.6	41.4	38.1	43.1	74.2	58.9	52.6	61.9	59.84
PPF-Det (Xie et al. 2024)	90.4	81.3	77.0	82.9	47.1	41.9	40.0	43.0	**82.2**	**68.7**	**61.1**	**70.7**	65.57
PVAFN(25'ESWA) (Li et al. 2025)	88.1	81.5	76.9	82.1	-	-	-	-	-	-	-	-	-
RAFDet (Zheng et al. 2025)	88.2	79.8	75.0	81.0	48.9	41.8	38.6	43.1	82.2	65.3	58.7	68.7	64.33
SSLFusion* (Ding et al. 2025)	90.2	81.3	76.5	82.7	52.3	44.7	42.3	46.4	78.0	62.3	56.4	65.6	65.35
RobusTor3D (Ours)	90.7	82.1	**79.3**	**84.0**	52.0	**44.9**	**42.6**	**46.5**	78.7	64.3	58.6	67.2	**65.97**

Table 1: Comparison of our model with state-of-the-art models on KITTI test set for the three classes. 'Mod.' indicates the moderate difficulty level. The mAP_{All} and * denote average precision across all classes and rep-implement results, respectively. The best results are highlighted in bold. For fair comparison, all the reproduced and our results are trained only on the train set.

Implementation Details

We train our RobusTor3D on KITTI, KITTI-C and CADC datasets, utilizing voxelization (Yan, Mao, and Li 2018) for point clouds, the CLIP image encoder ResNet50 (He et al. 2016), and the text encoder(12-layer Transformer) (Radford et al. 2019) for images and text descriptions, respectively. We reproduce Voxel RCNN (Deng et al. 2021) with focal sparse convolution (Chen et al. 2022) as the baseline(Voxel RCNN+). Our model is trained from scratch on an RTX3090 in an end-to-end manner using the Adam optimizer (Kingma and Ba 2015) and a one-cycle policy with a learning rate of 0.01 for 80 epochs. The batch size is set to 4 per GPU.

Robustness Experiment Comparisons

We compared the robustness of different 3D detectors across five key aspects (more robustness experiments are in the **Supp. material**). Compared with other methods, our method has superior robustness in all challenging scenarios.

Long-tail problem (KITTI test set). To evaluate the robustness of our RobusTor3D to long-tail distribution, we divide the categories in KITTI into head(car) and tail(pedestrian, cyclist) classes and report the per-class 3D mAP performance. Table 1 compares RobusTor3D with state-of-the-art multimodal 3D detectors on the KITTI test set. From Table 1, RobusTor3D achieves the best mAP_{All}, outperforming 1.64%, 0.62%, and 0.4% of latest methods, RAFDet (Zheng et al. 2025), SSLFusion (Ding et al. 2025), and PPF-Det (Xie et al. 2024), respectively. RobusTor3D achieves competitive performance across all categories. For the Car and Pedestrian classes, it ranks No.1 among the best-performing methods. For the Cyclist class, RobusTor3D achieves competitive performance. Although it has a lower 3D mAP in the cyclist class than PPF-Det and RAFDet, they have much lower performance on the pedestrian class than RobusTor3D. The pedestrian category has extremely few objects and the smallest size. RobusTor3D can achieve relatively balanced performance on rare categories. This highlights that our robustness-enhanced design of text descriptions and V-LKE enhances the recognition of rare classes.

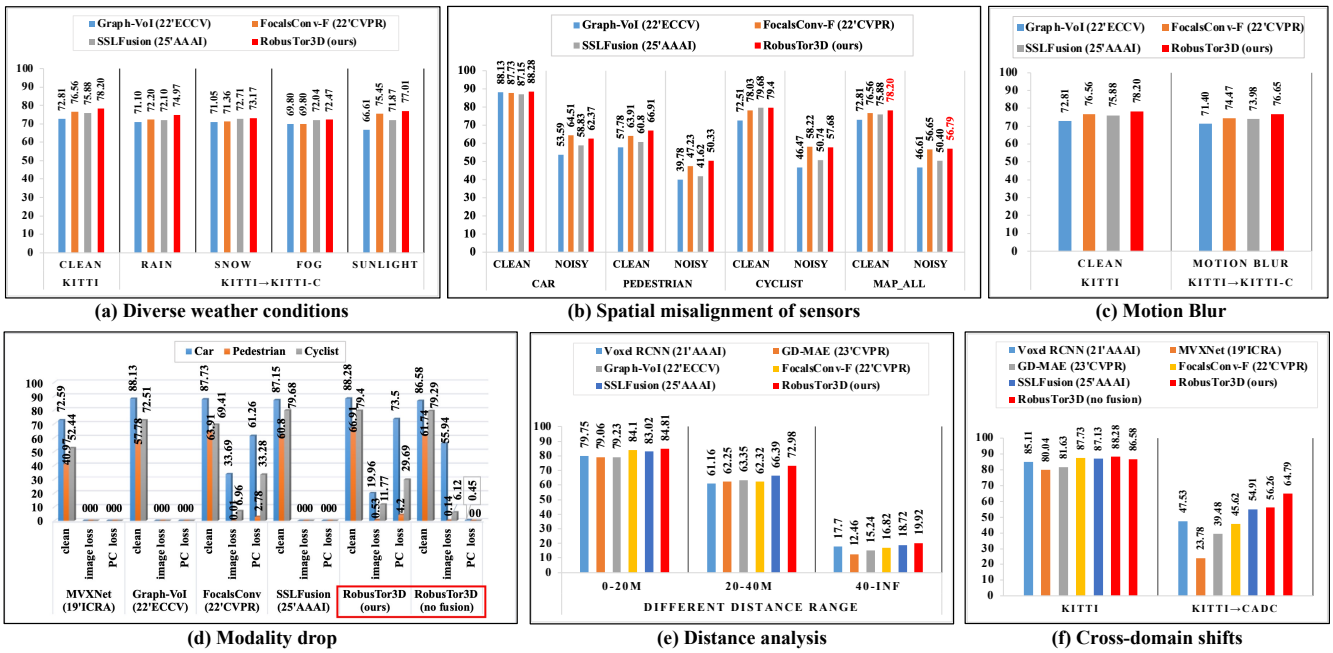
Method	Rain	Snow	Fog	SSun	Noise
Graph-VoI	72.66	72.93	71.54	71.79	47.21
FocalsConv	72.56	73.81	73.82	75.00	56.18
SSLFusion	72.83	71.60	72.34	76.68	48.68
RobusTor3D	**75.71**	**76.02**	**76.73**	**77.02**	**59.14**

Table 2: Comparison results of 3D mAP of all classes with other methods on KITTI-C val set under adverse weather conditions and sensor spatial misalignment. Noise and SSun represent sensor spatial alignment noise and strong sunlight.

Adverse weather conditions (KITTI-C val set). To validate the robustness of our RobusTor3D to the corruptions in the real environment in the out-of-distribution scene, we conduct extensive comparative experiments on the KITTI-C benchmark (various adverse weather conditions: rain, snow, fog, strong sunlight). From Table 2, RobusTor3D achieves the best 3D mAP performance, over 2.88%-3.15%, 2.21%-4.42%, and 2.91%-5.19%, than other methods in all challenging environments, proving that our RobusTor3D has stronger adaptability to various adverse weather conditions.

To further verify the generalization capability under adverse weather conditions of RobusTor3D, we also test multi-modal models trained on the clean KITTI dataset directly on the KITTI-C dataset without any fine-tuning. From Figure 4(a), RobusTor3D consistently outperforms other methods across all weather conditions, especially under strong sunlight, surpassing other methods by 1.56%-10.4%. These gains highlight the necessity of vision-language priors in maintaining semantic stability under adverse conditions.

Sensor-level problems (KITTI-C val set). To verify the noise resistance under sensor-level corruptions of RobusTor3D, we train different detectors on KITTI-C with sensor alignment noise. From Table 2, our RobusTor3D out-



(a) Diverse weather conditions (b) Spatial misalignment of sensors (c) Motion Blur

(d) Modality drop (e) Distance analysis (f) Cross-domain shifts

Figure 4: Comprehensive robustness experiment comparisons of different 3D detectors against real-world factors from the remaining four aspects, including (a) adverse weather conditions, (b) (c) sensor corruptions: spatial misalignment of sensors and motion blur, (d)(e) modality loss: modality drop and distance analysis, and (f) cross-domain shifts.

performs other methods by 2.96%-11.93%, proving that it is resistant to spatial misalignment of sensors. We also test multi-modal models trained on the clean KITTI dataset directly on the KITTI-C dataset under spatial misalignment of sensors and motion blur. From Figure 4(b), in both clean and noisy spatial alignment settings, RobusTor3D maintains superior performance, achieving the best 3D mAP, which outperforms 6.39% and 9.98% of SSLFusion and Graph-VoI. The robust performance under sensor noise suggests that the scene-level contrastive learning (SMCL) helps maintain consistency between modalities even when spatial correspondences are degraded. From Figure 4(c), RobusTor3D outperforms all other methods, with a minimal performance drop from clean to motion-blur conditions (only -0.74%), demonstrating the model's resilience against motion blur.

Modality loss (KITTI val set). To evaluate the modality robustness and spatial generalization of RobusTor3D, we conduct two controlled experiments on the KITTI val set. (a) Modality drop: we simulate image-drop and point cloud(PC)-drop scenarios by not fusing features of the corresponding branches during inference. From Figure 4(d), when the image or PC modality is dropped, most methods completely fail for all classes. Although FocalsConv-F has a high performance in the car class, it fails to perceive the pedestrian class. RobusTor3D still has perception ability for all categories in both cases. This shows that RobusTor3D has a strong ability to compensate for modalities and proves the actual robustness when the sensor is unavailable. In addition, we also conduct another experiment (without fusing image and text features during training) and find that it has a strong perception ability for cars when the image modality is dropped, far surpassing other methods. Although the image features are not fused, it can still have a certain perception ability for the car category when the PC modality is dropped, further proving the superiority of our SMCL strategy for modality loss. (b) Distance analysis: we divide the val set into three spatial ranges (0–20m, 20–40m, >40m) based on the ground-truth object distance, and report the performance in each range. From Figure 4(e), RobusTor3D shows stable and superior performance across all distance ranges. This illustrates RobusTor3D's balanced dependence on both near-field and far-field features, aided by SMCL and rich semantics of V-LKE.

Cross-domain shifts (KITTI→CADC, CADC). To validate cross-domain generalization under severe domain shifts of RobusTor3D, we train LiDAR-based and multi-modal models on the KITTI dataset and directly test them on the CADC dataset without any fine-tuning. From Figure 4(f), in the KITTI→CADC domain shift evaluation, RobusTor3D outperforms all other methods by 1.35%-32.48% in 3D mAP across three difficulty levels of the car. This demonstrates its domain generalization capacity, enhanced by VLM priors, enabling the model to better handle unseen distributions from CADC. Similarly, our other method's cross-domain performance surpasses all other methods.

To further evaluate generalization and weather robustness, we conduct experiments on the CADC dataset. We train some state-of-the-art 3D detectors, including Voxel RCNN (Deng et al. 2021), PV-RCNN (Shi et al. 2020), GD-MAE (Yang et al. 2023), SSLFusion, Graph-VoI, Focalsconv, and MVXNet, using the same dataset and code provided by MMDetection3D (Contributors 2020) and OpenPCDet

Method	Modal.	Easy	Mod.	Hard	mAP
Voxel RCNN	L	54.42	51.31	52.01	52.58
PV-RCNN		54.34	46.89	48.43	49.89
GD-MAE		40.30	32.79	33.75	35.61
MVXNet	Multi.	56.22	53.95	56.18	55.45
Graph-VoI		40.47	33.96	35.14	36.52
FocalsConv		60.15	57.34	**59.61**	59.03
SSLFusion		63.17	55.88	54.25	57.77
RobusTor3D		**63.99**	**59.27**	58.95	**60.74**

Table 3: Comparison results of the Car class with other methods on the CADC dataset. L and Multi. represent single LiDAR modality and multi-modal, respectively.

OTSG	V-LKE	ITCCL	IVCCL	Car	Ped.	Cyc.
✗	✗	✗	✗	86.78	60.34	78.47
✗	V	✗	✗	87.63	63.49	79.43
✓	✓	✗	✗	87.76	65.74	79.41
✓	✓	✓	✗	87.81	66.48	79.61
✓	✓	✓	✓	88.28	66.91	79.40
Improvements				+1.50	+6.57	+0.93

Table 4: Effect of each component of our RobusTor3D on KITTI val set. V indicates the use of image and I.E.

(Team 2020). The results for the car class are in Figure 4. It can be observed that RobusTor3D outperforms other SOTA methods in terms of 3D mAP of all difficulty levels, with a superiority of 24.22%, 1.71%, and 2.97% over Graph-VoI, FocalsConv-F, and SSLFusion, respectively.

Ablation Study

In this section, we perform an ablation study to analyze the impact of key components of our RobusTor3D, including the Offline Text Supervision Generation (OTSG), Vision-Language Knowledge Extraction (V-LKE), and Scene-level Multimodal Consistent Learning (SMCL) with ITCCL and IVCCL, on the KITTI val set for the car, pedestrian, and cyclist classes. The Voxel RCNN with focal sparse convolution is chosen as the baseline (Voxel RCNN+). The results are given in Table 4. It can be observed that compared with the baseline method, our RobusTor3D obtains an absolute gain of 1.50%, 6.57%, and 0.93% of average 3D mAP over three difficulty levels of three classes, respectively. The #params and inference time of baseline and ours are (7.8M, 0.1s) and (9.9M, 0.3s), not introducing too much complexity.

Effect of Offline Text Supervision Generation. From Table 4, by comparing the second and third rows, it can be observed that OTSG improves the 3D mAP on the easy, moderate, and hard difficulty levels by 0.13% and 2.25% of the car and pedestrian classes, respectively. The performance gain highlights the importance of introducing text modality, especially for long-tail or ambiguous instances like pedestrians, which enables the model to better exploit high-level semantics encoded in the vision-language space.

Effect of Vision-Language Knowledge Extraction. From Table 4, by comparing the 2nd and 1st rows, it is evident that the image input and CLIP vision encoder lead to performance gains of 0.85%, 3.25%, and 0.96% in terms of 3D mAP over the easy, moderate, and hard difficulty levels of the three classes, respectively. Similarly, comparing row 3 with row 1, using the full vision and language encoders consistently improves detection results notably. The mAP of cars, pedestrians, and cyclists increased by 0.98%, 4.40%, and 0.94%, which shows that V-LKE can effectively enrich the representation of specific modalities. This improvement is particularly evident in scenes with sparse semantics, confirming the importance of prior knowledge blending.

Effect of Scene-level Multimodal Consistent Learning. SMCL contains ITCCL and IVCCL. From Table 4, when ITCCL and IVCCL are added gradually (from 3rd row to 4th row and to 5th row), we observe that except for the cyclist category in the last row, which is slightly lower than the 4th row but almost unchanged compared to the third row (the reason for the analysis may be that the two losses affect each other during joint alignment), all other indicators are steadily improved. In particular, IVCCL further improves the pedestrian 3D mAP (+1.23%) based on ITCCL, reflecting its effectiveness in enhancing the scene-level alignment between image-text and LiDAR modalities. The overall improvement confirms that scene-level contrast objectives are crucial for learning stable and complementary multimodal features.

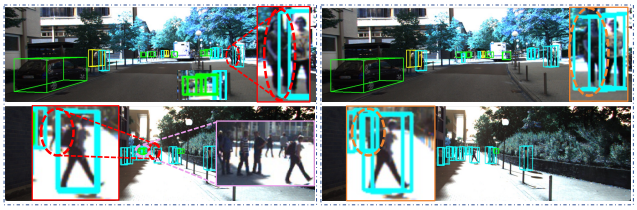


Figure 5: Comparison of visualization results on KITTI test set.

Visualization

We also provide visual comparisons (more visualizations are in the **Supp. material**) in Figure 6, where green, light blue, and yellow 3D boxes represent Cars, Pedestrians, and Cyclists, respectively. The missed, false, and correct detections are labeled with red, purple, and orange dashed ellipses, respectively. It shows that SSLFusion* has some missed detections of pedestrians (tail class) and false detections of cars. Our RobusTor3D can accurately detect pedestrians that are far away and severely obscured, further demonstrating the robustness of RobusTor3D.

Conclusion

In this paper, we propose RobusTor3D, a robust multimodal 3D detector designed for real-world autonomous driving scenarios by blending semantic priors from vision-language models at both the structural and supervisory levels. Extensive experiments on the KITTI, KITTI-C, and CADC benchmarks demonstrate that RobusTor3D achieves consistent and superior robustness across a diverse range of real-world perturbations, outperforming state-of-the-art methods.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (U2468223, 51827813), and in part by the China Scholarship Council under Grant No.202407090137.

References

Chang, G.; Roh, W.; Jang, S.; Lee, D.; Ji, D.; Oh, G.; Park, J.; Kim, J.; and Kim, S. 2024. Cmda: Cross-modal and domain adversarial adaptation for lidar-based 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 972–980.

Chen, L.; Wu, P.; Chitta, K.; Jaeger, B.; Geiger, A.; and Li, H. 2024. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, Y.; Li, Y.; Zhang, X.; Sun, J.; and Jia, J. 2022. Focal sparse convolutional networks for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5428–5437.

Contributors, M. 2020. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d.

Deng, J.; Shi, S.; Li, P.; Zhou, W.; Zhang, Y.; and Li, H. 2021. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1201–1209.

Ding, B.; Xie, J.; Nie, J.; and Cao, J. 2025. SSLFusion: Scale and Space Aligned Latent Fusion Model for Multimodal 3D Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2735–2743.

Dong, Y.; Kang, C.; Zhang, J.; Zhu, Z.; Wang, Y.; Yang, X.; Su, H.; Wei, X.; and Zhu, J. 2023. Benchmarking robustness of 3d object detection to common corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1022–1032.

Gao, H.; Wang, Z.; Li, Y.; Long, K.; Yang, M.; and Shen, Y. 2024. A survey for foundation models in autonomous driving. *arXiv preprint arXiv:2402.01105*.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

Greer, R.; Antoniussen, B.; Møgelmose, A.; and Trivedi, M. 2025. Language-driven active learning for diverse open-set 3d object detection. In *Proceedings of the Winter Conference on Applications of Computer Vision*, 980–988.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hegde, D.; Lohit, S.; Peng, K.-C.; Jones, M.; and Patel, V. 2025a. Multimodal 3D Object Detection on Unseen Domains. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2499–2509.

Hegde, D.; Yasarla, R.; Cai, H.; Han, S.; Bhattacharyya, A.; Mahajan, S.; Liu, L.; Garrepalli, R.; Patel, V. M.; and Porikli, F. 2025b. Distilling multi-modal large language models for autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 27575–27585.

Huang, C.; Jiang, A.; Feng, J.; Zhang, Y.; Wang, X.; and Wang, Y. 2024. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11375–11385.

Jiao, P.; Zhao, N.; Chen, J.; and Jiang, Y.-G. 2024. Unlocking textual and visual wisdom: Open-vocabulary 3d object detection enhanced by comprehensive guidance from text and image. In *European Conference on Computer Vision*, 376–392. Springer.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Li, Y.; Wen, J.; Gong, R.; Ren, B.; Li, W.; Cheng, C.; Liu, H.; and Sebe, N. 2025. Pvafn: Point-voxel attention fusion network with multi-pooling enhancing for 3d object detection. *Expert Systems with Applications*, 281: 127608.

Lin, C.; Tian, D.; Duan, X.; Zhou, J.; Zhao, D.; and Cao, D. 2022a. 3D-DFM: Anchor-free multimodal 3-D object detection with dynamic fusion module for autonomous driving. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12): 10812–10822.

Lin, C.; Tian, D.; Duan, X.; Zhou, J.; Zhao, D.; and Cao, D. 2022b. CL3D: Camera-LiDAR 3D object detection with point feature enhancement and point-guided fusion. *IEEE Transactions on Intelligent Transportation Systems*, 23(10): 18040–18050.

Liu, Z.; Huang, T.; Li, B.; Chen, X.; Wang, X.; and Bai, X. 2022. EPNet++: Cascade bi-directional fusion for multi-modal 3D object detection. *IEEE transactions on pattern analysis and machine intelligence*, 45(7): 8324–8341.

Pan, C.; Yaman, B.; Velipasalar, S.; and Ren, L. 2024. Clip-bevformer: Enhancing multi-view image-based bev detector with ground truth flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15216–15225.

Peng, Y.; Qin, Y.; Tang, X.; Zhang, Z.; and Deng, L. 2022. Survey on image and point-cloud fusion-based object detection in autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23(12): 22772–22789.

Pitropov, M.; Garcia, D. E.; Rebello, J.; Smart, M.; Wang, C.; Czarnecki, K.; and Waslander, S. 2021. Canadian adverse driving conditions dataset. *The International Journal of Robotics Research*, 40(4-5): 681–690.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Sapkota, R.; Roumeliotis, K. I.; Cheppally, R. H.; Calero, M. F.; and Karkee, M. 2025. A review of 3d object detection with vision-language models. *arXiv preprint arXiv:2504.18738*.

Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; and Li, H. 2020. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10529–10538.

Sindagi, V. A.; Zhou, Y.; and Tuzel, O. 2019. Mvx-net: Multimodal voxelnet for 3d object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, 7276–7282. IEEE.

Song, Z.; Liu, L.; Jia, F.; Luo, Y.; Jia, C.; Zhang, G.; Yang, L.; and Wang, L. 2024. Robustness-aware 3d object detection in autonomous driving: A review and outlook. *IEEE Transactions on Intelligent Transportation Systems*.

Song, Z.; Wei, H.; Bai, L.; Yang, L.; and Jia, C. 2023. Graphalign: Enhancing accurate feature alignment by graph matching for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3358–3369.

Team, O. D. 2020. OpenPCDet: An Open-source Toolbox for 3D Object Detection from Point Clouds. https://github.com/open-mmlab/OpenPCDet.

Vora, S.; Lang, A. H.; Helou, B.; and Beijbom, O. 2020. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4604–4612.

Wu, Y.; Xing, M.; Zhang, Y.; Xie, Y.; and Qu, Y. 2024. Clip2uda: Making frozen clip reward unsupervised domain adaptation in 3d semantic segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8662–8671.

Xia, C.; Li, X.; Gao, X.; Ge, B.; Li, K.-C.; Fang, X.; Zhang, Y.; and Yang, K. 2024. PCDR-DFF: multi-modal 3D object detection based on point cloud diversity representation and dual feature fusion. *Neural Computing and Applications*, 36(16): 9329–9346.

Xie, G.; Chen, Z.; Gao, M.; Hu, M.; and Qin, X. 2024. PPF-Det: point-pixel fusion for multi-modal 3D object detection. *IEEE Transactions on Intelligent Transportation Systems*, 25(6): 5598–5611.

Yan, Y.; Mao, Y.; and Li, B. 2018. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10): 3337.

Yang, H.; He, T.; Liu, J.; Chen, H.; Wu, B.; Lin, B.; He, X.; and Ouyang, W. 2023. Gd-mae: generative decoder for mae pre-training on lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9403–9414.

Yang, H.; Liu, Z.; Wu, X.; Wang, W.; Qian, W.; He, X.; and Cai, D. 2022. Graph r-cnn: Towards accurate 3d object detection with semantic-decorated local graph. In *European conference on computer vision*, 662–679. Springer.

Yoon, J. H.; Jung, J. W.; and Yoo, S. B. 2025. Equirectangular Point Reconstruction for Domain Adaptive Multimodal 3D Object Detection in Adverse Weather Conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9553–9561.

Yu, K.; Tao, T.; Xie, H.; Lin, Z.; Liang, T.; Wang, B.; Chen, P.; Hao, D.; Wang, Y.; and Liang, X. 2023. Benchmarking the robustness of lidar-camera fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3188–3198.

Zhang, D.; Li, C.; Zhang, R.; Xie, S.; Xue, W.; Xie, X.; and Zhang, S. 2024a. Fm-ov3d: Foundation model-based cross-modal knowledge blending for open-vocabulary 3d detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16723–16731.

Zhang, W.; Wang, Z.; and Loy, C. C. 2020. Exploring data augmentation for multi-modality 3d object detection. *arXiv preprint arXiv:2012.12741*.

Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; and Feng, J. 2023. Deep long-tailed learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9): 10795–10816.

Zhang, Z.; Meyer, G. P.; Lu, Z.; Shrivastava, A.; Ravichandran, A.; and Wolff, E. M. 2024b. Vlm-kd: Knowledge distillation from vlm for long-tail visual recognition. *arXiv preprint arXiv:2408.16930*.

Zheng, Z.; Huang, Z.; Zhao, J.; Lin, K.; Hu, H.; and Chen, D. 2025. RAFDet: Range view augmented fusion network for point-based 3D object detection. *IEEE Transactions on Multimedia*.