LiDAR-based All-weather 3D Object Detection via Prompting and Distilling 4D Radar

Yujeong Chae, Hyeonseong Kim, Changgyoon Oh, Minseok Kim, and Kuk-Jin Yoon

Visual Intelligence Lab., KAIST
{yujeong, brian617, changgyoon, alstjrx1x1, kjyoon}@kaist.ac.kr

Abstract. LiDAR-based 3D object detection models show remarkable performance, however their effectiveness diminishes in adverse weather. On the other hand, 4D radar exhibits strengths in adverse weather but faces limitations in standalone use. While fusing LiDAR and 4D radar seems to be the most intuitive approach, this method comes with limitations, including increased computational load due to radar pre-processing, situational constraints when both domain information is present, and the potential loss of sensor advantages through joint optimization. In this paper, we propose a novel LiDAR-only-based 3D object detection framework that works robustly in all-weather (normal and adverse) conditions. Specifically, we first propose 4D radar-based 3D prompt learning to inject auxiliary radar information into a LiDAR-based pre-trained 3D detection model while preserving the precise geometry capabilities of LiDAR. Subsequently, using the preceding model as a teacher, we distill weather-insensitive features and responses into a LiDAR-only student model through our four levels of inter-/intra-modal knowledge distillation. Extensive experiments demonstrate that our prompt learning effectively integrates the strengths of LiDAR and 4D radar, and our LiDAR-only student model even surpasses the detection performance of teacher and state-of-the-art models under various weather conditions.

Keywords: LiDAR · 3D object detection · Normal/adverse · 4D radar · Knowledge distillation · 3D prompt learning · Autonomous driving

1 Introduction

3D object detection is one of the fundamental tasks in autonomous driving, which aims to identify the targets and localize them in 3D coordinates. Therefore, numerous studies have been conducted in the field of 3D object detection, leveraging various sensors (*e.g.* camera, LiDAR, radar) [1,2,8,9,11,13,21,36,41]. Among them, studies utilizing LiDAR have demonstrated remarkable performance compared to other sensors, thanks to its precise mapping capabilities [21, 35, 53]. However, these methods suffer performance degradation under adverse weather conditions due to the noisy measurements and reduced detection range of the

* Code: https://github.com/yujeong-star/LOD_PDR.

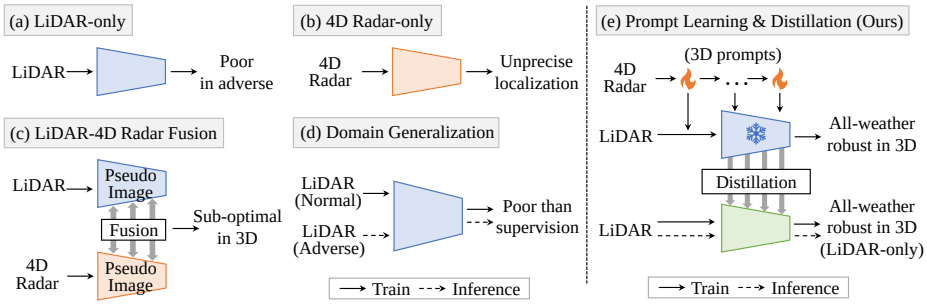

Fig. 1: Prior LiDAR- or 4D radar-based research follows the detection framework in (a)-(d). Single-modal methods often exhibit poor detection performance in normal or adverse weather conditions due to limitations inherent to each sensor. On the other hand, multi-modal approaches struggle to effectively fuse the strengths of 3D sensors, resulting in sub-optimal performance. In contrast, our method (e) proposes a effective multi-modal integration in 3D space via 4D radar-based prompts and distill the multi-modal weather-insensitive knowledge to LiDAR-only student model.

LiDAR [5,6,19,28]. While several approaches have attempted to address adverse weather conditions through domain generalization, their results still fall short compared to the performance of supervised LiDAR-based models across various weather conditions [7, 12, 34]. Conversely, radar sensors, utilizing radio waves, offer strengths in weather-insensitive and wide-range measurements [22,25]. The advent of a novel sensor, 4D radar, providing additional elevation information, further enhances the capabilities of radar technology for 3D object detection. Despite their advantages, radars face limitations in standalone use due to their inability to provide accurate depth information [11,25,26]. Indeed, directly integrating LiDAR and 4D radar information at the input or feature level appears to be an intuitive approach for designing a 3D object detection model under diverse weather conditions. However, this method comes with the following limitations.

Firstly, using radar alongside LiDAR requires data pre-processing due to the different formats, leading to increased computational load and time-consuming procedures [25,26]. To utilize information from radar sensors, multiple steps such as Fast Fourier Transform (from multiple FMCW signals to radar tensor), coordinate transform (from polar to Cartesian) and CFAR algorithm (from radar tensor to point cloud) are necessary to handle radar data effectively. Moreover, if the model processes LiDAR and radar through separate encoders, the model size and number of parameters may also increase. The second reason, while evident, is that a model designed to take both LiDAR and radar as input will only operate when both LiDAR and radar data are available simultaneously. This diminishes the generality of the developed algorithm. The final reason is that joint optimization can lead to sub-optimal results when simultaneously receiving and fusing data from both domains. While there can be performance improvements by leveraging the advantages of each sensor, the unprecise localization of

radar or the noise in LiDAR introduced by adverse conditions may hinder the algorithm from achieving optimal performance (see Table 2). These limitations make real-world applications a bit challenging.

In this paper, we propose a novel LiDAR-only-based 3D object detection framework that works robustly in various weather conditions, as shown in Fig. 1. Specifically, we first design a teacher model that takes LiDAR and auxiliary 4D radar as input. The 4D radar information is incorporated into a pre-trained LiDAR-based 3D object detection model through our 4D radar-based 3D prompt learning. Given the trained LiDAR features and untrained 4D radar data, we perform LiDAR self-calibration, globally aggregate it with radar information and create a 4D radar prompt. Then, by locally aggregating the neighboring prompts around the LiDAR queries, LiDAR features are efficiently updated in a weather-insensitive manner using minimal parameters while preserving their original precise geometric mapping capabilities. After that, we distill weather-insensitive features and responses into a LiDAR-only student model through our four levels of inter-/intra-modal knowledge distillation (KD). For intra-modal KD, we align the radar-aided LiDAR sparse voxel, dense bird-eye-view (BEV) features and responses from the teacher to closely resemble those of the student model. We further distill the knowledge from the 4D radar prompt to enhance the LiDAR features of the student, facilitating structural 3D scene understanding across inter-modalities. We evaluate the detection performance of the proposed teacher and student model on the K-Radar dataset [25] that provides LiDAR and 4D radar under various weather conditions. The experiments demonstrate that our prompt learning method efficiently integrates the strengths of 4D radar into LiDAR without compromising the inherent strengths of LiDAR. Furthermore, the results illustrate that our LiDAR-only student model successfully learns weather-insensitive knowledge from the teacher model, surpassing the performance of both the teacher and state-of-the-art models across various weather conditions.

Our main contributions are four-fold: (I) We present an innovative LiDAR-only-based 3D object detection framework designed for robust performance in various weather conditions by effectively utilizing of auxiliary 4D radar information during training. (II) We propose 4D radar prompt learning, which integrates the strengths of radar and LiDAR with minimal loss of information in 3D. (III) We propose four levels of inter-/intra-modal distillation method between 4D radar-LiDAR-based teacher and LiDAR-based student. (IV) We conduct extensive experiments on the K-Radar dataset, showing superior performance of the LiDAR-only student model across various weather conditions.

2 Related Works

2.1 3D Object Detection with LiDAR

LiDAR Only. LiDAR-based 3D object detection demonstrates remarkable performance owing to the precise mapping characteristics of LiDAR sensors. Existing research mainly processes LiDAR point clouds with point-based [23, 32, 46],

voxel-based [2,42,47,54,55] models or employs multiple representations [30,31] to derive the final 3D bounding boxes. Point-based models and voxel-based models extract 3D point or voxel features with point networks, graph networks, 3D sparse convolution networks and Transformers. Several works utilize pillar [13,15] or BEV representation [39,44] for efficient computation and execution. Among them, voxel-based methods are actively researched based on their favorable performance-speed balance. Most existing research has primarily focused on enhancing performance in clean weather conditions.

Adverse Weather. In adverse weather conditions, incorrect laser reflections occur in LiDAR around rain or snow particles. Consequently, a decrease in points captured on objects, an increase in noisy points, and a reduction in detection range contribute to the degradation of detection performance. To minimize the degradation of the LiDAR-only model in adverse weather, several research [19,27] focus on filtering noise in the LiDAR input or feature with supervision. Another line of research focuses on data augmentation and training strategies to enable the model trained in normal conditions to work robustly in adverse conditions. [7, 38] propose domain adaptation and semi-supervised framework, [34] designs normal to adverse distillation loss, and [12] recently proposes simulated corruption dataset along with domain generalization technique. Since LiDAR is vulnerable to adverse weather conditions, they do not yet exhibit robust performance in such challenging scenarios.

Fusion with Radar. Radar has the advantage of being robust in adverse weather conditions and capable of detecting objects at long distances, as it utilizes radio waves for object sensing. However, it lacks the capability to offer precise distance measurements or detailed 3D maps and struggles with standalone deployment [11, 25, 26]. Therefore, several research fuses LiDAR and radar for robust 3D object detection in various weather conditions with feature concatenation [24, 43] and self-/cross-attention [18, 28]. These works process multi-domain features with identical network encoders, even though the sensor characteristics are distinctly different. With 4D radar that provides an additional elevation dimension, [37] fuse LiDAR and 4D radar pillar features with the proposed interaction module. However, the use of pillar features led to some loss of crucial height information in 3D object detection. Unlike prior works, we propose a novel LiDAR-based 3D object detection framework aided with auxiliary 4D radar information, considering each sensor's characteristics under various weather conditions during fusion.

2.2 Multi-modal Prompt Learning

The prompt is widely used to guide the language models in generating specific outputs [29,33]. In particular, various studies have explored prompt tuning, employing soft prompts with frozen pre-trained models, and demonstrated the effectiveness over traditional finetuning methods [14, 16]. Likewise, vision researchers have employed prompts for vision-language models [51, 52]. Later, [4, 10] demonstrated high performance in classification tasks by only utilizing visual prompts.

Recently, [56] proposed a multi-modal visual prompt to extend the use of RGB-based pre-trained tracking models to multi-modal tracking, demonstrating its effective application across auxiliary domains beyond RGB without the need for an additional network branch. While prompts have been extensively explored in 2D and language applications, there is a lack of research on utilizing prompts for multi-modal information transfer in the 3D domain.

2.3 Knowledge Distillation for LiDAR-based 3D Object Detection

Research on knowledge distillation (KD) within the LiDAR modality can be broadly categorized into two main approaches. One line of research focuses on model compression that transfers the learned knowledge from a larger 3D object detection model (teacher) to a smaller model (student) [3,17,45,48]. Another line of research investigates how to effectively train a student model of the same size, utilizing meta-information obtained from the teacher model [40, 50]. Recently, there has been a significant increase in research on cross-modal KD for 3D object detection using LiDAR. [49,53] distill the rich semantic information of multi-view image feature to LiDAR-only detection model with crucial response mining. Unlike prior works that only transfer knowledge between LiDAR and images, we propose a novel cross-modal KD method between the previously unexplored 3D domains of LiDAR and 4D radar, enabling robust performance in adverse weather conditions using LiDAR-only information.

3 Methods

3.1 Framework Overview

The overall scheme of our framework is illustrated in Fig. 2. In the framework, we have the following components: the baseline model, the teacher model, and the student model. They will be explained in the following sequential order.

Baseline. The baseline LiDAR-based 3D object detection model is inspired by [25], which has a voxel embedding layer, three voxel and BEV encoders, and a detection head. Specifically, the LiDAR point cloud $L \in \mathbb{R}^{N_0 \times 3}$ is first mapped into a higher dimension of voxel features $L_0 \in \mathbb{R}^{N_0 \times C_0}$ by voxel embedding layer. Then the encoder layer E_l extracts $L_l \in \mathbb{R}^{N_l \times C_l}$ from L_{l-1}, where l is the layer index and $l \in \{1, 2, 3\}$. While L_l serves as the input of the next encoder layer, it is compressed through the BEV encoder and extracts l-th BEV feature $B_l \in \mathbb{R}^{H_d \times W_d \times D}$. Finally, the concatenation of all layers' BEV features $B_{total} \in \mathbb{R}^{H_d \times W_d \times 3D}$ passes the detection head, which is composed of the classification head and regression head. Each head outputs classification map $A_{cls} \in \mathbb{R}^{H_d \times W_d \times I}$ and regression map $A_{reg} \in \mathbb{R}^{H_d \times W_d \times J}$, respectively.

Teacher. The teacher model that takes LiDAR and 4D radar consists of a baseline LiDAR detection model and an additional 4D radar prompt module P_l. To indicate the features and layers belonging to the teacher model in the following context, we will denote them as T superscript. The 4D radar point

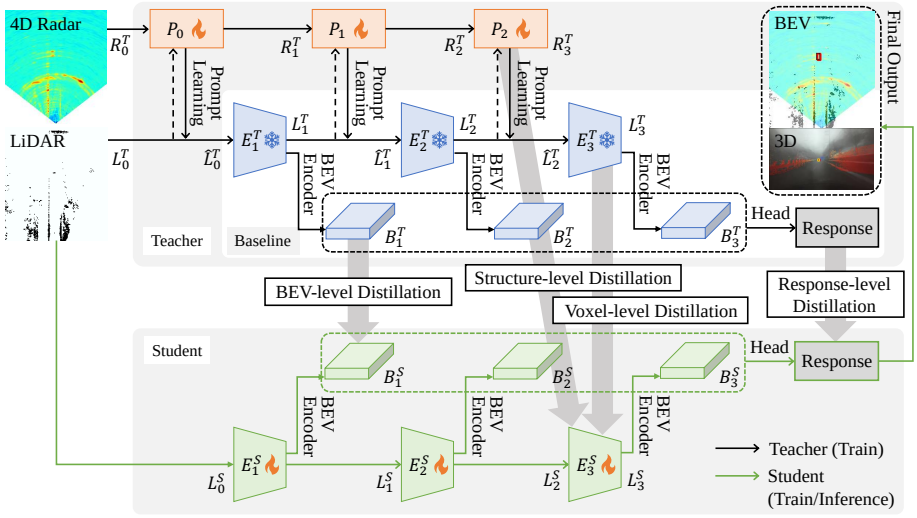

Fig. 2: Overall scheme of our LiDAR-based all-weather 3D object detection framework. Our framework consists of multi-modal teacher model and LiDAR-based student model. The teacher model efficiently fuses 4D radar information into LiDAR baseline using the proposed prompt learning, ensuring that both sensors' advantages are preserved. The student receives the rich and weather-insensitive knowledge from the teacher through the novel distillation methods, making it robust across various weather conditions.

cloud $R \in \mathbb{R}^{M_0 \times 3}$ is first mapped into radar voxel feature $R_0^T \in \mathbb{R}^{M_0 \times C_0}$. Then, our 4D radar-based 3D prompt learning module P_{l-1} updates L_{l-1}^T into enhanced LiDAR feature $\hat{L}_{l-1}^T$ by utilizing the radar feature R_{l-1}^T (see Sec. 3.2).

Student. The student model only takes LiDAR, and its structure is identical to the baseline model. Student's encoder E_l^S extracts L_l^S, student's BEV encoder extracts BEV feature B_l^S and head extracts the responses A_{cls}^S and A_{reg}^S, respectively. Our inter-/intra-modal KD losses align the features and responses of the student with those of the teacher (see Sec. 3.3). The final 3D detection results of the student model are robust regardless of the weather conditions.

3.2 4D Radar-based 3D Prompt Learning

In this section, we explain our 4D radar-based 3D prompt learning procedure in detail. LiDAR provides precise geometry information but has the disadvantage of being vulnerable to adverse weather conditions. In contrast, 4D radar is robust in adverse weather but has the disadvantage of being inaccurate due to imprecise geometry information, making it difficult to use alone. Therefore, we aim to train a robust teacher model under all weather conditions using LiDAR and 4D radar as an auxiliary. In particular, among the various ways of using 4D radar as an auxiliary to LiDAR, *e.g.* fusing 4D radar with LiDAR and jointly optimizing the model from scratch or fine-tuning, we propose prompting 4D radar information

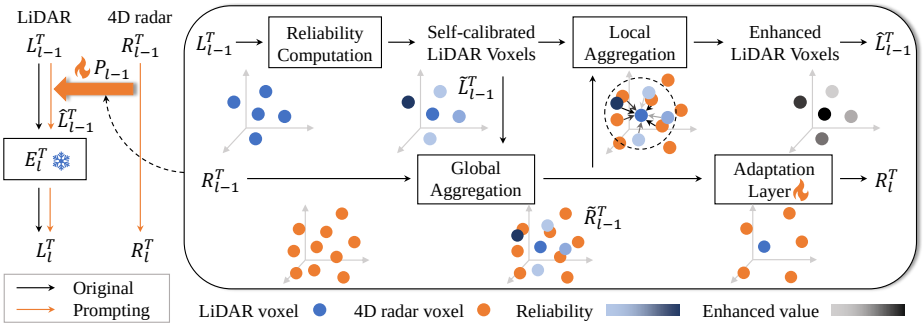

Fig. 3: Illustration of our 4D radar-based 3D prompt learning. The pre-trained LiDAR first undergoes self-calibration and is then globally aggregated with 4D radar to obtain a rich and robust 3D prompt. Subsequently, this self-calibrated LiDAR and 4D radar prompts go through local aggregation mechanism to enhance LiDAR feature.

on a pre-trained LiDAR-based model. As can be seen in Table 2, training a model from scratch (c, d, g) or fine-tuning (h) by fusing a 4D radar can degrade the LiDAR's precise localization ability due to the radar's imprecise geometry information. To alleviate this issue, we create learnable 3D prompts conditioned on the 4D radar with lightweight adaptation layers to make the LiDAR-only pre-trained model robust in various weather conditions without sacrificing precise geometric information. Unlike [56], which introduced multi-modal prompting in a pixel-wise, well-aligned 2D domain for multi-modal visual tracking, we propose a novel multi-modal 3D prompt learning. Our method effectively handles sparse LiDAR and 4D radar data that are not aligned in 3D space, considering the distinct characteristics of each sensor.

Specifically, as shown in Fig. 3, the proposed multi-modal 3D prompt learning module P_{l-1} takes processed LiDAR voxel feature L^T_{l-1} and 4D radar voxel feature R^T_{l-1} as input and output enhanced LiDAR voxel feature $\hat{L}^T_{l-1}$ and 4D radar voxel feature R^T_l. To further amplify the important information of LiDAR features from the pre-trained LiDAR-based model, we self-calibrate the LiDAR feature based on its reliability as follows:

$$\widetilde{L}^T_{l-1} = \phi(L^T_{l-1}) \cdot L^T_{l-1}, \quad \phi(L^T_{l-1}) = \sigma(\sum_{c=1}^{C_{l-1}} abs(\{L^T_{l-1}\}_c))\}, \tag{1}$$

where $\widetilde{L}^T_{l-1} \in \mathbb{R}^{N_{l-1} \times C_{l-1}}$ is self-calibrated LiDAR feature, $\phi(L^T_{l-1}) \in \mathbb{R}^{N_{l-1}}$ is reliability score, c denotes c-th channel feature, abs is absolute function and σ is a sigmoid function. Then, the self-calibrated LiDAR feature is globally aggregated into the 4D radar feature to fill in the missing geometrical information. To this end, we insert the LiDAR features $\widetilde{L}^T_{l-1}$ into the radar feature R^T_{l-1} in 3D space where only the LiDAR feature exists. We denote this globally aggregated feature as $\widetilde{R}^T_{l-1}$, which serves as a learnable prompt to enhance the LiDAR feature.

$$\widetilde{R}_{l-1}^{T} = [\widetilde{L}_{l-1}^{T}, R_{l-1}^{T}], \tag{2}$$

where $[\cdot, \cdot]$ denotes global aggregation.

Following that, we formulate the prompting process as a local aggregation. For each LiDAR voxel feature $(\widetilde{L}_{l-1}^{T})_i$, we first search the K_{l-1} nearest neighbor radar features around the LiDAR voxel feature location, denoted as V_{l-1}. Here, $(\cdot)_i$ refers to i-th value of each voxel feature. Then, we locally aggregate the obtained nearest neighbor radar features into the LiDAR voxel feature as follows:

$$(\hat{L}_{l-1}^{T})_i = (\widetilde{L}_{l-1}^{T})_i + \sum_{k=1}^{K_{l-1}} ((V_{l-1})_i)_k, \tag{3}$$

where $\hat{L}_{l-1}^{T}$ refers to enhanced LiDAR voxel feature. Meanwhile, the radar prompt $\widetilde{R}_{l-1}^{T}$ is fed into a lightweight convolutional layer for further adaptation.

$$R_{l}^{T} = \psi(\widetilde{R}_{l-1}^{T}), \tag{4}$$

where ψ is a 3D sparse convolution layer and is the only trainable component in P_{l-1}. It updates the prompt before entering the next layer, ensuring compatibility with L_{l}^{T} by adjusting the feature and spatial size.

3.3 Inter-/Intra-modal Knowledge Distillation

In Sec. 3.2, we train the robust teacher model under various weather conditions through the proposed 4D radar-based 3D prompt learning. In this section, we distill its weather-insensitive knowledge into the LiDAR-only-based student model. To this end, we propose a novel multi-modal KD method that successfully transfers the knowledge between 3D multi-modal domains. Our KD method consists of four levels of distillation that consider the characteristics of modality.

Response-level Distillation. We first distill the teacher model responses to those of the student model. To be specific, we use classification and regression maps as responses. L1 loss and smooth L1 loss are adopted for the classification and regression maps, respectively. Through the response-level distillation, the outputs of the student model can resemble the teacher model's outputs.

$$L_{resp.}^{KD} = L1(A_{cls}^{T}, A_{cls}^{S}) + SmoothL1(A_{reg}^{T}, A_{reg}^{S}). \tag{5}$$

BEV Feature-level Distillation. Next, we propagate the knowledge from the teacher to the student by aligning the multi-scale bird-eye-view (BEV) features. Since the BEV features are already dense and pixel-wise aligned along the X-Y axes, we easily make the BEV features of the two models resemble by utilizing an L1 loss. We directly use B_{total}, which is the concatenation of multi-scale BEV features with their sizes adjusted using transpose 2D convolution blocks.

$$L_{BEV}^{KD} = L1(B_{total}^{T}, B_{total}^{S}). \tag{6}$$

Voxel Feature-level Distillation. Aside from distilling the compressed high-level features and responses along the Z-axis (Eq. 5, 6), we design voxel feature-level loss to distill 3D knowledge from the 4D radar-aided LiDAR voxel feature of teacher model L_l^T to the LiDAR voxel feature of student model L_l^S. We believed that imposing strict constraints on voxel features between the teacher (benefiting from radar) and the student (receiving only LiDAR as input) might not be optimal. Therefore, we design a loss term L_{voxel}^{KD} to optimize the cosine similarity of each i-th voxel feature between the teacher and student to be $\mathbb{1}$. Additionally, we utilize only the features from the last layer (l=3) of the teacher model for the voxel feature-level loss, considering that the low-level voxel features in the teacher model did not sufficiently incorporate the LiDAR and 4D radar information.

$$L_{voxel}^{KD} = \sum_{i=1}^{N_3} \mathbb{1} - cos((L_3^T)_i, (L_3^S)_i). \tag{7}$$

Inter-modal Structure-level Distillation. Finally, our novel and most important component is an inter-modal structural distillation that transfers the learned 4D radar and LiDAR features of the teacher model to the LiDAR voxel feature of the student model. Since our learned 4D radar prompt has abundant structural information thanks to the global aggregation of two modalities, we utilize this prompt in our distillation. To distill the teacher's global scene structure information, we use densified 3D features instead of sparse voxel features. Additionally, considering that the sparse LiDAR and radar features are not aligned in 3D space, we adopt KL divergence loss by formulating 3D multi-modal distillation with distribution matching rather than pair-wise L1 loss. Along with the radar feature, we further use dense LiDAR features of the teacher model. Our inter-modal structural distillation is formulated as follows:

$$L_{struc.}^{KD} = KL(\delta(R_3^T)||\delta(L_3^S)) + KL(\delta(L_3^T)||\delta(L_3^S)), \tag{8}$$

where $KL(\cdot||\cdot)$ is the KL divergence between two distributions and δ is a dense operation that makes sparse features into 3D dense features.

All of these distillations work in a complementary manner to effectively transfer the multi-modal teacher's knowledge to the LiDAR-only student, resulting in learning a robust and superior student model that even surpasses the teacher model under various weather conditions.

3.4 Training and Inference

Training. The output of the detection head in the baseline model is first optimized with the ground truth bounding boxes. L_{cls}, which is Focal loss [20], is adopted to train the classification and L_{reg} minimizes the regression error with smooth L1 regularization. The total loss of the baseline model is:

$$L_{baseline} = L_{cls} + L_{reg}. \tag{9}$$

Then, the teacher model, which takes LiDAR and 4D radar inputs, is optimized using the ground truth bounding boxes. It is worth noting that the parameters of the baseline model are frozen, and only a few parameters from the 4D radar prompts are updated. The loss of the teacher model is $L_{teacher} = L_{cls} + L_{reg}$, which is same as the loss of the baseline model.

The student model, which only takes LiDAR as input, is optimized with the loss terms in Sec. 3.3 and Eq. 9 at last. To distill the teacher's inter-/intra-modal knowledge, the total objective of the student model is formulated as follows:

$$L_{student} = L_{cls} + L_{reg} + \lambda_1 L^{KD}_{resp.} + \lambda_2 L^{KD}_{BEV} + \lambda_3 L^{KD}_{voxel} + \lambda_4 L^{KD}_{struc.}, \quad (10)$$

where $\lambda_1 \sim \lambda_4$ are hyper-parameters used to balance each loss.

Inference. During inference, only the LiDAR domain and the student model are required (as indicated by the green line in Fig. 2), and neither the radar nor the teacher model is utilized. From the output of the student model, robust final 3D bounding boxes are estimated under various weather conditions.

4 Experiments

4.1 Experimental Setup

Dataset and Metrics. The K-Radar dataset [25] is a large-scale 3D object detection benchmark that contains 17,458 training and 17,536 testing scenes and provides multi-sensor measurements (*e.g.* 4D radar, LiDAR, stereo camera, GPS) under seven weather conditions (normal, overcast, fog, rain, sleet, light snow, heavy snow). The K-Radar dataset is the only benchmark that provides 4D radar and LiDAR measurements under adverse weather. We adopt widely-used metrics for 3D object detection, AP_{3D} and AP_{BEV} of the class "Sedan" at IoU=0.3, following the evaluation protocol in [25]. We additionally provide the detection results at IoU=0.5 for a more comprehensive analysis.

Implementation Details. The input LiDAR data contains 64 laser beams, and the input 4D radar is pre-processed by selecting the top 10% data points with high power measurement as in [25]. We set the voxel size as (0.4m, 0.4m, 0.4m) and point cloud range as [0m, 72m] along X-axis, [-6.4m, -6.4m] along Y-axis and [-2m, 6m] along Z-axis, setting the consistent environment with [25]. The hyper-parameters $\lambda_1 \sim \lambda_4$ in Eq. 10 are 10, 10, 1, 1, respectively. All training stages are optimized with Adam optimizer, lr=1e-3, β_1=0.9 and β_2=0.999. For more details, please refer to the supplementary material.

4.2 Main Results

We compare our method with state-of-the-art 4D radar-based 3D object detection method (RTNH [25]), LiDAR-based methods (RTNH* [25], PointPillars [13], VoxelNext [2]), LiDAR-4D radar fusion-based methods (InterFusion [37], BEV-Fusion [21]) and LiDAR-based domain generalization method (Robo3D [12]). The variant of RTNH* (baseline), which takes LiDAR, and BEVFusion*, which

Table 1: Quantitative results of LiDAR and 4D radar-based 3D object detection methods on K-Radar dataset [25] (L: LiDAR, 4DR: 4D radar) at IoU=0.3. We present the detailed performance for each weather condition. Best in **bold**, second in underline.

Methods	Modality	Metric	Total	Normal	Overcast	Fog	Rain	Sleet	Lightsnow	Heavysnow
RTNH [25]	4DR	AP_{BEV}	59.1	60.1	67.3	68.7	58.0	50.8	61.1	58.2
		AP_{3D}	47.5	50.3	57.8	50.3	41.3	33.7	57.5	48.5
RTNH* [25]	L	AP_{BEV}	80.5	81.5	88.1	85.4	83.4	49.1	87.3	55.7
		AP_{3D}	70.6	72.9	76.4	83.0	68.9	43.2	85.1	50.3
PointPillars [13]	L	AP_{BEV}	50.6	58.4	61.4	43.6	61.2	19.6	44.4	29.6
		AP_{3D}	45.7	52.7	58.2	34.4	49.1	13.9	42.1	29.5
VoxelNext [2]	L	AP_{BEV}	78.4	80.8	92.9	82.6	82.7	44.0	86.5	57.1
		AP_{3D}	67.7	69.9	74.5	71.8	66.3	37.2	83.3	52.6
InterFusion [37]	4DR+L	AP_{BEV}	71.9	71.9	85.8	69.8	76.1	31.6	79.0	55.6
		AP_{3D}	59.2	58.0	73.5	62.2	61.5	28.1	76.7	53.6
BEVFusion* [21]	4DR+L	AP_{BEV}	81.5	81.7	95.3	87.3	**84.9**	**55.3**	88.4	**67.8**
		AP_{3D}	69.4	69.9	77.4	66.8	70.5	43.2	85.2	**61.6**
Robo3D [12]	L	AP_{BEV}	(66.2)	-	85.2	61.9	77.2	27.7	77.5	48.0
		AP_{3D}	(56.4)	-	84.4	50.0	65.9	25.0	72.6	44.4
Ours-T	4DR+L	AP_{BEV}	82.1	83.9	88.3	**87.7**	84.4	52.4	**89.1**	59.6
		AP_{3D}	73.2	74.4	76.9	**85.5**	70.3	**45.9**	**87.8**	54.4
Ours-S	L	AP_{BEV}	**83.3**	**85.3**	**96.0**	87.4	**84.9**	46.8	88.9	57.2
		AP_{3D}	**75.6**	**81.6**	**86.5**	85.4	**75.9**	43.2	86.6	53.6

takes LiDAR and 4D radar, are additionally adopted for comparison. Our teacher and student model will be denoted as Ours-T and Ours-S in the following. The detection results at IoU=0.5 can be seen in the supplementary material.

Table 1 shows the quantitative comparison results under various weather conditions. Ours-S outperforms single- and multi-modal 3D object detection models and even surpasses Ours-T under all metrics. Specifically, Ours-S surpasses the second-best model, BEVFusion*, up to a 6.2 AP increase, demonstrating that Ours-S successfully distilled the weather-insensitive features from the multi-modal teacher through proposed KD losses. Moreover, Ours-T reports higher performance than BEVFusion* under multiple metrics, showing the effectiveness of the proposed 4D radar-based prompt learning that integrates the strengths of both domains. Under sleet and heavy snow, Ours-S shows slightly lower performance than BEVFusion*. This might be due to the presence of many scenes with too little LiDAR data, severely limiting the amount of information in the scene (see Fig. 4). Overall, Ours-S exhibits robust performance, overcoming challenges posed by noisy or sparser LiDAR data in various weather conditions. Additionally, our two models exhibit less performance degradation as the IoU threshold increases from 0.3 to 0.5. This suggests that our model excels at accurately placing 3D bounding boxes in object locations compared to other models.

We visualize the results of our two models, the top three performing models following ours (BEVFusion*, RTNH*, VoxelNext) and the radar-only-based model (RTNH) in Fig. 4. We can see that radar-based models suffer from false positives, and LiDAR-based models struggle with adverse weather conditions or

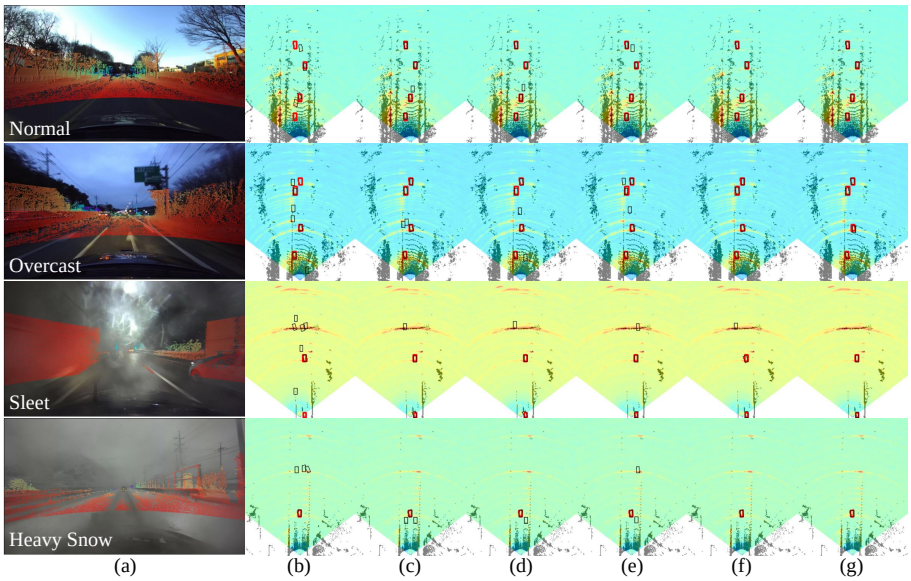

Fig. 4: Qualitative results under various weather conditions. (a) shows the image and projected LiDAR in range view. (b) to (g) shows the 4D radar, LiDAR and detection results of (b) RTNH [25], (c) RTNH* [25], (d) VoxelNext [2], (e) BEVFusion* [21], (f) Ours-T and (g) Ours-S in bird-eye-view. Red boxes are GT and black boxes are prediction of each method. More results with various weather conditions are in supplementary material. Best viewed when zoomed in with colors.

distant objects. Multi-modal-based model, BEVFusion*, fails to efficiently select and utilize information from each domain, resulting in sub-optimal outcomes. On the contrary, Ours-T successfully leverages the strength from both domains, and Ours-S even outperforms other single- and multi-modal-based methods. In sleet scenarios, the quantitative results of Ours-S may not be as favorable, but as shown in Fig. 4, Ours-S demonstrates superior accuracy in predicting bounding boxes compared to other methods when there are at least some LiDAR points available, even in situations with a limited number of points.

4.3 Model Analysis

We analyze and validate the effects of our prompt learning and KD losses by comparing them with existing works. We also discuss the impact of each component and variants of prompt learning and KD losses quantitatively and qualitatively.
Effect of Prompt Learning. Table 2 demonstrates the impact of employing prompt learning to integrate LiDAR and 4D radar. We first compare the performance with other fusion methods, including concatenation, attention, and the modified version of the existing prompt learning approach, ViPT-3D* and ViPT-BEV* [56]. Among them, Ours-T shows the overall best detection performance.

Table 2: Comparison of our prompt learning with multi-modal fusion methods, other prompt learning, and training strategies. Best in **bold**, second best in underline.

Methods	Multi-Modality	Prompt Learning	IoU=0.3		IoU=0.5		Number of Params.(M)	Model Size(MB)
			AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}		
(a) RTNH [25]			47.5	59.1	14.1	44.4	17.2	65.7
(b) RTNH* [25]			70.6	80.5	37.2	65.8	17.2	65.7
(c) Concatenation	✓		**73.3**	81.3	35.1	69.4	34.4	131.4
(d) Attention	✓		73.2	80.2	35.0	65.7	34.4	131.4
(e) ViPT-3D* [56]	✓	✓	50.6	64.4	19.1	42.4	19.9	73.3
(f) ViPT-BEV* [56]	✓	✓	69.4	81.0	31.0	64.8	20.2	77.1
(g) Ours-T (Scratch)	✓		69.8	81.0	31.5	68.5	18.4	70.3
(h) Ours-T (Finetune)	✓		71.5	81.5	35.8	70.8	18.4	70.3
(i) Ours-T	✓	✓	73.2	**82.1**	**40.4**	**71.3**	18.4	70.3

Table 3: Ablation study on the proposed 4D radar-based 3D prompt learning. Best in **bold**, second best in underline.

Methods	Local Aggregation	Self-calibration	Global Aggregation	IoU=0.3		IoU=0.5	
				AP_{3D}	AP_{BEV}	AP_{3D}	AP_{BEV}
(a)	✓			71.9 ± 0.30	80.9 ± 0.25	36.5 ± 0.36	70.1 ± 0.15
(b)	✓	✓		72.5 ± 0.40	81.1 ± 0.20	37.0 ± 1.05	70.6 ± 1.08
(c)	✓		✓	72.7 ± 0.23	81.3 ± 0.41	36.6 ± 0.88	70.7 ± 0.30
Ours-T	✓	✓	✓	**73.0** ± 0.34	**81.9** ± 0.34	**40.3** ± 0.17	**71.1** ± 0.20

This indicates that the direct fusion of LiDAR and 4D radar or existing spatial fovea-based prompts for the 2D domain does not lead to optimal performance, highlighting the effectiveness of our proposed 3D prompt-based information exchange approach that leverages the strengths of each sensor in 3D object detection task. Our model is even more efficient with only 1.2M additional parameters and an additional model size of 4.6MB for prompt learning, making the entire model smaller and more efficient for training than other models. Moreover, we compare Ours-T with models trained from scratch or finetuned in rows 7 and 8 of the table. Ours-T trained with prompt learning outperforms other two models, demonstrating that our proposed prompt learning training strategy proves to be efficient in enabling each sensor to utilize its strengths optimally.

Component Analysis of Prompt Learning. We analyze the effect of each components used in prompt learning in Table 3. The best performance is achieved when the LiDAR voxel is self-calibrated and globally aggregated with the 4D radar voxel. It suggests that self-calibration enhances the pre-trained LiDAR voxel, and global aggregation increases the amount of useful information within the scene, thereby improving the understanding of 3D sparse scenes.

Comparison with Other KD. To demonstrate the effectiveness of our KD approach, we compare it with the LiDAR-based KD method PointDistiller [48] and the multi-modal KD method UniDistill [53] in Table 4. UniDistill and PointDistiller show lower performance than Ours-S (rows 2 and 3). UniDistill distills features only near the ground truth bounding box in the BEV space, which may be insufficient for comprehensive 3D KD in diverse weather conditions.

Table 4: Comparison of our KD approach with SoTA KD methods. Effect of each components in our KD losses are further reported. Best in **bold**, second best in underline.

Methods	Resp. KD	BEV KD	Voxel KD	Struc. KD	IoU=0.3 AP_{3D} (↑)	IoU=0.3 AP_{BEV} (↑)	IoU=0.5 AP_{3D} (↑)	IoU=0.5 AP_{BEV} (↑)
Ours-T	-	-	-	-	73.2	82.1	40.4	71.3
UniDistill [53]	-	-	-	-	67.1 (-6.1)	78.6 (-3.5)	34.0 (-6.4)	68.3 (-3.0)
PointDistiller [48]	-	-	-	-	71.3 (-1.9)	81.0 (-1.1)	34.1 (-6.3)	64.0 (-7.3)
(a)	✓				72.5 (-0.7)	81.7 (-0.4)	39.7 (-0.7)	71.6 (+0.3)
(b)		✓			72.2 (-1.0)	81.3 (-0.8)	38.6 (-1.8)	71.2 (-0.1)
(c)	✓	✓			72.2 (-1.0)	82.0 (-0.1)	40.3 (-0.1)	72.0 (+0.7)
(d)	✓	✓	✓		73.1 (-0.1)	81.8 (-0.3)	40.5 (+0.1)	71.7 (+0.4)
(e)	✓	✓		✓	74.4 (+1.2)	82.9 (+0.8)	41.2 (+0.8)	**72.4** (+1.1)
Ours-S	✓	✓	✓	✓	**75.6** (+2.4)	**83.3** (+1.2)	**43.3** (+2.9)	**72.4** (+1.1)

PointDistiller, distilling graph features of 3D voxels, outperforms UniDistill but lags behind our approach. This comparison underscores the effectiveness of our method, considering four levels of distillation and inter-/intra-modal aspects.

KD Component Analysis. The rows in (a)∼(e) in Table 4 illustrate the impact of each KD loss. When both response KD and BEV KD are utilized as in (c), the performance surpasses that of each used individually, reaching a level similar to Ours-T. Applying voxel KD further slightly improves the performance in most metrics while utilizing inter-modal structural KD significantly improves the overall performance. When incorporating all four levels of distillations, it is demonstrated that Ours-S is effectively trained to be weather-insensitive.

Variant of KD Losses. We experiment with variations of the voxel KD loss and different teacher voxel features utilized in the structural KD loss. The quantitative results in the supplementary material demonstrate that our design choices of KD losses are beneficial for enhancing the LiDAR-only student model.

Objectness Score and Feature Visualization. As depicted in supplementary material, the objectness score map from the classification head indicates that the student model robustly predicts object locations compared to the baseline and teacher models, aided by our KD method. Feature activation analysis reveals that our prompt learning effectively aggregates LiDAR and radar information, leading to higher activations at locations where the model needs to focus.

5 Conclusion

We propose a robust LiDAR-based 3D object detection model under various weather conditions aided by 4D radar during training. First, our approach efficiently integrates 4D radar information into LiDAR features using prompt learning, allowing us to leverage the strengths of both sensors with minimal parameters and achieve optimal performance. Then, using the preceding model as a teacher, we distill inter-/intra-modal knowledge into a LiDAR-only student model with our novel four levels of distillation. Our LiDAR-only student model outperforms both the teacher and state-of-the-art models that utilize both modalities, demonstrating robust performance across various weather conditions.

Acknowledgements

This work was supported by the Technology Innovation Program (1415187329, 20024355, Development of autonomous driving connectivity technology based on sensor-infrastructure cooperation) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF2022R1A2B5B03002636).

References

1. Arnold, E., Al-Jarrah, O.Y., Dianati, M., Fallah, S., Oxtoby, D., Mouzakitis, A.: A survey on 3d object detection methods for autonomous driving applications. IEEE Transactions on Intelligent Transportation Systems **20**(10), 3782–3795 (2019). https://doi.org/10.1109/TITS.2019.2892405
2. Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J.: Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21674–21683 (June 2023)
3. Cho, H., Choi, J., Baek, G., Hwang, W.: itkd: Interchange transfer-based knowledge distillation for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13540–13549 (June 2023)
4. Dong, B., Zhou, P., Yan, S., Zuo, W.: Lpt: Long-tailed prompt tuning for image classification. arXiv preprint arXiv:2210.01033 (2022)
5. Gupta, H., Kotlyar, O., Andreasson, H., Lilienthal, A.J.: Robust object detection in challenging weather conditions. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7523–7532 (January 2024)
6. Hahner, M., Sakaridis, C., Bijelic, M., Heide, F., Yu, F., Dai, D., Van Gool, L.: Lidar snowfall simulation for robust 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16364–16374 (June 2022)
7. Hegde, D., Kilic, V., Sindagi, V., Cooper, A.B., Foster, M., Patel, V.M.: Source-free unsupervised domain adaptation for 3d object detection in adverse weather. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 6973–6980 (2023). https://doi.org/10.1109/ICRA48891.2023.10161341
8. Huang, K., Shi, B., Li, X., Li, X., Huang, S., Li, Y.: Multi-modal sensor fusion for auto driving perception: A survey. ArXiv **abs/2202.02703** (2022), https://api.semanticscholar.org/CorpusID:246634264
9. Huang, K.C., Wu, T.H., Su, H.T., Hsu, W.H.: Monodtr: Monocular 3d object detection with depth-aware transformer. In: CVPR (2022)
10. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022)
11. Kim, Y., Shin, J., Kim, S., Lee, I.J., Choi, J.W., Kum, D.: Crn: Camera radar net for accurate, robust, efficient 3d perception. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
12. Kong, L., Liu, Y., Li, X., Chen, R., Zhang, W., Ren, J., Pan, L., Chen, K., Liu, Z.: Robo3d: Towards robust and reliable 3d perception against corruptions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19994–20006 (October 2023)

13. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 12689–12697 (2018), https://api.semanticscholar.org/CorpusID:55701967
14. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021)
15. Li, J., Luo, C., Yang, X.: Pillarnext: Rethinking network designs for 3d object detection in lidar point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17567–17576 (June 2023)
16. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 (2021)
17. Li, Y., Xu, S., Lin, M., Yin, J., Zhang, B., Cao, X.: Representation disparity-aware distillation for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6715–6724 (October 2023)
18. Li, Y.J., Park, J., O'Toole, M., Kitani, K.: Modality-agnostic learning for radar-lidar fusion in vehicle detection. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 908–917 (2022). https://doi.org/10.1109/CVPR52688.2022.00099
19. Lin, J., Yin, H., Yan, J., Ge, W., Zhang, H., Rigoll, G.: Improved 3d object detector under snowfall weather condition based on lidar point cloud. IEEE Sensors Journal **22**(16), 16276–16292 (2022). https://doi.org/10.1109/JSEN.2022.3188985
20. Lin, T.Y., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 2999–3007 (2017), https://api.semanticscholar.org/CorpusID:47252984
21. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., Han, S.: Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: IEEE International Conference on Robotics and Automation (ICRA) (2023)
22. Meyer, M., Kuschk, G.: Automotive radar dataset for deep learning based 3d object detection. In: 2019 16th European Radar Conference (EuRAD). pp. 129–132 (2019)
23. Ngiam, J., Caine, B., Han, W., Yang, B., Chai, Y., Sun, P., Zhou, Y., Yi, X., Alsharif, O., Nguyen, P., Chen, Z., Shlens, J., Vasudevan, V.: Starnet: Targeted computation for object detection in point clouds. CoRR **abs/1908.11069** (2019), http://arxiv.org/abs/1908.11069
24. Nobis, F., Shafiei, E., Karle, P., Betz, J., Lienkamp, M.: Radar voxel fusion for 3d object detection. Applied Sciences **11**(12) (2021), https://www.mdpi.com/2076-3417/11/12/5598
25. Paek, D.H., Kong, S.H., Wijaya, K.T.: K-radar: 4d radar object detection for autonomous driving in various weather conditions. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022), https://openreview.net/forum?id=W_bsDmzwaZ7
26. Palffy, A., Pool, E., Baratam, S., Kooij, J.F.P., Gavrila, D.M.: Multi-class road user detection with 3+1d radar in the view-of-delft dataset. IEEE Robotics and Automation Letters **7**(2), 4961–4968 (2022)
27. Piroli, A., Dallabetta, V., Kopp, J., Walessa, M., Meissner, D., Dietmayer, K.: Towards robust 3d object detection in rainy conditions (2023)
28. Qian, K., Zhu, S., Zhang, X., Li, L.E.: Robust multimodal vehicle detection in foggy weather using complementary lidar and radar signals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 444–453 (June 2021)
29. Schick, T., Schütze, H.: Exploiting cloze questions for few shot text classification and natural language inference. arXiv preprint arXiv:2001.07676 (2020)

30. Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., Li, H.: Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
31. Shi, S., Jiang, L., Deng, J., Wang, Z., Guo, C., Shi, J., Wang, X., Li, H.: Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. International Journal of Computer Vision **131**, 531–551 (2021), https://api.semanticscholar.org/CorpusID:231741181
32. Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
33. Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. arXiv preprint arXiv:2010.15980 (2020)
34. The Do, A., Yoo, M.: Lossdistillnet: 3d object detection in point cloud under harsh weather conditions. IEEE Access **10**, 84882–84893 (2022). https://doi.org/10.1109/ACCESS.2022.3197765
35. Wang, C., Ma, C., Zhu, M., Yang, X.: Pointaugmenting: Cross-modal augmentation for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11794–11803 (2021)
36. Wang, L., Zhang, X., Song, Z., Bi, J., Zhang, G., Wei, H., Tang, L., Yang, L., Li, J., Jia, C., Zhao, L.: Multi-modal 3d object detection in autonomous driving: A survey and taxonomy. IEEE Transactions on Intelligent Vehicles **8**(7), 3781–3798 (2023). https://doi.org/10.1109/TIV.2023.3264658
37. Wang, L., Zhang, X., Xv, B., Zhang, J., Fu, R., Wang, X., Zhu, L., Ren, H., Lu, P., Li, J., Liu, H.: Interfusion: Interaction-based 4d radar and lidar fusion for 3d object detection. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 12247–12253 (2022). https://doi.org/10.1109/IROS47612.2022.9982123
38. Wang, Y., Yin, J., Li, W., Frossard, P., Yang, R., Shen, J.: Ssda3d: Semi-supervised domain adaptation for 3d object detection from point cloud. In: Proceedings of the AAAI Conference on Artificial Intelligence (2023)
39. Wang, Y., Deng, J., Hou, Y., Li, Y., Zhang, Y., Ji, J., Ouyang, W., Zhang, Y.: Club: Cluster meets BEV for liDAR-based 3d object detection. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
40. Wang, Y., Solomon, J.M.: Object dgcnn: 3d object detection using dynamic graphs. Advances in Neural Information Processing Systems **34**, 20745–20758 (2021)
41. Wu, H., Wen, C., Shi, S., Wang, C.: Virtual sparse convolution for multimodal 3d object detection. In: CVPR (2023)
42. Yan, Y., Mao, Y., Li, B.: Second: Sparsely embedded convolutional detection. Sensors **18**(10) (2018), https://www.mdpi.com/1424-8220/18/10/3337
43. Yang, B., Guo, R., Liang, M., Casas, S., Urtasun, R.: Radarnet: Exploiting radar for robust perception of dynamic objects. In: European Conference on Computer Vision (2020), https://api.semanticscholar.org/CorpusID:220831382
44. Yang, B., Luo, W., Urtasun, R.: Pixor: Real-time 3d object detection from point clouds. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
45. Yang, J., Shi, S., Ding, R., Wang, Z., Qi, X.: Towards efficient 3d object detection with knowledge distillation. Advances in Neural Information Processing Systems **35**, 21300–21313 (2022)

46. Yang, Z., Sun, Y., Liu, S., Jia, J.: 3dssd: Point-based 3d single stage object detector. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
47. Yin, T., Zhou, X., Krahenbuhl, P.: Center-based 3d object detection and tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11784–11793 (June 2021)
48. Zhang, L., Dong, R., Tai, H.S., Ma, K.: Pointdistiller: Structured knowledge distillation towards efficient and compact 3d detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21791–21801 (2023)
49. Zheng, W., Hong, M., Jiang, L., Fu, C.W.: Boosting 3d object detection by simulating multimodality on point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13638–13647 (2022)
50. Zheng, W., Tang, W., Jiang, L., Fu, C.W.: Se-ssd: Self-ensembling single-stage object detector from point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14494–14503 (2021)
51. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)
52. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
53. Zhou, S., Liu, W., Hu, C., Zhou, S., Ma, C.: Unidistill: A universal cross-modality knowledge distillation framework for 3d object detection in bird's-eye view. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5116–5125 (2023)
54. Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
55. Zhou, Z., Zhao, X., Wang, Y., Wang, P., Foroosh, H.: Centerformer: Center-based transformer for 3d object detection. In: ECCV (2022)
56. Zhu, J., Lai, S., Chen, X., Wang, D., Lu, H.: Visual prompt multi-modal tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9516–9526 (2023)