HVA-Fusion: Hierarchical Velocity-Aware 4D Radar-LiDAR Fusion for Robust 3D Object Detection

Jingxian Wu[1,3], Lu Wang[1]*, Lisheng Xu[1], and Jun Cheng[2]

[1] Northeastern University, Shenyang, China
[2] Chinese Academy of Sciences, Shenzhen, China
[3] Carizon, Beijing, China
wujx6@mails.neu.edu.cn, wanglu@mail.neu.edu.cn, xuls@bmie.neu.edu.cn, jun.cheng@siat.ac.cn

Abstract. 3D object detection is a core task in autonomous driving perception, with mainstream methods primarily based on LiDAR. However, its robustness deteriorates significantly in adverse weather conditions. While 4D radar offers all-weather robustness and unique Doppler velocity, it is hindered by its inherent sparsity and noise interference. Therefore, fusing LiDAR and 4D radar is highly promising. However, due to the significant differences in semantics and density between the data of these two modalities, single feature space fusion cannot fully leverage the advantages of each modality and the complementarity between modalities. To address this, we propose a two-stage fusion detection framework—HVA-Fusion. Specifically, motion velocity is first estimated through a Motion Velocity Estimation and Encoding module to enhance dynamic perception. Then, in the pillar feature space, a Cross-Modal Progressive Adaptation module is introduced for local feature alignment and fusion (Stage-I fusion) to enrich representations and mitigate semantic differences between modalities. After that, to alleviate the sparsity of 4D radar point cloud, a Radar Cross-Section (RCS) Guided Gaussian Generation module is developed to densify the feature representations. Subsequently, in the Bird's Eye View (BEV) space, the enhanced multi-modal features are adaptively fused through a Multi-Scale Bi-Directional Deformable Attention Gate module (Stage-II fusion) to address feature misalignment and sensor degradation caused by adverse weather. Extensive experiments demonstrate that HVA-Fusion achieves highly competitive performance across the K-Radar, VoD, and TJ4DRadSet datasets, while exhibiting robustness in adverse weather conditions.

Keywords: Velocity Aware · Adaptive Fusion · 3D Object Detection

1 Introduction

3D object detection is a fundamental task in perception systems for autonomous driving [6] and robotic navigation [12]. It aims to accurately localize and recog-

* Corresponding author.

nize traffic participants and represent them with 3D bounding boxes parameterized by position, size, and category [2, 11, 26, 27]. To build effective and robust perception systems, camera, LiDAR, and 4D radar have been widely adopted for 3D object detection. Although camera provides rich semantic information [1,23], their performance is limited by unreliable depth perception, sensitivity to lighting conditions, and severe occlusions [16]. In contrast, LiDAR has become the mainstream modality [10, 43, 44], as it provides dense 3D point clouds that precisely capture geometric structures while being largely insensitive to ambient illumination. However, due to its relatively short operating wavelength, LiDAR suffers from limited sensing range and substantial performance degradation under adverse weather conditions [15, 17]. Compared with LiDAR, 4D radar can provide 3D coordinates, valuable Doppler velocity, and RCS at relatively low cost [30]. Moreover, owing to its operating wavelength being much larger than atmospheric particles, 4D radar demonstrates superior robustness under adverse weather conditions [4, 13, 21]. Nevertheless, hardware limitations lead to coarse angular resolution, resulting in sparse and noisy point clouds.

LiDAR provides precise geometric structures but degrades under adverse weather. In contrast, although 4D radar produces sparse point clouds, it remains robust in complex environments and additionally offers Doppler velocity and RCS. Motivated by their complementarity, many works explore LiDAR–4D radar fusion for 3D object detection. InterFusion [42], M^2Fusion [41], and 3D-LRF [7] demonstrate advantages over single-modality methods, while RadarPillars [28] and MutualForce [31] further incorporate Doppler information to enhance dynamic perception. Nevertheless, limitations persist. First, existing methods [28,31] only apply rudimentary transformations (e.g., magnitude or squaring) to raw radial velocities as auxiliary features, failing to explicitly model the true motion state of objects. Second, the intrinsic discrepancies between LiDAR and 4D radar data pose significant challenges. While LiDAR provides dense, precise point clouds, 4D radar outputs sparse data enriched with Doppler velocity and RCS. Although existing methods [18, 38] attempt early-stage fusion to mitigate these gaps, they rely on precise spatial alignment and fail to fully model the complementary relationships in geometry, physics, and semantics. Third, the inherent sparsity of 4D radar hinders cross-modal fusion. While some approaches [34, 40] aggregate temporal frames to increase density, this often introduces feature misalignment in dynamic scenarios. Fourth, maintaining multi-modal consistency in complex environments is difficult due to noise and sensor degradation. Current methods [7, 31, 41, 42] typically rely on simple concatenation in the BEV space, lacking robust dynamic alignment and effective complementary fusion.

To address these challenges, we propose Hierarchical Velocity-Aware (HVA)-Fusion, a novel two-stage multi-modal 3D object detection framework. By leveraging hierarchical cross-modal interactions, HVA-Fusion synergizes high-precision LiDAR geometry with 4D radar's motion sensing to achieve robust detection. Specifically, the Motion Velocity Estimation and Encoding (MVEE) module first predicts point-wise motion directions to estimate true velocities, which are then encoded as point features. Subsequently, the Cross-Modal Progressive Adapta-

tion (CPA) module aligns and propagates multi-modal features at the pillar level to enrich representations and construct semantic associations via local fusion. In the BEV stage, the RCS-Guided Gaussian Generation (RGG) module leverages RCS and range priors to construct dense radar representations, effectively alleviating sparsity. Building upon this, the Multi-Scale Bi-Directional Deformable Attention Gate (MS-BDDG) achieves dynamic alignment and robust global feature fusion. In summary, our main contributions are as follows:

- We introduce a two-stage multi-modal fusion framework that performs local feature alignment and propagation in pillar space, followed by global interaction in BEV space to achieve deep, robust complementary fusion.
- We design the MVEE module, which enhances dynamic perception by estimating true motion velocities via motion direction prediction and suppresses radar noise through foreground probability filtering.
- We propose the RGG module, which incorporates RCS and range priors to construct dense radar features, effectively mitigating the negative impact of 4D radar sparsity on multi-modal fusion.
- We conduct extensive evaluations on the K-Radar [29], VoD [30], and TJ4DRadSet [58] datasets, demonstrating effectiveness and robustness of HVA-Fusion across diverse and adverse weather conditions.

2 Related Work

Single-Modal 3D Object Detection. LiDAR-based 3D object detection is the predominant approach, categorized into point-based [39, 51, 54], voxel-based [8, 20,22], and point-voxel [19,36,37] methods according to their encoding strategies. However, LiDAR performance degrades significantly under adverse weather. In contrast, 4D radar is nearly weather-invariant and provides additional Doppler velocity and RCS, exhibiting superior robustness. Recent studies have extensively explored 4D radar for 3D detection. RadarPillars [28] decomposes radial velocity and introduces a sparse-adaptive PillarAttention mechanism. SMURF [25] leverages kernel density estimation and Gaussian mixture features to mitigate sparsity and noise. MVFAN [49] jointly models geometric and dynamic features in BEV using RCS and Doppler, while MUFASA [32] complements local geometry with global multi-view context. Furthermore, MAFF-Net [5] utilizes sparse and velocity-clustered cross-attention for denoising. In summary, LiDAR-based methods lack robustness in inclement weather, and while existing 4D radar research improves feature representation through Doppler and RCS, these approaches remain constrained by the inherent limits of a single modality.

Multi-Modal 3D Object Detection. To overcome single-sensor limitations, researchers have shifted toward LiDAR-4D radar fusion, which can be categorized into two paradigms. Voxel-space fusion methods focus on local interactions. InterFusion [42] first utilized self-attention to adaptively fuse pillar features. To address its noise sensitivity, M^2Fusion [41] introduced Gaussian preprocessing and multi-scale fusion, while 3D-LRF [7] extended this to the image with a weather-gating mechanism. Despite enhancing local interactions, these

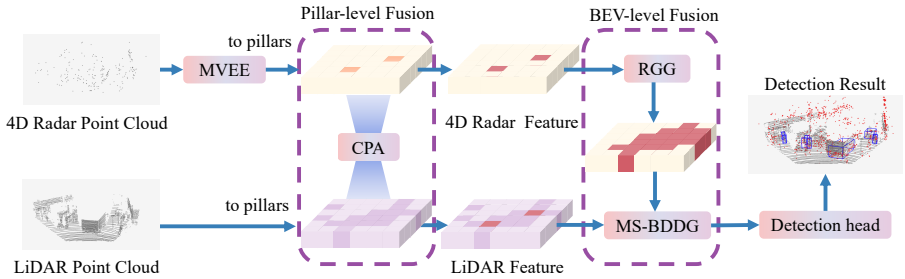

Fig. 1: Overview of the HVA-Fusion framework. First, the MVEE module estimates and encodes the true motion velocity to capture discriminative dynamic representations. Following individual pillar-based encoding for both LiDAR and radar, the CPA module performs local multi-modal integration at the pillar-level. Subsequently, the RGG module is introduced to construct dense radar features, based on which MS-BDDG performs robust and effective global complementary fusion within the BEV space.

methods often rely on simple concatenation in the BEV space, lacking global dynamic awareness and robust integration. BEV-space fusion methods focus on weighted global semantic integration after generating high-quality BEV features. LiRaFusion [38] blends multi-modal points and employs a gating network for adaptive BEV fusion, while L4DR [18] enhances this via cross-modal encoding interactions. RLNet [47] introduces spatial adaptive weights and velocity compensation. MoRAL [34] proposes motion-attention gate to focus on dynamic targets. Furthermore, MutualForce [31] utilizes Doppler and RCS to guide geometric extraction, mitigating density discrepancies via a shape-aware network. However, current research remains constrained by two primary limitations. First, the utilization of Doppler is predominantly restricted to radial velocities, lacking deep modeling of the object's true motion state. Second, these methods depend on precise modality alignment but lack local interaction mechanisms; consequently, the model cannot adaptively rectify small-scale spatial offsets, leading to semantic ambiguity in fused features and degraded detection performance. To address these challenges, we propose to recover true motion velocities by predicting point-wise motion directions, thereby fully exploiting 4D radar's inherent dynamic sensing potential. Furthermore, we leverage RCS to adaptively densify radar features. Specifically, we introduce a two-stage fusion framework that performs local feature propagation at the pillar level and dynamic global alignment in BEV space, achieving deep and robust complementary fusion.

3 Method

3.1 Network Framework

To achieve effective and robust multi-modal integration, we propose a two-stage fusion framework—HVA-Fusion. As shown in Fig. 1, HVA-Fusion integrates Li-

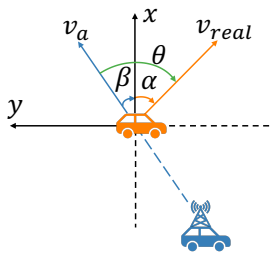

Fig. 2: Relationship between absolute radial velocity and true motion velocity.

DAR and 4D radar point clouds comprehensively through a two-stage fusion process. In particular, the Motion Velocity Estimation and Encoding (MVEE) module first predicts point-wise motion directions to recover true velocities, which are then incorporated into point features to obtain more discriminative dynamic features. Then, at the pillar level, the Cross-Modal Progressive Adaptation (CPA) module performs multi-modal feature alignment, enhances representations via co-located feature propagation, and utilizes a density-aware cross-modal neighborhood attention mechanism for local fusion, establishing more effective multi-modal semantic correlations. To mitigate the inherent sparsity of 4D radar data, the RCS-Guided Gaussian Generation (RGG) module leverages RCS and range priors to construct dense radar features. After that, in the BEV space, the Multi-Scale Bi-Directional Deformable Attention Gate (MS-BDDG) employs a three-layer hierarchical structure with multi-head deformable cross-attention to dynamically establish multi-modal correlations, alleviating the semantic ambiguity caused by feature misalignment. Furthermore, a gating network is integrated to achieve robust complementary fusion, upon which the detection head performs the final category prediction and 3D bounding box regression.

3.2 Motion Velocity Estimation and Encoding (MVEE)

As shown in Fig. 2, raw 4D radar point clouds contain both the relative radial velocity v_r and the ego-compensated absolute radial velocity v_a, which represent the radial component of the true motion velocity v_{real} along the line-of-sight β and vary with the orientation angle between the point and the sensor. However, v_a cannot reflect the actual motion state of the object because the true motion direction α is unknown. When the deviation between α and β is significant, v_a becomes much smaller than v_{real}. For instance, a pedestrian moving laterally yields a negligible v_a component along β, making them indistinguishable from the background. While existing methods [28, 31] utilize v_a as auxiliary features via magnitude or decomposition, they remain insufficient to characterize the authentic motion state. To address this, we propose MVEE to estimate v_{real} by means of predicting α, thereby obtaining more accurate and discriminative velocity information and fully unleashing the dynamic sensing potential of 4D radar.

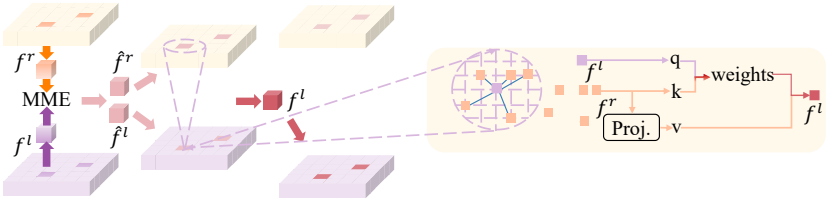

Fig. 3: The Structure of Cross-Modal Progressive Adaptation (CPA) module.

Given a 4D radar point cloud $P^r \in \mathbb{R}^{N\times 7}$, each point consists of features $[x, y, z, \Omega, v_r, v_a, T]$, where Ω is the RCS, and T is the timestamp. First, we employ the PointNet++ [35] with a direction prediction head to estimate α (defined as the angle between the true motion velocity vector and the positive x-axis). During training, the orientation label $\alpha_{heading}$ of the corresponding bounding box serves as supervision. The direction prediction loss L_{dir} is defined as:

$$L_{dir} = \frac{1}{N}\sum_{i=1}^{N}(1 - \cos(\alpha - \alpha_{heading})), \tag{1}$$

where N is the total number of positive samples. The radial direction is defined as $\beta = \arctan(y/x)$. Consequently, the deviation angle between the motion and radial directions is $\theta = \alpha - \beta$. The estimated true motion velocity v_{pred} is then computed as follows:

$$v_{pred} = \begin{cases} \frac{v_a}{\cos\theta}, & \text{if } |\cos\theta| > 10^{-1} \\ v_a, & \text{otherwise} \end{cases} \tag{2}$$

which is integrated as an additional feature to enrich the point representation. Experimental results (see Tab. 2) demonstrate that v_{pred} boosts performance, particularly for pedestrian and cyclist. This stems from the discriminative dynamic information provided by the true velocity. In particular, while dynamic objects exhibit stable velocity distributions, static objects or background noise show near-zero or stochastic patterns. Such velocity discrepancies empower the model to distinguish small dynamic objects from complex backgrounds.

To reduce radar clutter noise, we introduce a denoising operation that employs a PointNet++ [35] with a segmentation head to predict foreground probability for each 4D radar point. Points below a pre-defined confidence threshold are removed, yielding a refined point cloud that facilitates effective feature extraction and fusion for 3D object detection.

3.3 Cross-Modal Progressive Adaptation (CPA)

LiDAR and 4D radar exhibit fundamental discrepancies in data characteristics: LiDAR provides dense, precise geometric structures, whereas 4D radar yields sparse point clouds with Doppler velocity and RCS. Although widely employed

fusion strategies during encoding [18, 38] can partially narrow the representational discrepancy, they rely heavily on precise spatial alignment and fail to fully capture the complementary relationships across geometry, physics, and semantics. To address this, we propose the CPA module, as illustrated in Fig. 3. Instead of static feature aggregation, CPA dynamically calibrates the sampling size of cross-modal neighborhood search by perceiving the local distribution density of radar features. Within the pillar level, this localized adaptive mechanism performs cross-modal fusion and enriches feature representations based on radar distribution characteristics, thereby providing a high-quality foundation for subsequent global BEV-level fusion.

We first construct pillar encoding features for LiDAR and 4D radar as $\boldsymbol{h}^l = [\boldsymbol{p}^l, \boldsymbol{d}^l, \boldsymbol{o}^l, \lambda]$ and $\boldsymbol{h}^r = [\boldsymbol{p}^r, \boldsymbol{d}^r, \boldsymbol{o}^r, \boldsymbol{V}, \Omega]$, where $\boldsymbol{p}^l(\boldsymbol{p}^r)$ denotes the 3D coordinates of LiDAR (4D radar) points, $\boldsymbol{d}^l(\boldsymbol{d}^r)$ denotes the arithmetic mean distance from a LiDAR (4D radar) point to all other LiDAR (4D radar) points within its corresponding pillar, $\boldsymbol{o}^l(\boldsymbol{o}^r)$ denotes the offset of a LiDAR (4D radar) point relative to its pillar center, λ is the LiDAR intensity, and $\boldsymbol{V} = [v_r, v_{a_x}, v_{a_y}, v_{pred}]$, where $v_{a_x}(v_{a_y})$ is the x-axis(y-axis) component of v_a. Then Multi-Modal Encoder (MME) [18] is employed for cross-modal feature propagation within the same pillar coordinates, yielding enriched features $\hat{\boldsymbol{h}}^l = [\boldsymbol{p}^l, \boldsymbol{d}^l, \boldsymbol{d}^{l\to r}, \boldsymbol{o}^l, \lambda, \bar{\boldsymbol{V}}, \bar{\Omega}]$ and $\hat{\boldsymbol{h}}^r = [\boldsymbol{p}^r, \boldsymbol{d}^{r\to l}, \boldsymbol{d}^r, \boldsymbol{o}^r, \bar{\lambda}, \boldsymbol{V}, \Omega]$, where $\boldsymbol{d}^{l\to r}$ ($\boldsymbol{d}^{r\to l}$) is the arithmetic mean distance from a LiDAR (4D radar) point to all 4D radar (LiDAR) points within the same pillar, and the overline denotes the average of all point features from the other modality within the same pillar. $\hat{\boldsymbol{h}}^l$ and $\hat{\boldsymbol{h}}^r$ are further processed via linear layers and max-pooling to obtain pillar features f^l and f^r. While MME's early-stage bidirectional fusion enriches representation, it relies heavily on precise spatial alignment. To mitigate performance degradation caused by spatial misalignments and establish effective semantic connections, we propose a cross-modal neighborhood adaptive fusion strategy within CPA. Specifically, for each LiDAR pillar feature f_i^l, we estimate the local density of radar pillar features within a radius r_l around its spatial location. The estimated density dynamically determines the sampling size k_l for adaptive k-nearest neighbor (k-NN) search, yielding a radar neighborhood $\mathcal{N}_i^r$ for f_i^l. Cross-modal attention is then performed by taking f_i^l as the query and the neighboring radar features $\{f_j^r \mid j \in \mathcal{N}_i^r\}$ as keys and values. The resulting attention weights are applied to these neighboring radar features to generate the enhanced LiDAR feature $\hat{f}_i^l$. Through this attention-weighting mechanism, LiDAR features in regions with dense radar support receive stronger responses, effectively indicating higher foreground confidence. Conversely, features in sparse regions receive constrained enhancement. This adaptive strategy bridges the semantic gap between LiDAR and 4D radar and establishes more effective cross-modal semantic associations.

3.4 RCS-Guided Gaussian Generation (RGG)

The inherent density discrepancy between sparse 4D radar points and dense LiDAR points often leads to LiDAR features dominating the fusion, causing radar

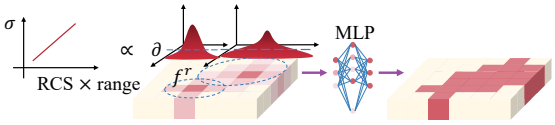

Fig. 4: The Structure of RGG module.

to be suppressed as noise. While temporal accumulation [34, 40] mitigates sparsity, it introduces potential motion misalignment. To address this, we propose the RGG module. Leveraging the positive correlation between an object's physical size and its RCS [24], RGG generates features via an adaptive Gaussian distribution.

For each non-empty radar feature f^r, RGG constructs the Gaussian distribution $G_i(x, y) = \exp(-((x - x_i)^2 + (y - y_i)^2)/2\sigma_i^2)$, where (x_i, y_i) are the coordinates of f^r, σ_i is the scale coefficient which is conditioned on its RCS and range attributes, as shown in Fig. 4. Positions with weights below $\delta = 0.1$ are filtered. Specifically, we first derive a physically-constrained baseline diffusion scale $\sigma_i^{base} = \sigma_{min} + (\sigma_{max} - \sigma_{min}) \cdot \text{Sigmoid}\left(\kappa(RCS_i - \mu)\right)$ based on the RCS, where κ controls the slope of the Sigmoid function, and μ represents the mean offset of the RCS. The hyperparameters σ_{min} and σ_{max} define the lower and upper bounds of the scale control coefficients, respectively. This mapping ensures that σ_i varies monotonically with the RCS, thereby maintaining physical plausibility. To compensate for the limitations of the physical prior in complex scenarios, we introduce a learnable refinement term $\sigma_i^{learn} = \sigma_{min} + (\sigma_{max} - \sigma_{min}) \cdot \text{MLP}(\text{Sigmoid}(RCS_i))$, where the MLP takes the Sigmoid-normalized RCS as input to adaptively refine the diffusion scale. Considering the range-dependent sparsity of 4D radar in the BEV space, we define a linear modulation factor based on the Euclidean distance d_i relative to the sensor origin, where d_i is normalized to $[0, 1]$. The final diffusion scale $\sigma_i = \text{clip}\left(d_i \cdot (\gamma\sigma_i^{base} + (1 - \gamma)\sigma_i^{learn}), \sigma_{min}, \sigma_{max}\right)$ is obtained by a weighted fusion of σ_i^{base} and σ_i^{learn}, where $\text{clip}(\cdot)$ is the truncation operator that constrains σ_i within the interval $[\sigma_{min}, \sigma_{max}]$. This mechanism ensures localized precision for near-range objects while expanding the diffusion scale for far-range or high-RCS targets to alleviate sparsity. The generated feature f^g is obtained via Gaussian-weighted averaging as follows:

$$f^g = \frac{\sum_{i \in \mathcal{P}} w_i \cdot f_i^r}{\sum_{i \in \mathcal{P}} w_i}, \tag{3}$$

where $\mathcal{P}$ is the set of radar pillars covering the current location and w_i is the Gaussian weight. Finally, f^g is refined by a lightweight MLP and projected into the BEV space, yielding the densified 4D radar BEV feature F^r. The RGG module effectively densifies 4D radar features and mitigates the inherent sparsity of radar point clouds. This leads to substantial performance gains, particularly for long-range object detection.

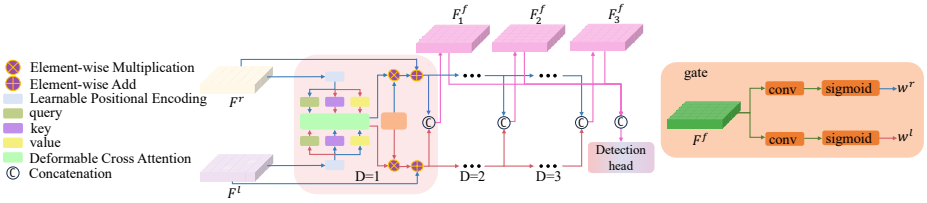

Fig. 5: The Structure of MS-BDDG module.

3.5 Multi-Scale Bi-Directional Deformable Attention Gate (MS-BDDG)

Despite the natural complementarity between LiDAR and 4D radar, achieving strict alignment in the BEV space remains challenging due to inaccurate calibration and inherent azimuthal noise in radar measurements. Such misalignment hinders effective cross-modal interaction and limits the performance ceiling of fusion frameworks. To address this, we propose the MS-BDDG module, which achieves dynamic alignment and robust fusion through multi-scale bidirectional interactions with learnable spatial offsets and gated weights.

As illustrated in Fig. 5, we construct a three-layer multi-scale framework to progressively implement cross-modal interactions: shallower layers prioritize geometric complementarity, while deeper layers enforce semantic consistency. The bidirectional design prevents modal dominance, ensuring balanced multi-modal representations. Within each layer, deformable cross-attention [59] is introduced to dynamically establish correspondences. For LiDAR BEV features F^l and 4D radar BEV features F^r, we first employ learnable position embeddings and decoupled content-position encoding to enhance representation. Transforming F^l into the query sequence f_q^l with reference points p_q^l, and F^r as keys and values, the deformable attention for layer $D \in \{1, 2, 3\}$ is formulated as follows:

$$\text{DeformAttn}(f_q^l, p_q^l, F_D^r) = \sum_{h=1}^{H} W_h \left[\sum_{k=1}^{K} A_{hqk} \cdot W_h' F_D^r(p_q^l + \Delta p_{hqk}) \right], \quad (4)$$

where H and K is the number of heads and sampling points, and Δp_{hqk} is the sampling offset. Subsequently, we exchange the roles of the modalities to perform the bidirectional cross-modal interaction:

$$\begin{cases} F_D^l = \text{DeformAttn}(f_q^l, p_q^l, F_D^r) + F_D^l \\ F_D^r = \text{DeformAttn}(f_q^r, p_q^r, F_D^l) + F_D^r. \end{cases} \quad (5)$$

Recognizing that LiDAR degrades in adverse weather and 4D radar suffers from multipath reflections, multi-modal fusion may yield negative effects. Inspired by LiRaFusion [38], we introduce a Modal Adaptive Gating mechanism. Specifically, independent convolutional blocks extract features from F^l, F^r, and the fusion feature F^f in each layer before entering the gating network. Within

the gating network, gating weights w^m are obtained as follows:

$$w^m = \text{Sigmoid}(\text{block}(F_D^f)), \quad m \in \{l, r\}, \tag{6}$$

where block is the 3×3 convolution. These weights adaptively modulate F^l and F^r via element-wise multiplication:

$$F_D^m = F_D^m \cdot w^m, \quad m \in \{l, r\} \tag{7}$$

suppressing noise from low-quality modalities while highlighting complementary information. Finally, the features from all scales are concatenated.

3.6 Loss Function and Training Strategy

We trained our HVA-Fusion with the following losses:

$$L_{all} = \beta_{cls} L_{cls} + \beta_{loc} L_{loc} + \beta_{mask} L_{mask} + \beta_{dir} L_{dir}, \tag{8}$$

where $\beta_{cls} = 1.0$, $\beta_{loc} = 2.0$, $\beta_{mask} = 2.0$, $\beta_{dir} = 1.1$. As with L4DR [18], we adopt Focal Loss for object classification and foreground point classification, Smooth L1 Loss for object location regression and L_{dir} in Eq. (1).

Random Modal Dropout. To enhance the model's robustness against sensor failure or data absence, we introduce a Random Modal Dropout strategy during the training phase. Specifically, a random variable $p_1 \sim U(0, 1)$ is sampled; if $p_1 \leq P_{drop}$, a second variable $p_2 \sim U(0, 1)$ determines the dropout modality. If $p_2 > P_{lidar\_thres}$, the 4D radar BEV feature F^r is zeroed out; otherwise, the LiDAR BEV feature F^l is discarded. Simulating single-modality failures promotes robust feature learning and improves generalization across diverse scenarios.

4 Experiments

4.1 Implement Details

We implement HVA-Fusion with L4DR [18] and conduct experiments on 2 NVIDIA H20 GPUs. We set κ as 0.5, $\mu = 0$, $\gamma = 0.9$, $\sigma_{min} = 0.3$, $\sigma_{max} = 1.2$, $H \in \{8, 4, 4\}, K \in \{8, 4, 4\}$, and $P_{drop} = P_{lidar\_thres} = 0.2$. We use Adam optimizer with lr = 1e-3, $\beta_1 = 0.9$, $\beta_2 = 0.999$.

4.2 Dataset and Evaluation Metrics

K-Radar dataset. The K-Radar dataset [29] contains 35K frames from 58 sequences collected under diverse weather conditions (e.g., overcast, fog, and heavy snow). The data is partitioned into 17,458 training and 17,536 testing frames. We report AP_{3D} and AP_{BEV} for the class "Sedan" at IoU = 0.5.

View-of-Delft (VoD) dataset. The VoD dataset [30] contains 8,693 frames, with 5,139 frames for training and 1,296 for validation. Evaluation follows official metrics across two regions: 1) Entire Area (camera FoV up to 50 meters)

Table 1: Experimental Results on the K-Radar dataset. For each method, the sensor modalities (L: LiDAR, R: 4D radar) and detailed performance (in %) are reported. Best in **bold**, second in underline.

Methods	Modality	Metrics	Total	Normal	Overcast	Fog	Rain	Sleet	Lightsnow	Heavysnow
RTNH [29]	R	AP_{BEV}	41.1	41.0	44.6	45.4	32.9	50.6	81.5	56.3
		AP_{3D}	37.4	37.6	42.0	41.2	29.2	49.1	63.9	43.1
PointPillars [22]	L	AP_{BEV}	49.1	48.2	53.0	45.4	44.2	45.9	74.5	53.8
		AP_{3D}	22.4	21.8	28.0	28.2	27.2	22.6	23.2	12.9
RTNH [29]	L	AP_{BEV}	66.3	65.4	87.4	83.8	73.7	48.8	78.5	48.1
		AP_{3D}	37.8	39.8	46.3	59.8	28.2	31.4	50.7	24.6
InterFusion [42]	L+R	AP_{BEV}	52.9	50.0	59.0	80.3	50.0	22.7	72.2	53.3
		AP_{3D}	17.5	15.3	20.5	47.6	12.9	9.33	56.8	25.7
3D-LRF [7]	L+R	AP_{BEV}	73.6	72.3	88.4	86.6	76.6	47.5	79.6	**64.1**
		AP_{3D}	45.2	45.3	55.8	51.8	38.3	23.4	60.2	36.9
L4DR [18]	L+R	AP_{BEV}	77.5	**76.8**	88.6	89.7	78.2	59.3	80.9	53.8
		AP_{3D}	53.5	53.0	64.1	73.2	53.8	46.2	52.4	37.0
HVA-Fusion(Ours)	L+R	AP_{BEV}	**78.2**	76.7	**89.9**	**90.7**	**80.6**	**64.1**	**87.5**	60.8
		AP_{3D}	**65.7**	**63.8**	**87.7**	**79.7**	**65.5**	**60.7**	**77.1**	**57.7**

and 2) Driving Area ($[-4m < x < +4m, z < +25m]$). IoU are set to 0.5, 0.25, and 0.25 for car, pedestrian, and cyclist, respectively. Additionally, we evaluate on **VoD-Fog** [18], which simulates fog effects [15] on LiDAR points (fog level $L = [0, 1, 2, 3, 4]$, fog density $D = [0.00, 0.03, 0.06, 0.10, 0.20]$) while keeping 4D radar intact. The KITTI official metrics are used to analyze the performance of objects with different difficulties.

TJ4DRadSet. The TJ4DRadSet [58] provides 7,757 high-quality annotated frames (5,717 for training, 2,040 for testing). Performance is measured by AP_{3D} and AP_{BEV}, with IoU of 0.5 for car and truck, 0.25 for pedestrian and cyclist.

4.3 Experimental Results

Results on K-Radar Adverse Weather Dataset. Tab. 1 summarizes the results on the K-Radar dataset. By effectively synergizing fine-grained geometric structures from LiDAR with Doppler velocity information from 4D radar, our multi-modal approach consistently outperforms single-modality methods, especially in adverse weather conditions. Instead of relying on naive fusion, our method incorporates motion-aware priors to enhance feature representations. Furthermore, our two-stage fusion framework bridges the density and semantic gaps at both the pillar and BEV levels. This hierarchical alignment facilitates robust complementary fusion, achieving superior performance across diverse environmental conditions.

Results on VoD Dataset. As shown in Tab. 2, HVA-Fusion establishes a new state-of-the-art (SOTA) on the Entire Area subset with 75.51% mAP. Notably, it excels in pedestrian (72.46% AP) and cyclist (83.96% AP) detection. This performance stems from: 1) the MVEE module, which provides discriminative dynamic semantic features by estimating true motion states; 2) the CPA

Table 2: Experimental Results on the VoD dataset.

Methods	Modality	Entire Area AP				Driving Area AP			
		Car	Ped.	Cyc.	mAP	Car	Ped.	Cyc.	mAP
MVFAN [49]	R	38.12	30.96	66.17	45.08	71.45	40.21	86.63	66.10
SMURF [25]	R	42.31	39.09	71.50	50.97	71.74	50.54	86.87	69.72
MUFASA [32]	R	43.10	38.97	68.65	50.24	72.50	50.28	88.51	70.43
RadarPillars [28]	R	41.10	38.60	72.60	50.70	71.10	52.30	87.90	70.50
SCKD [48]	R	41.89	43.51	70.83	52.08	77.54	51.06	86.89	71.80
UniBEVFusion [55]	C+R	42.22	47.11	72.94	54.09	72.10	57.71	93.29	74.37
Doracamom-S [57]	C+R	53.35	48.94	76.99	59.76	-	-	-	-
HGSFusion [14]	C+R	51.67	52.64	72.58	58.96	88.28	62.61	87.49	79.46
TL-4DRCF [53]	C+R	43.71	40.11	64.22	49.35	79.49	53.76	76.50	69.92
LXLv2 [46]	C+R	47.81	49.30	77.15	58.09	-	-	-	-
SGDet3D [3]	C+R	53.16	49.98	76.11	59.75	81.13	60.91	90.22	77.42
PointPillars [22]	L	65.55	55.71	72.96	64.74	81.10	67.92	88.96	79.33
LXL-Pointpillars [45]	L	66.60	56.10	75.10	65.90	-	-	-	-
InterFusion [42]	L+R	67.50	63.21	78.79	69.83	88.11	74.80	87.50	83.47
L4DR [18]	L+R	69.10	66.20	82.80	72.70	90.80	76.10	95.50	87.47
MutualForce [31]	L+R	71.67	66.26	77.35	71.76	**92.31**	76.79	89.97	86.36
CM-FA [9]	L+R	**77.52**	67.99	75.97	73.83	90.91	80.78	87.80	86.50
SVEFusion [50]	L+R	71.83	67.85	**83.97**	74.55	90.91	79.37	96.31	**88.86**
MoRAL [34]	L+R	71.23	69.67	79.01	73.30	90.91	78.90	96.25	88.68
ELMAR [33]	L+R	76.41	69.34	78.91	74.89	91.74	**81.35**	93.01	88.70
HVA-Fusion(Ours)	L+R	70.12	**72.46**	83.96	**75.51**	90.91	78.03	**96.51**	88.48

module, which strengthens dynamic perception via local cross-modal feature interactions; and 3) the RGG module, which leverages RCS to recover semantic integrity for distant, sparse objects. Nevertheless, two scenarios exhibit comparatively limited gains. The first concerns car detection in the Entire Area. VoD cars are mostly static and distant [30], benefiting less from dynamic semantics emphasized by our framework. The remaining gap is mainly due to backbone representation: point-based backbones preserve finer geometry for distant, sparse objects, whereas our pillar-based framework is more affected by spatial quantization. The second pertains to pedestrian detection in the Driving Area, where the isotropic Gaussian kernels in RGG are misaligned with the small and irregular spatial footprint of nearby pedestrians, thereby blurring their precise contours. We leave the exploration of shape-aware anisotropic distributions to future work.

Results on VoD-Fog Simulated Dataset. Tab. 3 summarizes the results on VoD-Fog under simulated adverse weather. Multi-modal methods incorporating 4D radar consistently outperform LiDAR-only baselines across all difficulty levels, with the performance gap widening as fog density increases. Notably, HVA-Fusion demonstrates robust performance across fog levels 0–2 by effectively exploiting Doppler velocity, RCS, and robust complementary fusion. However, under extreme fog conditions, our method is surpassed by L4DR [18] in the cyclist category. This discrepancy suggests that while the CPA module enhances semantic consistency for cars and pedestrians via local cross-modal interactions, the complex Doppler patterns of cyclists—stemming from rotating wheels and metallic frames—pose a challenge. This effect is further exacerbated

Table 3: Quantitative comparison of various 3D object detection methods on the VoD-Fog dataset using KITTI evaluation metrics under different fog intensities.

Fog Level	Methods	Modality	**mAP**	**Car**			**Pedestrian**			**Cyclist**		
				Easy	**Mod.**	**Hard**	**Easy**	**Mod.**	**Hard**	**Easy**	**Mod.**	**Hard**
0	PointPillars [22]	L	70.8	84.9	73.5	67.5	62.7	58.4	53.4	85.5	79.0	72.7
	InterFusion [42]	L+R	73.2	67.6	65.8	58.8	73.7	70.1	64.7	90.3	87.0	81.2
	L4DR [18]	L+R	78.9	85.0	76.6	69.4	74.4	72.3	65.7	93.4	90.4	83.0
	HVA-Fusion(Ours)	L+R	**81.2**	**86.3**	**77.3**	**69.8**	**78.8**	**77.3**	**71.7**	**94.3**	**91.4**	**84.6**
1	PointPillars [22]	L	69.0	79.9	72.7	67.0	59.9	55.6	50.5	85.5	78.2	72.0
	InterFusion [42]	L+R	73.0	66.1	64.0	56.9	74.0	70.6	64.5	91.6	87.4	82.0
	L4DR [18]	L+R	77.9	77.9	73.2	67.8	75.4	72.1	66.7	93.8	**91.0**	83.2
	HVA-Fusion(Ours)	L+R	**79.6**	**81.6**	**74.8**	**69.1**	**75.6**	**74.6**	**69.5**	**95.3**	90.6	**85.3**
2	PointPillars [22]	L	55.0	67.0	51.4	44.4	53.1	47.2	42.7	69.6	62.7	57.2
	InterFusion [42]	L+R	59.3	56.0	48.5	41.5	63.2	57.8	52.9	77.3	71.1	66.2
	L4DR [18]	L+R	64.0	68.5	56.4	49.3	63.1	59.9	55.1	**82.7**	70.8	**70.7**
	HVA-Fusion(Ours)	L+R	**65.8**	**72.2**	**59.8**	**52.4**	**64.1**	**61.1**	**56.2**	81.1	**76.4**	69.4
3	PointPillars [22]	L	39.6	44.5	31.9	27.0	40.2	37.7	34.0	53.2	46.7	41.8
	InterFusion [42]	L+R	46.5	41.2	33.1	27.0	52.9	49.2	44.8	59.9	57.7	53.1
	L4DR [18]	L+R	**52.5**	46.2	41.4	34.6	**53.5**	50.6	46.2	**72.2**	**67.7**	**60.9**
	HVA-Fusion(Ours)	L+R	52.4	**53.3**	**44.7**	**37.7**	52.4	**51.1**	**46.6**	66.8	62.4	57.3
4	PointPillars [22]	L	8.8	13.0	8.77	7.19	10.6	12.9	11.3	6.15	4.89	4.57
	InterFusion [42]	L+R	14.3	15.2	10.8	8.40	25.7	25.1	22.6	6.68	7.95	6.99
	L4DR [18]	L+R	**28.0**	26.9	26.2	21.6	**33.1**	30.7	**27.9**	**30.3**	**29.7**	**26.3**
	HVA-Fusion(Ours)	L+R	25.8	**37.5**	**32.7**	**26.2**	30.7	**31.6**	26.3	15.7	16.0	15.9

Table 4: Experimental Results on the TJ4DRadSet dataset. † denotes results reproduced by us using the official code under the same experimental settings.

Methods	Modality	$AP_{\mathbf{3D}}$					$AP_{\mathbf{BEV}}$				
		Car	Ped.	Cyc.	Tru.	**mAP**	Car	Ped.	Cyc.	Tru.	**mAP**
CenterPoint [52]	R	22.03	25.02	53.32	15.92	29.07	33.03	27.87	58.74	25.09	36.18
SMURF [25]	R	28.47	26.22	54.61	22.64	32.99	43.13	29.19	58.81	32.80	40.98
RCFusion [56]	C+R	29.72	27.17	54.93	23.56	33.85	40.89	30.95	58.30	28.92	39.76
UniBEVFusion [55]	C+R	44.26	27.92	51.11	27.75	37.76	50.43	29.57	56.48	35.22	42.92
HGSFusion [14]	C+R	-	-	-	-	37.21	-	-	-	-	43.23
SGDet3D [3]	C+R	59.43	26.57	51.30	30.00	41.82	66.38	29.18	53.72	39.36	47.16
Doracamom-S [57]	C+R	52.18	33.61	56.38	34.79	44.24	-	-	-	-	-
CM-FA† [9]	L+R	**86.94**	47.34	89.03	**42.26**	66.39	**89.41**	47.91	89.05	**54.39**	70.19
L4DR† [18]	L+R	83.54	58.86	85.59	35.00	65.74	89.13	59.98	85.69	48.82	70.90
HVA-Fusion(Ours)	L+R	83.62	**67.54**	**89.41**	39.79	**70.09**	87.94	**68.80**	**89.98**	51.87	**74.64**

under extreme fog, which degrades local cross-modal correspondence and makes the associations in CPA unreliable. In contrast, L4DR avoids such unreliable associations by bypassing local fusion. To probe this, we conducted preliminary inference-time experiments on Fog Level 4 by varying the CPA interaction strength (0/0.5/1.0). Disabling CPA drops Entire Area mAP by 1.76%, while tuning down the interaction strength gains 0.31%. These results provide preliminary support for the feasibility of Signal-to-Noise Ratio (SNR)-aware adaptive interaction to improve perception upper bounds in extreme environments.

Table 5: Ablation study of different components on the VoD dataset.

Module				Entire Area	Driving Area
MVEE	CPA	RGG	MS-BDDG	mAP	mAP
				70.99	83.95
✓				73.36	85.35
✓	✓			73.71	88.14
✓	✓	✓		74.11	88.24
✓			✓	74.02	88.01
✓		✓	✓	74.20	87.92
✓	✓		✓	74.77	88.43
✓	✓	✓	✓	**75.51**	**88.48**

Results on TJ4DRadSet Dataset. To evaluate the generalization capability of HVA-Fusion, we conducted additional experiments on the TJ4DRadSet dataset. As demonstrated in Tab. 4, our method consistently outperforms existing LiDAR-4D radar fusion approaches in both AP_{3D} and AP_{BEV}. Furthermore, it surpasses models based on other types of modality combinations by a notable margin, establishing a new SOTA performance.

Inference Time. On VoD with an NVIDIA H20 GPU, HVA-Fusion has a total inference time of 148 ms: 85 ms for the L4DR baseline, plus 26, 5, 4, and 28 ms for MVEE, CPA, RGG, and MS-BDDG, respectively.

4.4 Ablation study

We conduct ablation study on the VoD dataset and the results are shown in Tab. 5. The 1$^{\text{st}}$ row represents the performance of L4DR [18], where LiDAR and 4D radar BEV features are fused via direct concatenation.

Effects of MVEE. Building upon the baseline (Row 1), the MVEE module recovers true point velocities via motion direction prediction, which serves as an auxiliary encoding to provide discriminative dynamic features. This yields mAP gains of 2.37% and 1.40% in the Entire and Driving Areas, respectively.

Effects of CPA. Adding CPA on top of MVEE, the module performs feature propagation at the pillar level to enrich multi-modal representations, further narrowing semantic discrepancies through cross-modal neighborhood adaptive fusion. This early-stage local integration yields additional mAP gains of 0.35% and 2.79% in the Entire and Driving Areas, respectively. Notably, the more pronounced improvement in the feature-rich Driving Area demonstrates that CPA establishes effective semantic correlations between LiDAR's precise geometric features and 4D radar's dynamic attributes through adaptive local interaction. Removing CPA from the complete HVA-Fusion degrades the Entire Area mAP by 1.31%, as the absence of pillar-level cross-modal propagation forces subsequent modules to fuse insufficiently aligned features, thereby hindering effective multi-modal complementarity.

Effects of RGG. Further incorporating RGG upon the MVEE+CPA configuration, the module adaptively densifies 4D radar features based on RCS and range-aware priors. This mechanism yields mAP gains of 0.40% and 0.10% in the

Entire and Driving Areas, respectively, demonstrating its capacity to adaptively expand the diffusion scale for long-range or high-RCS pillars, thereby effectively mitigating the sparsity problem. In contrast, removing RGG from the complete HVA-Fusion decreases the Entire Area mAP by 0.74%, since without the RCS and range-aware densification the 4D radar features remain overly sparse and fail to provide a sufficiently informative basis for the subsequent fusion.

Effects of MS-BDDG. Finally, integrating MS-BDDG to form the complete HVA-Fusion, the module synergizes deformable cross-attention with a gating network to achieve dynamic alignment and robust complementary fusion of high-level multi-modal semantics. By dynamically aligning features and suppressing environmental noise, MS-BDDG effectively mitigates the negative impact of occlusions and adverse weather, particularly in non-driving areas. This mechanism yields mAP gains of 1.40% and 0.24% in the Entire and Driving Areas, respectively. Removing MS-BDDG from the complete HVA-Fusion degrades the Entire Area mAP by 1.40%, confirming that its gated deformable alignment plays an effective and robust role in high-level multi-modal semantic fusion.

5 Conclusion

In this paper, we introduce a novel two-stage fusion framework—HVA-Fusion. Within the HVA-Fusion framework, we first introduce the MVEE module, which bolsters dynamic object perception by estimating true motion states to obtain discriminative dynamic representations. Subsequently, the CPA module facilitates feature propagation and neighborhood cross-modal attention in the pillar space, enriching feature representations while establishing robust semantic associations. The RGG module then leverages RCS and range priors modulation to achieve effective 4D radar feature densification. Finally, the MS-BDDG module implements dynamic feature alignment and robust complementary fusion within the BEV space. Extensive experimental results demonstrate that HVA-Fusion achieves highly competitive performance across the K-Radar, VoD, and TJ4DRadSet datasets. Future work will explore shape-aware anisotropic feature generation and SNR-based adaptive fusion strategies to further push the perception boundaries in extreme all-weather environments. Code & Weights: `https://github.com/wujingxian1/HVA-Fusion`.

Acknowledgements

This research was supported in part by Guangdong Major Project of Basic and Applied Basic Research (2023B0303000016), the National Natural Science Foundation of China (U21A20487, 62273082), Guangdong Technology Project (2023TX07Z126), Shenzhen High-tech Zone Development Special Plan Innovation Platform Construction Project, the proof of concept center for high precision and high resolution 4D imaging, CAS Key Technology Talent Program.

References

1. Alaba, S.Y., Ball, J.E.: A survey on deep-learning-based lidar 3d object detection for autonomous driving. Sensors **22**(24), 9577 (2022)
2. Aung, N.H.H., Sangwongngam, P., Jintamethasawat, R., Shah, S., Wuttisittikulkij, L.: A review of lidar-based 3d object detection via deep learning approaches towards robust connected and autonomous vehicles. IEEE Transactions on Intelligent Vehicles **10**(1), 526–547 (2025)
3. Bai, X., Yu, Z., Zheng, L., Zhang, X., Zhou, Z., Zhang, X., Wang, F., Bai, J., Shen, H.L.: Sgdet3d: Semantics and geometry fusion for 3d object detection using 4d radar and camera. IEEE Robotics and Automation Letters **10**(1), 828–835 (2025)
4. Balal, N., Pinhasi, G.A., Pinhasi, Y.: Atmospheric and fog effects on ultra-wide band radar operating at extremely high frequencies. Sensors **16**(5), 751 (2016)
5. Bi, X., Weng, C., Tong, P., Fan, B., Eichberge, A.: Maff-net: enhancing 3d object detection with 4d radar via multi-assist feature fusion. IEEE Robotics and Automation Letters **10**(5), 4284–4291 (2025)
6. Bijelic, M., Gruber, T., Mannan, F., Kraus, F., Ritter, W., Dietmayer, K., Heide, F.: Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11679–11689 (2020)
7. Chae, Y., Kim, H., Yoon, K.J.: Towards robust 3d object detection with lidar and 4d radar fusion in various weather conditions. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15162–15172 (2024)
8. Chen, Y., Liu, J., Zhang, X., Qi, X., Jia, J.: Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21674–21683 (2023)
9. Deng, J., Chan, G., Zhong, H., Lu, C.X.: Robust 3d object detection from lidar-radar point clouds via cross-modal feature augmentation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 6585–6591. IEEE (2024)
10. Deng, J., Ye, W., Wu, H., Huang, X., Xia, Q., Li, X., Fang, J., Li, W., Wen, C., Wang, C.: Cmd: A cross mechanism domain adaptation dataset for 3d object detection. In: Computer Vision – ECCV 2024. pp. 219–236. Springer (2025)
11. Ghasemieh, A., Kashef, R.: 3d object detection for autonomous driving: Methods, models, sensors, data, and challenges. Transportation Engineering **8**, 100115 (2022)
12. Ghita, A., Antoniussen, B., Zimmer, W., Greer, R., Creß, C., Møgelmose, A., Trivedi, M.M., Knoll, A.C.: Activeanno3d-an active learning framework for multi-modal 3d object detection. In: 2024 IEEE Intelligent Vehicles Symposium (IV). pp. 1699–1706. IEEE (2024)
13. Golovachev, Y., Etinger, A., Pinhasi, G.A., Pinhasi, Y.: Millimeter wave high resolution radar accuracy in fog conditions—theory and experimental verification. Sensors **18**(7), 2148 (2018)
14. Gu, Z., Ma, J., Huang, Y., Wei, H., Chen, Z., Zhang, H., Hong, W.: Hgsfusion: Radar-camera fusion with hybrid generation and synchronization for 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 3185–3193 (2025)
15. Hahner, M., Sakaridis, C., Dai, D., Van Gool, L.: Fog simulation on real lidar point clouds for 3d object detection in adverse weather. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15263–15272 (2021)

16. Hu, H., Wang, F., Su, J., Wang, Y., Hu, L., Fang, W., Xu, J., Zhang, Z.: Ea-lss: Edge-aware lift-splat-shot framework for 3d bev object detection. arXiv preprint arXiv:2303.17895 (2023)
17. Huang, X., Wu, H., Li, X., Fan, X., Wen, C., Wang, C.: Sunshine to rainstorm: Cross-weather knowledge distillation for robust 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 2409–2416 (2024)
18. Huang, X., Xu, Z., Wu, H., Wang, J., Xia, Q., Xia, Y., Li, J., Gao, K., Wen, C., Wang, C.: L4dr: Lidar-4dradar fusion for weather-robust 3d object detection. In: Proceedings of the AAAI conference on Artificial Intelligence. pp. 3806–3814 (2025)
19. Jiang, Y., Xie, Q., Li, J., Xu, J., Liu, Y., Ma, Y.: Dpa-rcnn: Dual position aware 3d object detector for point cloud. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–9. IEEE (2024)
20. Jin, X., Liu, K., Ma, C., Yang, R., Hui, F., Wu, W.: Swiftpillars: High-efficiency pillar encoder for lidar-based 3d detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 2625–2633 (2024)
21. Kim, Y., Shin, J., Kim, S., Lee, I.J., Choi, J.W., Kum, D.: Crn: Camera radar net for accurate, robust, efficient 3d perception. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17569–17580 (2023)
22. Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12689–12697 (2019)
23. Li, H., Qu, H.: Dassf: dynamic-attention scale-sequence fusion for aerial object detection. In: International Conference on Computational Visual Media. pp. 212–227. Springer (2025)
24. Lin, Z., Liu, Z., Xia, Z., Wang, X., Wang, Y., Qi, S., Dong, Y., Dong, N., Zhang, L., Zhu, C.: Rcbevdet: Radar-camera fusion in bird's eye view for 3d object detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14928–14937 (2024)
25. Liu, J., Zhao, Q., Xiong, W., Huang, T., Han, Q.L., Zhu, B.: Smurf: Spatial multi-representation fusion for 3d object detection with 4d imaging radar. IEEE Transactions on Intelligent Vehicles **9**(1), 799–812 (2024)
26. Ma, X., Ouyang, W., Simonelli, A., Ricci, E.: 3d object detection from images for autonomous driving: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(5), 3537–3556 (2024)
27. Mao, J., Shi, S., Wang, X., Li, H.: 3d object detection for autonomous driving: A comprehensive survey. International Journal of Computer Vision **131**, 1909–1963 (2023)
28. Musiat, A., Reichardt, L., Schulze, M., Wasenmüller, O.: Radarpillars: Efficient object detection from 4d radar point clouds. In: 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC). pp. 1656–1663. IEEE (2024)
29. Paek, D.H., Kong, S.H., Wijaya, K.T.: K-radar: 4d radar object detection for autonomous driving in various weather conditions. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. pp. 3819–3829 (2022)
30. 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)

31. Peng, X., Sun, H., Bierzynski, K., Fischbacher, A., Servadei, L., Wille, R.: Mutualforce: Mutual-aware enhancement for 4d radar-lidar 3d object detection. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
32. Peng, X., Tang, M., Sun, H., Bierzynski, K., Servadei, L., Wille, R.: Mufasa: Multi-view fusion and adaptation network with spatial awareness for radar object detection. In: Artificial Neural Networks and Machine Learning – ICANN 2024. pp. 168–184. Springer (2024)
33. Peng, X., Tang, M., Sun, H., Kay, B., Servadei, L., Wille, R.: Elmar: Enhancing lidar detection with 4d radar motion awareness and cross-modal uncertainty. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4939–4945. IEEE (2025)
34. Peng, X., Wang, Y., Tang, M., Kay, B., Servadei, L., Wille, R.: Moral: Motion-aware multi-frame 4d radar and lidar fusion for robust 3d object detection. In: 2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC). pp. 2913–2919 (2025)
35. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 5105–5114 (2017)
36. 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: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10526–10535 (2020)
37. 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 (2023)
38. Song, J., Zhao, L., Skinner, K.A.: Lirafusion: Deep adaptive lidar-radar fusion for 3d object detection. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 18250–18257. IEEE (2024)
39. Song, X., Zhou, Z., Zhang, L., Lu, X., Hei, X.: Psns-ssd: Pixel-level suppressed nonsalient semantic and multicoupled channel enhancement attention for 3d object detection. IEEE Robotics and Automation Letters **9**(1), 603–610 (2024)
40. Tan, B., Ma, Z., Zhu, X., Li, S., Zheng, L., Chen, S., Huang, L., Bai, J.: 3-d object detection for multiframe 4-d automotive millimeter-wave radar point cloud. IEEE Sensors Journal **23**(11), 11125–11138 (2023)
41. Wang, L., Zhang, X., Li, J., Xv, B., Fu, R., Chen, H., Yang, L., Jin, D., Zhao, L.: Multi-modal and multi-scale fusion 3d object detection of 4d radar and lidar for autonomous driving. IEEE Transactions on Vehicular Technology **72**(5), 5628–5641 (2023)
42. 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. IEEE (2022)
43. Wu, H., Zhao, S., Huang, X., Wen, C., Li, X., Wang, C.: Commonsense prototype for outdoor unsupervised 3d object detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14968–14977 (2024)
44. Xia, Q., Ye, W., Wu, H., Zhao, S., Xing, L., Huang, X., Deng, J., Li, X., Wen, C., Wang, C.: Hinted: Hard instance enhanced detector with mixed-density feature fusion for sparsely-supervised 3d object detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15321–15330 (2024)

45. Xiong, W., Liu, J., Huang, T., Han, Q.L., Xia, Y., Zhu, B.: Lxl: Lidar excluded lean 3d object detection with 4d imaging radar and camera fusion. IEEE Transactions on Intelligent Vehicles **9**(1), 79–92 (2024)
46. Xiong, W., Zou, Z., Zhao, Q., He, F., Zhu, B.: Lxlv2: Enhanced lidar excluded lean 3d object detection with fusion of 4d radar and camera. IEEE Robotics and Automation Letters **10**(3), 2862–2869 (2025)
47. Xu, R., Xiang, Z.: Rlnet: Adaptive fusion of 4d radar and lidar for 3d object detection. In: Computer Vision – ECCV 2024 Workshops. pp. 181–194. Springer (2025)
48. Xu, R., Xiang, Z., Zhang, C., Zhong, H., Zhao, X., Dang, R., Xu, P., Pu, T., Liu, E.: Sckd: Semi-supervised cross-modality knowledge distillation for 4d radar object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 8933–8941 (2025)
49. Yan, Q., Wang, Y.: Mvfan: Multi-view feature assisted network for 4d radar object detection. In: Neural Information Processing. pp. 493–511. Springer (2024)
50. Yang, P., Wu, F., Liu, M., Zhong, T., Zhou, F.: Beyond pillars: Advancing 3d object detection with salient voxel enhancement of lidar-4d radar fusion. Pattern Recognit. **173**, 112841 (2026)
51. Yang, Z., Sun, Y., Liu, S., Jia, J.: 3dssd: Point-based 3d single stage object detector. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11037–11045 (2020)
52. Yin, T., Zhou, X., Krähenbühl, P.: Center-based 3d object detection and tracking. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11779–11788 (2021)
53. Zhang, H., Wu, K., Chen, R., Wu, Z., Zhong, Y., Li, W.: Tl-4drcf: A two-level 4-d radar–camera fusion method for object detection in adverse weather. IEEE Sensors Journal **24**(10), 16408–16418 (2024)
54. Zhang, Y., Hu, Q., Xu, G., Ma, Y., Wan, J., Guo, Y.: Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18931–18940 (2022)
55. Zhao, H., Guan, R., Wu, T., Man, K.L., Yu, L., Yue, Y.: Unibevfusion: Unified radar-vision bevfusion for 3d object detection. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 6321–6327. IEEE (2025)
56. Zheng, L., Li, S., Tan, B., Yang, L., Chen, S., Huang, L., Bai, J., Zhu, X., Ma, Z.: Rcfusion: Fusing 4-d radar and camera with bird's-eye view features for 3-d object detection. IEEE Transactions on Instrumentation and Measurement **72**, 1–14 (2023)
57. Zheng, L., Liu, J., Guan, R., Yang, L., Lu, S., Li, Y., Bai, X., Bai, J., Ma, Z., Shen, H.L., Zhu, X.: Doracamom: Joint 3d detection and occupancy prediction with multi-view 4d radars and cameras for omnidirectional perception. IEEE Transactions on Circuits and Systems for Video Technology **36**(6), 7630–7645 (2026)
58. Zheng, L., Ma, Z., Zhu, X., Tan, B., Li, S., Long, K., Sun, W., Chen, S., Zhang, L., Wan, M., Huang, L., Bai, J.: Tj4dradset: A 4d radar dataset for autonomous driving. In: 2022 IEEE 25th international conference on intelligent transportation systems (ITSC). pp. 493–498. IEEE (2022)
59. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: International Conference on Learning Representations (2021)