HGSFusion: Radar-Camera Fusion with Hybrid Generation and Synchronization for 3D Object Detection

Zijian Gu[1*], Jianwei Ma[1*], Yan Huang[1†], Honghao Wei[2†], Zhanye Chen[1], Hui Zhang[1†], Wei Hong[1]

[1]Southeast University
[2]Washington State University
{zijian_gu, jianwei_ma, yan_huang, chenzhanye, huizhang, weihong}@seu.edu.cn, honghao.wei@wsu.edu,

Abstract

Millimeter-wave radar plays a vital role in 3D object detection for autonomous driving due to its all-weather and all-lighting-condition capabilities for perception. However, radar point clouds suffer from pronounced sparsity and unavoidable angle estimation errors. To address these limitations, incorporating a camera may partially help mitigate the shortcomings. Nevertheless, the direct fusion of radar and camera data can lead to negative or even opposite effects due to the lack of depth information in images and low-quality image features under adverse lighting conditions. Hence, in this paper, we present the radar-camera fusion network with Hybrid Generation and Synchronization (HGSFusion), designed to better fuse radar potentials and image features for 3D object detection. Specifically, we propose the Radar Hybrid Generation Module (RHGM), which fully considers the Direction-Of-Arrival (DOA) estimation errors in radar signal processing. This module generates denser radar points through different Probability Density Functions (PDFs) with the assistance of semantic information. Meanwhile, we introduce the Dual Sync Module (DSM), comprising spatial sync and modality sync, to enhance image features with radar positional information and facilitate the fusion of distinct characteristics in different modalities. Extensive experiments demonstrate the effectiveness of our approach, outperforming the state-of-the-art methods in the VoD and TJ4DRadSet datasets by 6.53% and 2.03% in RoI AP and BEV AP, respectively.

Code — https://github.com/garfield-cpp/HGSFusion

Introduction

3D object detection is a critical task in autonomous driving, focusing on accurately determining the location, dimensions, and orientation of surrounding objects (Mao et al. 2023; Ma et al. 2023; Ghasemieh and Kashef 2022; Aung et al. 2024). Multiple sensors, such as camera, radar, and LiDAR, have been widely used for object detection with distinct data structures and properties. To achieve accurate and effective object detection, both semantic information, provided by the camera, and positional information, offered by radar or LiDAR, are crucial (Wu et al. 2024).

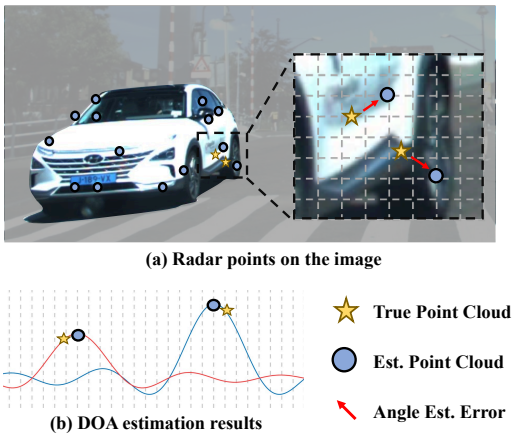

Figure 1: Illustration of angle estimation errors in obtaining radar point clouds. (a) True points and estimated points are shown in the image. (b) True points and estimated points are shown in the radar DOA estimation. The estimated points fall on the beamforming peaks, deviating from the true points.

Initially, camera-based methods were used for object detection, and they are still a hot topic in recent years (Reading et al. 2021; Li et al. 2023; Huang et al. 2021; Philion and Fidler 2020). The semantic information in images facilitates the differentiation of object categories and the identification of small targets (Alaba and Ball 2022; Li and Qu 2024). However, the lack of depth information in images makes it challenging to accurately localize objects with images alone (Hu et al. 2023). Moreover, adverse weather conditions can easily affect the performance of cameras (Bhadoriya, Vegamoor, and Rathinam 2021), reducing the robustness of the detection system. Hence, how to leverage the rich semantic information in images while compensating for the deficiencies in depth and robustness has become an urgent issue. Positional information can be provided by either radar (Tan et al. 2022a; Yan and Wang 2023) or LiDAR (Zhang et al. 2022; Huang et al. 2024). Radar systems, in particular, offer additional velocity and enhanced robustness in adverse weather conditions at a lower cost (Kim et al. 2023b). However, compared with LiDAR, radar point clouds exhibit more pronounced sparsity degrading the detection performance, yet potential solutions for this issue are quite limited. Meth-

*These authors contributed equally.
†Corresponding authors.

ods designed to handle the sparsity of LiDAR points (Yin, Zhou, and Krähenbühl 2021) fail to achieve optimal performance when directly applied to radar points. Moreover, the conventional radar signal processing to obtain radar point clouds involves applying the Constant False Alarm Rate (CFAR) algorithm to radar echo signals and then performing angle estimation of the detected target through CFAR. As shown in Figure 1(b), the beamforming peak of the DOA estimation is the estimated angle of radar points, deviating from true radar points. And in Figure 1(a), this deviation is projected on the image, where the estimation error of radar points may degrade the detection performance.

The sparsity of radar point clouds can result in only a few points on the target, and angle estimation errors can cause the point cloud to be distributed in incorrect locations. Both factors significantly degrade the detection performance of radar-based methods.

To further improve detection performance, an increasing number of studies focus on leveraging complementary information from different modalities through fusion approaches. Although a straightforward concatenation of features from various modalities can yield some improvement (Liu et al. 2023), challenges arise due to the limited angle resolution of radar and the absence of depth information in images, leading to feature misplacement. Therefore, developing effective strategies for feature fusion across modalities and mitigating the misalignment of features have emerged as critical issues that require immediate attention.

In this paper, we introduce a radar-camera fusion network named HGSFusion (**H**ybrid **G**eneration and **S**ynchronization), designed to fully leverage the potential of radar and facilitate the integration of camera and radar data for 3D object detection. In particular, the proposed Radar Hybrid Generation Module (RHGM) generates denser radar points with estimated points falling into masks, also known as foreground points. During the generation process, different probability distributions are employed to mitigate the impact of angle errors brought by DOA estimation. Subsequently, features from both image and radar are extracted by separate backbones and transformed into one unified Bird's Eye View (BEV) space. Then, the Dual Sync Module (DSM) utilizes spatial sync to enhance image features with position information in radar features and modality sync to alleviate the influences of image features under adverse lighting conditions. Extensive experiments conducted on VoD and TJ4DRadSet datasets achieve state-of-the-art (SOTA) performance, verifying the effectiveness and robustness of the proposed hybrid generation and Dual Sync.

The main contributions of our work are listed as follows

- We propose a novel radar-camera fusion network HGSFusion to boost the fusion of radar points and images.
- Radar Hybrid Generation Module (RHGM) leverages the distribution of point clouds derived from the radar point cloud imaging process to generate denser and higher-quality radar point clouds.
- Dual Sync Module (DSM) guides 3D image features with positional information from radar and utilizes complementary information to produce fused BEV features.
- Extensive experiments on the VoD and TJ4DRadSet datasets demonstrate the effectiveness of the network and each component, outperforming state-of-the-art View of Delft (VoD) and TJ4DRadSet datasets by 6.53% and 2.03% in RoI AP and BEV AP, respectively.

Related Works

Single-Modality 3D Object Detection

Existing camera-based detection methods typically require transforming the image features from Perspective View (PV) to BEV to ensure consistency between the input feature space and the output space. The transformation can be categorized into splatting and sampling. Splatting methods (Philion and Fidler 2020) project each pixel of the image to 3D space along the corresponding 3D rays and place image features to voxels passed by 3D rays. Sampling methods (Harley et al. 2023) project the center of voxels to images, and then sample the voxel features based on the positions they fall on the image features.

On the other hand, both radar and LiDAR can provide input for point-based object detection. Several previous works (Li, Luo, and Yang 2023; Meng et al. 2023; Hu, Kuai, and Waslander 2022; Li, Wang, and Wang 2021) convert the LiDAR point cloud into voxels to realize regular shapes. Then, feature extraction is usually conducted on these regular tensors. Unlike LiDAR, conventional automotive radar provides additional physical information, such as velocity and Radar Cross Section (RCS), but with sparser points and lower angle resolution, making it challenging to perform object detection on radar alone (Dreher et al. 2020; Popov et al. 2023; Ulrich et al. 2022). The emergence of 4D imaging radar eases these issues with more radar points and elevation angle (Dong et al. 2020; Liu et al. 2024a; Köhler et al. 2023). RadarMFNet (Tan et al. 2022b) conducts 3D object detection using a multi-frame 4-D radar point cloud to handle the sparsity in radar point clouds and shows that incorporating temporal and spatial features can improve detection capabilities. Moreover, in RPFA-Net (Xu et al. 2021) pillar-based design is employed to alleviate the influence of error in elevation angle. In addition to point cloud-based radar detection methods, recently, methods based on raw radar echo signals have also received more attention (Liu et al. 2024b; Paek, Kong, and Wijaya 2022; Rebut et al. 2022).

3D Object Detection with Multi-Modality Fusion

Recent advancements in 3D object detection focus on fusing image-based and point-based sensors to enhance system robustness and accuracy (Jiao et al. 2023; Zhang et al. 2024a; Yan et al. 2023; Yang et al. 2022). Notably, BEVFusion (Liu et al. 2023) introduces a technique that builds detection schemes for image and LiDAR in a unified BEV space, improving robustness in scenarios. The advancement of radar enables it as a key point-based sensor in autonomous driving (Stäcker et al. 2022; Kim et al. 2023a). FUTR3D (Chen et al. 2023) employs transformer-based query mechanisms to integrate features from camera, radar, and LiDAR in autonomous driving, presenting a robust fusion approach.

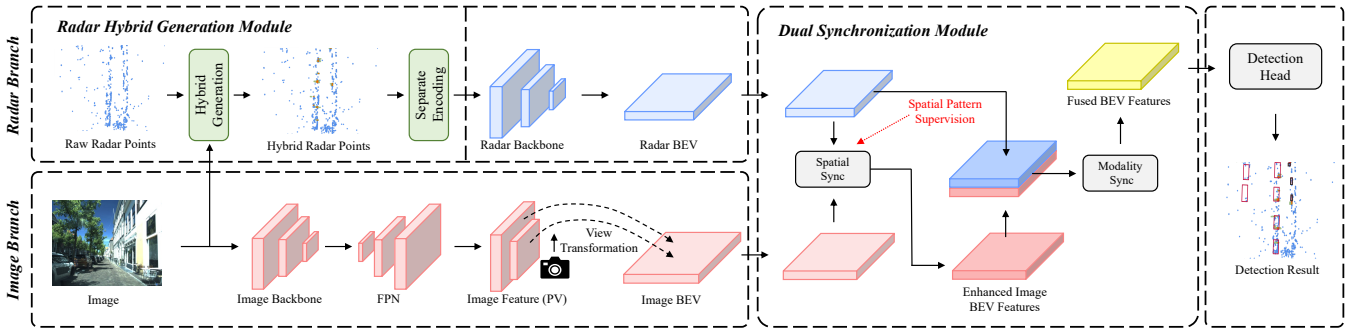

Figure 2: Overall framework of the proposed HGSFusion. In the radar branch, the RHGM utilizes raw radar points and images to generate hybrid radar points (generated points, foreground points, and raw radar points shown in green, orange, and blue points, respectively). Then the hybrid radar points are encoded and passed through the radar backbone to produce radar BEV features. In the image branch, images are processed through image backbone and view transformation, producing image BEV features. Subsequently in DSM, the image and radar features undergo dual sync to obtain fused BEV features for object detection.

The advent of 4D millimeter-wave radar with elevation angle makes it feasible to select only radar and camera as sensors. Particularly designed for 4D millimeter-wave radar with elevation angle, RCFusion (Zheng et al. 2023) develops a novel radar point cloud extraction backbone and implements interactive attention mechanisms to efficiently fuse radar features. In contrast, LXL (Xiong et al. 2023) and CRN (Kim et al. 2023b) refine the process of transforming 2D image features into 3D space by utilizing depth predictions and radar point clouds. Meanwhile, RCBEVDet (Lin et al. 2024) further exploits radar with an RCS-aware BEV encoder and multi-layer cross-attention fusion. The robustness of radar is verified in TL4DRCF (Zhang et al. 2024b) by conducting experiments in the VoD-Fog dataset. In this paper, we will explore how to leverage the position information in radar to enhance image features and improve robustness under adverse lighting conditions for radar-camera fusion detection.

Method

In this section, we first introduce the overall architecture of HGSFusion. Then, we present the details of the proposed RHGM and DSM, illustrating how the RHGM generates denser and more accurate radar points, and how the DSM facilitates effective fusion between radar and camera features.

Overall Architecture

The overall architecture of HGSFusion is shown in Figure 2. In the radar branch, the RHGM utilizes raw radar points and images to obtain foreground points and generate denser radar points. These hybrid points (generated points, foreground points, and raw radar points) are encoded and sent to the radar backbone to generate radar BEV features and spatial patterns. In the image branch, corresponding monocular images are passed through the image backbone to obtain multi-scale image features for subsequent 2D-to-3D view transformation and height compression, yielding image BEV features. The image and radar BEV features are then fused in DSM before being fed into the detection head.

Radar Hybrid Generation Module (RHGM)

Point Cloud Generation. Point cloud generation primarily involves three steps: obtaining foreground points, acquiring probability distribution, and generating hybrid points. The overall process is illustrated in Figure 3.

1) Obtaining foreground points. With the radar-camera transformation matrix and the camera intrinsic matrix, raw radar points are projected onto the corresponding images (Yin, Zhou, and Krähenbühl 2021). The i-th raw point can be expressed as $\mathcal{P}_{\text{raw},i} = [u_i, v_i, d_i, \mathbf{f}_i]$, where u_i and v_i are the pixel coordinates in the image, d_i is the depth, and $\mathbf{f}_i$ represents other physical features such as RCS and velocity. Next, corresponding image instance masks are predicted via a semantic segmentation network. Projected points that fall within these masks are identified as foreground points. Similar to raw points, the i-th foreground point is represented as $\mathcal{P}_{\text{fore},i} = [u_i, v_i, d_i, \mathbf{f}_i, \mathbf{s}_i]$, where $\mathbf{s}_i$ is a one-hot semantic feature indicating class labels after semantic segmentation.

2) Acquiring probability distribution. Next, the key challenge is to generate denser point clouds with higher quality based on the distribution of these foreground points. We overcome the difficulties by considering point generation as a sampling processing, where the probability distribution is characterized by the Probability Density Function (PDF) of the generated points in the given region. A straightforward method is to set a uniform distribution as the PDF of generated points within each mask, formulated as

$$f_U(u, v) = \frac{1}{A}, \tag{1}$$

where A is the area of the uniform distribution. This method can leverage image information to increase the number of points, which may yield good performance for LiDAR, as LiDAR point clouds typically follow a consistent pattern, especially in mechanical scanning systems. However, radar point clouds exhibit different distributions, since they are derived from CFAR detection and DOA estimation, which inherently include estimation errors. Consequently, uniform generation may not be a good choice for radar points.

Due to the fact that radar points are more likely to be distributed near foreground points, regions with and without

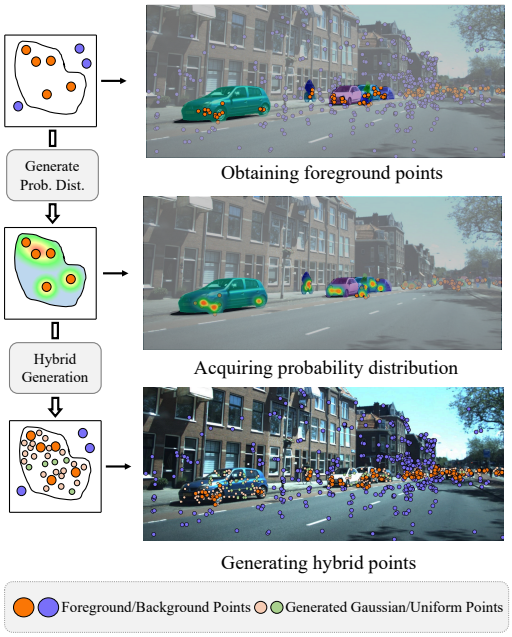

Figure 3: Point cloud generation in RHGM. Initially, raw radar points are projected onto the image, and points falling inside the mask are selected as foreground points. Subsequently, these foreground points are used to produce a generation probability distribution. Finally, the probability distribution is utilized to create the hybrid radar points composed of raw radar points (points in/out mask), foreground points, and generated Gaussian/uniform points.

nearby foreground radar points should be considered separately. Specifically, the areas centered around foreground points (u_i, v_i) with a radius of r pixels are referred to as regions with nearby foreground points, defined as

$$R_i(u,v) = \{(u,v) \in R_m | (u-u_i)^2 + (v-v_i)^2 < r^2\}, \tag{2}$$

where R_m is the region of instance masks. Then the areas out of these regions are considered to have no nearby points.

For the area near foreground points, the PDF of generated points should satisfy two properties: i) The generation probability near foreground points should be higher than areas without foreground points nearby. ii) The probability increases monotonically with the decreased distance from the foreground points. In our method, the generation probability distribution of the regions near foreground points (u_i, v_i) is modeled by the Gaussian distribution as

$$f_G(u,v) = \frac{1}{2\pi b_1 b_2} \exp\left[-\frac{1}{2}\left(\frac{(u-u_i)^2}{b_1^2} + \frac{(v-v_i)^2}{b_2^2}\right)\right], \tag{3}$$

where b_i is the standard deviation. In the regions without foreground points nearby, radar points can be generated via uniform distribution $f_U(u,v)$ due to the inexistence of prior information.

3) Generating hybrid points. Then the generation probability distribution for the entire region can be formulated as

$$f_H(u,v) = \begin{cases} f_G(u,v) & (u,v) \in R_i(u,v), \\ f_U(u,v) & (u,v) \in \complement_{R_m} R_i(u,v), \\ 0 & (u,v) \notin R_m, \end{cases} \tag{4}$$

where $\complement_{R_m} R_i(u,v)$ is the complementary regions of $R_i(u,v)$, i.e., the regions without foreground points nearby.

The generation probability distribution $f_H(u,v)$ is used to generate points $\mathcal{G}_i = [u_i, v_i]$ that lie within the image. To obtain the depth and features of these generated points, we calculate the distances from each generated point $\mathcal{G}$ to the foreground points $\mathcal{P}_{\text{fore}}$. The depth and feature of the nearest foreground point are then assigned to each corresponding generated point, resulting in $\mathcal{G}_i = [u_i, v_i, d_i, \mathbf{f}_i, \mathbf{s}_i]$. To enable these generated points to serve as input for the network, they need to be projected back into the radar coordinate system using the camera intrinsic matrix and the camera-radar transformation matrix, resulting in the generated points in radar coordinates $\mathcal{G}_i = [x_i, y_i, z_i, \mathbf{f}_i, \mathbf{s}_i]$, where x_i, y_i, and z_i are the coordinates in radar coordinate system.

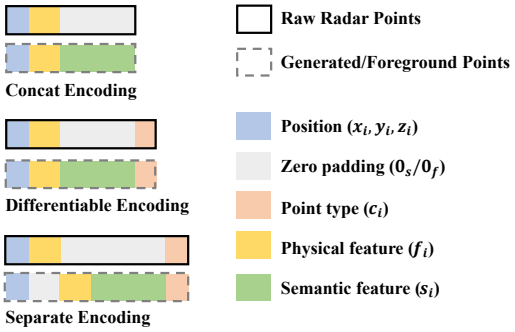

Figure 4: Different encoding strategies of RHGM. Generated and foreground points share the same encoding scheme.

Separate Radar Point Encoding. To retain as much information as possible from the point clouds, the generated radar points $\mathcal{G}$, raw points $\mathcal{P}_{raw}$, and foreground points $\mathcal{P}_{fore}$ are all used as the input. However, the lack of semantic features in $\mathcal{P}_{raw}$ results in an inconsistency in feature length and incompatibility for direct network input. Although it is possible to use two separate radar backbones for distinct feature extraction, it would introduce additional computational costs and risks of overfitting. Therefore, it is necessary to encode radar points with equal-length features before inputting them into the network.

Raw radar points $\mathcal{P}_{raw}$ only contain positions and radar physical features while generated radar points $\mathcal{G}$ and radar foreground points $\mathcal{P}_{fore}$ encompass the additional semantic feature. One simple encoding strategy, namely Concat Encoding, to align these features is to pad zeros at the end of raw point features, formulated as $[x_i, y_i, z_i, \mathbf{f}_i, \mathbf{0_s}]$, where $\mathbf{0_s}$ is the zero padding, with the generated points and foreground points invariant. Another enhanced encoding strategy, namely Differentiable Encoding, is formulated as $[x_i, y_i, z_i, \mathbf{f}_i, \mathbf{0_s}, \mathbf{c}_i]$ and $[x_i, y_i, z_i, \mathbf{f}_i, \mathbf{s}_i, \mathbf{c}_i]$, where $\mathbf{c}_i$ is the point type using one-hot encoding to distinguish different points. Generated points and foreground points share the same encoding but with different point types.

However, both Concat Encoding and Differentiable Encoding may be limited in representation, since pillar-based (Shi, Li, and Ma 2022; Lang et al. 2019) detectors mix points through average operation in each pillar, making it difficult to distinguish different point types when features are placed at the same position. Herein, we place the physical and semantic features of points in different places to help distinguish points, defined as distributed features. Then, for the points, which lack the corresponding features, zero padding is employed to ensure that they share the same length. The entire process is referred to as Separate Encoding, and it can enhance the distinction between different types of points and shield them from interfering with the features of other points. Specifically, the proposed encoding of raw points $\mathcal{P}_{raw}$ can be represented as $[x_i, y_i, z_i, \mathbf{f}_i, \mathbf{0_f}, \mathbf{0_s}, \mathbf{c}_i]$. Similarly, the encoding of $\mathcal{G}$ and $\mathcal{P}_{fore}$ can be represented as $[x_i, y_i, z_i, \mathbf{0_f}, \mathbf{f}_i, \mathbf{s}_i, \mathbf{c}_i]$ with different point types. Finally, encoded radar points are concatenated, pillarized, and fed into the radar backbone, yielding radar BEV features $F_R \in \mathbb{R}^{C\times X\times Y}$, where C is the number of channels, and X and Y denote the dimensions of BEV feature map, respectively. All the above encoding strategies are illustrated in Figure 4.

Dual Sync Module

The lack of depth information in images and the low-quality features under adverse lighting conditions present significant challenges for 3D object detection. In this subsection, we will introduce the DSM comprising spatial sync and modality sync to address these issues.

Spatial Sync. Radar point clouds encompass spatial information that is absent in images, allowing for the enhancement of image features by using radar. In spatial sync, radar features are utilized to explicitly predict the probability of object presence at various spatial locations, referred to as spatial patterns. Notably, we incorporate the atrous convolution to enlarge the receptive field, since objects, which are relatively large compared to the pillar size, may span a large region of the feature map. The entirety of the spatial pattern prediction $S_R \in \mathbb{R}^{1\times X\times Y}$ can be formulated as

$$S_R = \sigma\left(\mathrm{Conv}\left(\mathrm{AtrousConv}\left(F_R\right)\right)\right), \tag{5}$$

where σ is the sigmoid activation function. The radar spatial pattern is supervised by focal loss with ground truth generated by bounding boxes. Then the radar spatial pattern is multiplied with image BEV features $F_I \in \mathbb{R}^{C\times X\times Y}$. The enhanced image BEV features F_I can be formulated as

$$F_I' = S_R' \otimes F_I, \tag{6}$$

where S_R' is the spatial pattern broadcasted along the channel dimension and $\otimes$ is the element-wise multiplication.

Modality Sync. The enhanced image features and radar features are in two separate modalities and need to be fused. It is observed that in adverse lighting conditions such as darkness or shiny lighting, the quality of image features is significantly degraded. In contrast, radar features are less affected by lighting conditions. To leverage the distinct characteristics of different modalities, modality sync is employed to tackle this issue by predicting the importance of different modalities. In modality sync, the radar and image BEV features are first concated and fused with convolution layers, formulated as

$$F_{concat} = \mathrm{Conv}\left(F_R \copyright F_I'\right), \tag{7}$$

where $\copyright$ is the concatenation operation along channel dimension. Then, the feature weights $\mathcal{V} \in \mathbb{R}^{2C}$ measuring the varying importance of the feature map are predicted from the concated features $F_{concat} \in \mathbb{R}^{2C\times X\times Y}$, formulated as

$$\mathcal{V} = \sigma\left(\mathrm{Conv}\left(\mathrm{AvgPooling}\left(F_{concat}\right)\right)\right). \tag{8}$$

After the whole Modality Sync process, the final fused BEV feature map can be formulated as

$$F = \mathcal{V}' \otimes F_{concat}, \tag{9}$$

where $\mathcal{V}'$ is the feature weights broadcasted along the spatial dimensions of feature maps. Finally, the fused BEV features F are used for the downstream 3D object detection.

Experiments

Dataset and Metrics

In our study, we conduct experiments on 4D millimeter wave radar datasets, VoD dataset (Palffy et al. 2022) and TJ4DRadSet dataset (Zheng et al. 2022). We adopt the official split schemes of the datasets. For the VoD dataset, the official evaluation metrics are AP in Entire Annotated Area AP (EAA AP) and AP in the Driving Corridor (RoI AP). They are conducted in the Entire Annotated Area and the Driving Corridor area ranging $(-4\mathrm{m} < x < 4\mathrm{m}, z < 25\mathrm{m})$ in camera coordinates. IoU thresholds are set to 0.5, 0.25, and 0.25 for cars, pedestrians, and bicycles, respectively. The TJ4DRadSet dataset includes AP in both 3D and BEV space. Evaluation is limited to targets within 70 meters away from the sensor. IoU thresholds for car, pedestrian, and cyclist are the same as those of the VoD dataset. For the truck category, the IoU threshold is set to 0.5.

Implementation Details

ResNet-101 is employed as the image backbone with pretrained weight from DeepLabV3 and frozen to prevent overfitting. Mask2former (Cheng et al. 2022) is utilized as the segmentation network, and Radar PillarNet (Zheng et al. 2023) is utilized as the radar backbone. The radius r is set to 51, and in each mask, 250 enhanced point clouds are generated, where 50 via Gaussian generation and 200 via uniform generation. The detection head adopts an anchor-based approach. Horizontal flipping, global rotation, and global scaling are applied as data augmentation during training. We use AdamW as the optimizer and train the proposed network for 25 epochs with a learning rate of 0.001 and batch size of 4.

SOTA Comparison

VoD Validation Set. Our method is tested on the validation set of the VoD dataset, with the results presented in Table 1. The EAA AP and RoI AP achieve the best performance, surpassing the SOTA LXL by 2.65% and 6.53%, respectively. In terms of Car and Pedestrian categories, our

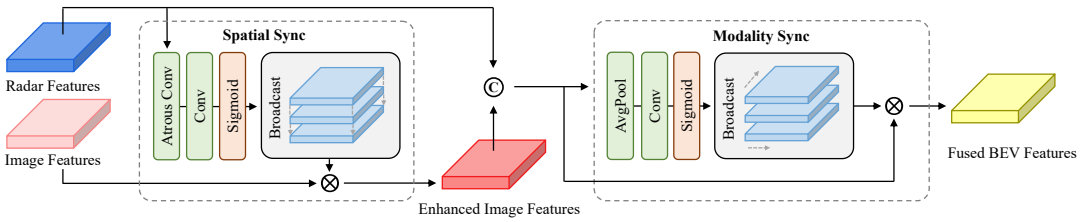

Figure 5: Internal structure of DSM. In Spatial Sync, radar features enhance image features with position information in radar features. Then the enhanced image features and radar features undergo Modality Sync, resulting in the fused BEV features.

Method	Modality	Entire Annotated Area				In Driving Corridor			
		Car	Pedestrian	Cyclist	mAP	Car	Pedestrian	Cyclist	mAP
PointPillars (Lang et al. 2019)	R	37.06	35.04	63.44	45.18	70.15	47.22	85.07	67.48
RadarPillarNet (Zheng et al. 2023)	R	39.30	35.10	63.63	46.01	71.65	42.80	83.14	65.86
FUTR3D (Chen et al. 2023)	R+C	46.01	35.11	65.98	49.03	78.66	43.10	86.19	69.32
BEVFusion (Liu et al. 2023)	R+C	37.85	40.96	68.95	49.25	70.21	45.86	**89.48**	68.52
RCFusion (Zheng et al. 2023)	R+C	41.70	38.95	68.31	49.65	71.87	47.50	88.33	69.23
LXL (Xiong et al. 2023)	R+C	42.33	49.48	**77.12**	56.31	72.18	58.30	88.31	72.93
TL-4DRCF (Zhang et al. 2024b)	R+C	43.71	40.11	64.22	49.35	79.49	53.76	76.50	69.92
RCBEVDet (Lin et al. 2024)	R+C	40.63	38.86	70.48	49.99	72.48	49.89	87.01	69.80
HGSFusion(Ours)	R+C	**51.67**	**52.64**	72.58	**58.96**	**88.28**	**62.61**	87.49	**79.46**

Table 1: Performance comparison on validation set of VoD dataset.

method achieves the best performance. Especially for the Car category, the proposed HGSFusion can greatly densify radar points and promote fusion between radar and camera, yielding an improvement of 5.66% and 8.79% compared with FUTR3D and TL-4DRCF. However, a performance decline is observed in the Cyclist category. This decline is due to the presence of various bicycle-like objects in the VoD dataset scenes, such as parked bicycles, bicycle racks, and scooters. These objects are difficult to distinguish, affecting the quality of the generated radar point clouds and consequently leading to a drop in performance.

TJ4DRadSet Test Set. To validate the generalization capability of the proposed model, we also conduct experiments on the TJ4DRadSet dataset, with the results presented in Table 2. The model surpasses the SOTA LXL by 0.89% in 3D mAP and by 2.03% in BEV mAP. These improvements indicate that the model effectively integrates images to generate denser radar point clouds and effectively fuse features of different modalities.

Method	Modality	3D(%)	BEV(%)
RPFA-Net (2021)	R	29.91	38.94
RadarPillarNet (2023)	R	30.37	39.24
SMURF(2024a)	R	32.99	40.98
FUTR3D (2023)	R+C	32.42	37.51
BEVFusion (2023)	R+C	32.71	41.12
RCFusion (2023)	R+C	33.85	39.76
LXL (2023)	R+C	36.32	41.20
HGSFusion(Ours)	R+C	**37.21**	**43.23**

Table 2: Performance comparison on test set of TJ4DRadSet dataset.

Comprehensive Analysis

ID	Modality	RHGM	DSM	EAA AP	RoI AP
1	R			47.70	66.88
2	C			22.40	42.74
3	R+C			54.82	73.27
4	R+C	✓		57.23	74.83
5	R+C		✓	55.99	74.45
6	R+C	✓	✓	**58.96**	**79.46**

Table 3: Ablation study results of proposed components on the validation set of VoD dataset.

Ablation of Proposed Components. Ablation experiments are performed on the VoD validation set to evaluate the impact of different modalities and the proposed modules, with the results presented in Table 3. As can be seen, the direct fusion of features from both modalities (#3) yields promising results compared to single modality (#1-2), indicating the existence of complementary information between images and radar points. In addition, the separated introduction of the RHGM (#4) and DSM (#5) improves network performance with 2.41% and 1.17% in EAA AP and 1.56% and 1.18% in RoI AP, respectively, Hence, both RHGM and DSM can help the network boost detection performance. The complete HGSFusion (#6), which utilizes RHGM and DSM, achieves the best performance, outperforming the baseline by 4.14% and 6.19% in EAA AP and RoI AP, respectively. This stems from the fact that the hybrid radar points incorporate semantic information from images, improving the quality of radar features. Additionally, the DSM leverages position information from radar to enhance image features while alleviating the impact of low-quality image features under adverse lighting conditions.

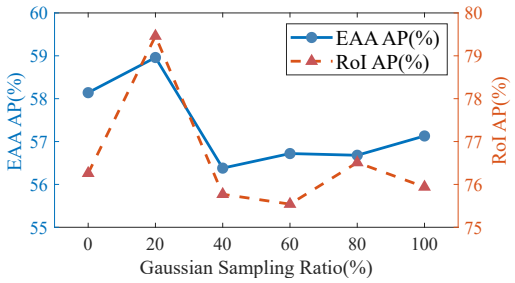

Figure 6: Performance of different generation scheme.

Comparisons between Radar Point Generation Schemes. Herein, we investigate the impacts of different generation schemes on detection performance by fixing other parameters and adjusting the ratio of Gaussian generation points to the total points. The results are presented in Figure 6. As can be observed, using a purely uniform generation method does not yield the best performance, lagging behind the proposed hybrid scheme by 0.82% and 3.20% in EAA AP and RoI AP, respectively. This is because uniform generation only brings segmentation information into the generated radar points without considering the angle estimation errors introduce by the DOA estimation algorithm. However, adopting a pure Gaussian generation also fails to achieve optimal performance, falling behind the hybrid scheme by 1.83% and 3.52% in EAA AP and RoI AP, respectively. This may arise from that pure Gaussian generation makes the generated points distributed near foreground points, yielding almost no points in the area without foreground points. As a result, the hybrid generation approach combining uniform and Gaussian generation effectively mitigates these shortcomings and achieves the best performance.

Discussion on Radar Point Cloud Encoding. Similar to the raw radar points, the enhanced radar points possess both positions and physical features. These points are encoded and fed into the radar backbone. Comprehensive experiments are conducted to investigate the impact of encoding strategy on overall performance. The encoding strategies are visualized in Figure 4 and their corresponding experimental results are shown in Table 4. One simple encoding strategy, referred to as Concat Encoding, is to mix these points together indiscriminately. The achieved performance improvement is attributed to semantic information contained in generated points brought by the proposed RHGM. As aforementioned, the Differentiable Encoding that incorporates one-hot encoding, which is point type, can better distinguish these points. As can be seen, HGSFusion with Differentiable Encoding achieves higher performance improvement by 0.56% and 2.47% in EAA AP and RoI AP, respectively. Look into the process of radar feature extraction, and it can be observed that points within the same pillar are grouped together, limiting the discriminative ability when features are placed at the same location. Hence, the adoption of the distributed features and zero-padding in Separate Encoding outperforms the baseline by 2.97% and 5.01% in EAA AP and RoI AP, respectively.

H.P.	Encoding Strategy	EAA AP	RoI AP
		55.99	74.45
✓	C.E.	56.22	74.46
✓	D.E.	56.55	76.92
✓	S.E. (Ours)	**58.96**	**79.46**

Table 4: Performance of point encoding strategies. H.P., C.E., D.E., and S.E. are abbreviations for Hybrid Points, Concat Encoding, Differentiable Encoding, and Separate Encoding, respectively.

	3D mAP(%)			BEV mAP(%)		
Method	Dark	Normal	Shiny	Dark	Normal	Shiny
Base-R	13.19	12.97	18.94	19.61	16.18	28.25
Base-R+C	4.27	33.50	22.37	8.42	38.93	28.70
HGSFusion	**15.68**	**35.82**	**25.28**	**19.73**	**42.05**	**31.83**

Table 5: Performances under different lighting conditions on the test set of TJ4DRadSet dataset.

Influences of Lighting Conditions. By considering the varying lighting conditions across different sequences in TJ4DRadSet, we divide the whole dataset into three subsets: dark, normal, and shiny. Then, we evaluate our proposed HGSFusion on the subsets, as well as two baseline networks, Base-R and Base-R+C (excluding RHGM and DSM). Base-R uses only raw radar point clouds as input, while Base-R+C uses both raw radar point clouds and the image. The results are presented in Table 5. As listed in Table 5, the fusion network outperforms the baseline network for all lighting scenarios. In "Dark" scenes, the information captured by the camera is limited and may even contain errors. Hence, the performance can degrade when the camera input is incorporated. However, our proposed HGSFusion network can leverage radar features to enhance image features and achieve performance improvement by 11.41% and 11.31% in 3D mAP and BEV mAP, respectively. Conversely, in the "Normal" and "Shiny" conditions, the image contains more information, leading to performance improvement when incorporated. Our proposed HGSFusion network can utilize images to generate denser radar point clouds, further boosting performance up to 2.91% and 3.13% in 3D mAP and BEV mAP, respectively. The improvements demonstrate the robustness of our fusion network in all lighting conditions.

Conclusion

In this paper, we propose HGSFusion, a pioneering network that fuses 4D imaging radar and images to enhance 3D object detection. The sparsity of radar points and angle estimation errors are mitigated by innovatively using RHGM hybrid generation that considers DOA estimation errors. In DSM, Spatial Sync leverages the position information from radar to enhance the image features, compensating for lack of depth in an image. Moreover, DSM also employs Modality Sync to measure the importance of different features and thus reduce the impact of low-quality image features under adverse lighting. Extensive experimental results demonstrate that HGSFusion achieves state-of-the-art performance in prevalent VoD and TJ4DRadSet datasets.

References

Alaba, S. Y.; and Ball, J. E. 2022. A survey on deep-learning-based lidar 3d object detection for autonomous driving. *Sensors*, 22(24): 9577.

Aung, N. H. H.; Sangwongngam, P.; Jintamethasawat, R.; Shah, S.; and Wuttisittikulkij, L. 2024. A Review of LiDAR-based 3D Object Detection via Deep Learning Approaches towards Robust Connected and Autonomous Vehicles. *IEEE Transactions on Intelligent Vehicles*.

Bhadoriya, A. S.; Vegamoor, V. K.; and Rathinam, S. 2021. Object detection and tracking for autonomous vehicles in adverse weather conditions. Technical report, SAE Technical Paper.

Chen, X.; Zhang, T.; Wang, Y.; Wang, Y.; and Zhao, H. 2023. Futr3d: A unified sensor fusion framework for 3d detection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 172–181.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Dong, X.; Wang, P.; Zhang, P.; and Liu, L. 2020. Probabilistic oriented object detection in automotive radar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 102–103.

Dreher, M.; Erçelik, E.; Bänziger, T.; and Knoll, A. 2020. Radar-based 2D car detection using deep neural networks. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 1–8. IEEE.

Ghasemieh, A.; and Kashef, R. 2022. 3D object detection for autonomous driving: Methods, models, sensors, data, and challenges. *Transportation Engineering*, 8: 100115.

Harley, A. W.; Fang, Z.; Li, J.; Ambrus, R.; and Fragkiadaki, K. 2023. Simple-bev: What really matters for multi-sensor bev perception? In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2759–2765. IEEE.

Hu, H.; Wang, F.; Su, J.; Wang, Y.; Hu, L.; Fang, W.; Xu, J.; and Zhang, Z. 2023. EA-LSS: Edge-aware Lift-splat-shot Framework for 3D BEV Object Detection. *arXiv preprint arXiv:2303.17895*, 2.

Hu, J. S.; Kuai, T.; and Waslander, S. L. 2022. Point density-aware voxels for lidar 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8469–8478.

Huang, J.; Huang, G.; Zhu, Z.; Ye, Y.; and Du, D. 2021. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*.

Huang, K.-C.; Lyu, W.; Yang, M.-H.; and Tsai, Y.-H. 2024. PTT: Point-Trajectory Transformer for Efficient Temporal 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14938–14947.

Jiao, Y.; Jie, Z.; Chen, S.; Chen, J.; Ma, L.; and Jiang, Y.-G. 2023. Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 21643–21652.

Kim, Y.; Kim, S.; Choi, J. W.; and Kum, D. 2023a. Craft: Camera-radar 3d object detection with spatio-contextual fusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1160–1168.

Kim, Y.; Shin, J.; Kim, S.; Lee, I.-J.; Choi, J. W.; and Kum, D. 2023b. Crn: Camera radar net for accurate, robust, efficient 3d perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17615–17626.

Köhler, D.; Quach, M.; Ulrich, M.; Meinl, F.; Bischoff, B.; and Blume, H. 2023. Improved multi-scale grid rendering of point clouds for radar object detection networks. In *2023 26th International Conference on Information Fusion (FUSION)*, 1–8. IEEE.

Lang, A. H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; and Beijbom, O. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12697–12705.

Li, H.; and Qu, H. 2024. DASSF: Dynamic-Attention Scale-Sequence Fusion for Aerial Object Detection. *arXiv preprint arXiv:2406.12285*.

Li, J.; Luo, C.; and Yang, X. 2023. 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*, 17567–17576.

Li, Y.; Ge, Z.; Yu, G.; Yang, J.; Wang, Z.; Shi, Y.; Sun, J.; and Li, Z. 2023. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1477–1485.

Li, Z.; Wang, F.; and Wang, N. 2021. Lidar r-cnn: An efficient and universal 3d object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7546–7555.

Lin, Z.; Liu, Z.; Xia, Z.; Wang, X.; Wang, Y.; Qi, S.; Dong, Y.; Dong, N.; Zhang, L.; and Zhu, C. 2024. RCBEVDet: Radar-camera Fusion in Bird's Eye View for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14928–14937.

Liu, J.; Zhao, Q.; Xiong, W.; Huang, T.; Han, Q.-L.; and Zhu, B. 2024a. SMURF: Spatial multi-representation fusion for 3D object detection with 4D imaging radar. *IEEE Transactions on Intelligent Vehicles*.

Liu, Y.; Wang, F.; Wang, N.; and ZHANG, Z.-X. 2024b. Echoes beyond points: Unleashing the power of raw radar data in multi-modality fusion. *Advances in Neural Information Processing Systems*, 36.

Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D. L.; and Han, S. 2023. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, 2774–2781. IEEE.

Ma, X.; Ouyang, W.; Simonelli, A.; and Ricci, E. 2023. 3d object detection from images for autonomous driving: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Mao, J.; Shi, S.; Wang, X.; and Li, H. 2023. 3D object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 131(8): 1909–1963.

Meng, Z.; Xia, X.; Xu, R.; Liu, W.; and Ma, J. 2023. HYDRO-3D: Hybrid object detection and tracking for cooperative perception using 3D LiDAR. *IEEE Transactions on Intelligent Vehicles*, 8(8): 4069–4080.

Paek, D.-H.; Kong, S.-H.; and Wijaya, K. T. 2022. K-radar: 4d radar object detection for autonomous driving in various weather conditions. *Advances in Neural Information Processing Systems*, 35: 3819–3829.

Palffy, A.; Pool, E.; Baratam, S.; Kooij, J. F. P.; and Gavrila, D. M. 2022. Multi-Class Road User Detection With 3+1D Radar in the View-of-Delft Dataset. *IEEE Robotics and Automation Letters*, 7(2): 4961–4968.

Philion, J.; and Fidler, S. 2020. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 194–210. Springer.

Popov, A.; Gebhardt, P.; Chen, K.; and Oldja, R. 2023. Nvradarnet: Real-time radar obstacle and free space detection for autonomous driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 6958–6964. IEEE.

Reading, C.; Harakeh, A.; Chae, J.; and Waslander, S. L. 2021. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8555–8564.

Rebut, J.; Ouaknine, A.; Malik, W.; and Pérez, P. 2022. Raw high-definition radar for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17021–17030.

Shi, G.; Li, R.; and Ma, C. 2022. Pillarnet: Real-time and high-performance pillar-based 3d object detection. In *European Conference on Computer Vision*, 35–52. Springer.

Stäcker, L.; Heidenreich, P.; Rambach, J.; and Stricker, D. 2022. Fusion point pruning for optimized 2d object detection with radar-camera fusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3087–3094.

Tan, B.; Ma, Z.; Zhu, X.; Li, S.; Zheng, L.; Chen, S.; Huang, L.; and Bai, J. 2022a. 3-D object detection for multiframe 4-D automotive millimeter-wave radar point cloud. *IEEE Sensors Journal*, 23(11): 11125–11138.

Tan, B.; Ma, Z.; Zhu, X.; Li, S.; Zheng, L.; Chen, S.; Huang, L.; and Bai, J. 2022b. 3d object detection for multi-frame 4d automotive millimeter-wave radar point cloud. *IEEE Sensors Journal*.

Ulrich, M.; Braun, S.; Köhler, D.; Niederlöhner, D.; Faion, F.; Gläser, C.; and Blume, H. 2022. Improved orientation estimation and detection with hybrid object detection networks for automotive radar. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 111–117. IEEE.

Wu, D.; Yang, F.; Xu, B.; Liao, P.; and Liu, B. 2024. A Survey of Deep Learning Based Radar and Vision Fusion for 3D Object Detection in Autonomous Driving. *arXiv preprint arXiv:2406.00714*.

Xiong, W.; Liu, J.; Huang, T.; Han, Q.-L.; Xia, Y.; and Zhu, B. 2023. LXL: LiDAR excluded lean 3D object detection with 4D imaging radar and camera fusion. *IEEE Transactions on Intelligent Vehicles*.

Xu, B.; Zhang, X.; Wang, L.; Hu, X.; Li, Z.; Pan, S.; Li, J.; and Deng, Y. 2021. RPFA-Net: A 4D radar pillar feature attention network for 3D object detection. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 3061–3066. IEEE.

Yan, J.; Liu, Y.; Sun, J.; Jia, F.; Li, S.; Wang, T.; and Zhang, X. 2023. Cross modal transformer: Towards fast and robust 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18268–18278.

Yan, Q.; and Wang, Y. 2023. Mvfan: Multi-view feature assisted network for 4d radar object detection. In *International Conference on Neural Information Processing*, 493–511. Springer.

Yang, Z.; Chen, J.; Miao, Z.; Li, W.; Zhu, X.; and Zhang, L. 2022. Deepinteraction: 3d object detection via modality interaction. *Advances in Neural Information Processing Systems*, 35: 1992–2005.

Yin, T.; Zhou, X.; and Krähenbühl, P. 2021. Multimodal virtual point 3d detection. *Advances in Neural Information Processing Systems*, 34: 16494–16507.

Zhang, H.; Liang, L.; Zeng, P.; Song, X.; and Wang, Z. 2024a. SparseLIF: High-Performance Sparse LiDAR-Camera Fusion for 3D Object Detection. *arXiv preprint arXiv:2403.07284*.

Zhang, H.; Wu, K.; Chen, R.; Wu, Z.; Zhong, Y.; and Li, W. 2024b. TL-4DRCF: A two-level 4D radar-camera fusion method for object detection in adverse weather. *IEEE Sensors Journal*.

Zhang, Y.; Hu, Q.; Xu, G.; Ma, Y.; Wan, J.; and Guo, Y. 2022. Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18953–18962.

Zheng, L.; Li, S.; Tan, B.; Yang, L.; Chen, S.; Huang, L.; Bai, J.; Zhu, X.; and Ma, Z. 2023. Rcfusion: Fusing 4d radar and camera with bird's-eye view features for 3d object detection. *IEEE Transactions on Instrumentation and Measurement*.

Zheng, L.; Ma, Z.; Zhu, X.; Tan, B.; Li, S.; Long, K.; Sun, W.; Chen, S.; Zhang, L.; Wan, M.; Huang, L.; and Bai, J. 2022. TJ4DRadSet: A 4D Radar Dataset for Autonomous Driving. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 493–498.