

RaLiBEV: Radar and LiDAR BEV Fusion Learning for Anchor Box Free Object Detection Systems

Yanlong Yang^{ID}, *Member, IEEE*, Jianan Liu^{ID}, Tao Huang^{ID}, *Senior Member, IEEE*,
Qing-Long Han^{ID}, *Fellow, IEEE*, Gang Ma, and Bing Zhu^{ID}, *Senior Member, IEEE*

Abstract—In autonomous driving, LiDAR and radar are crucial for environmental perception. LiDAR offers precise 3D spatial sensing information but struggles in adverse weather like fog. Conversely, radar signals can penetrate rain or mist due to their specific wavelength but are prone to noise disturbances. Recent state-of-the-art works reveal that the fusion of radar and LiDAR can lead to robust detection in adverse weather. Current approaches typically fuse features from various data sources using basic convolutional/transformer network architectures and employ straightforward label assignment strategies for object detection. However, these methods have two main limitations: they fail to adequately capture feature interactions and lack consistent regression constraints. In this paper, we propose a bird’s-eye view fusion learning-based anchor box-free object detection system. Our approach introduces a novel interactive transformer module for enhanced feature fusion and an advanced label assignment strategy for more consistent regression, addressing key limitations in existing methods. Specifically, experiments show that, our approach’s average precision ranks 1st and significantly outperforms the state-of-the-art method by 13.1% and 19.0% at Intersection of Union (IoU) of 0.8 under “Clear+Foggy” training conditions for “Clear” and “Foggy” testing, respectively. Our code repository is available at: <https://github.com/yx75/RaLiBEV>.

Index Terms—Autonomous driving, anchor box free object detection, LiDAR, point cloud, radar, range-azimuth heatmap,

label assignment, bird eye’s view fusion, interactive transformer, deep learning.

I. INTRODUCTION

AUTONOMOUS driving systems typically employ sensors such as LiDAR, cameras, and radar to construct their perception functionality. Such perception scheme has been widely demonstrated in numerous works, covering various aspects such as data collection/generation [1], [2], algorithm development [3], and real-world testing [4]. This multi-modality hardware perception system with a developed software finally helps solving downstream task such as trajectory prediction and ego motion planning [5], [6], etc. LiDAR, a key sensor used in Advanced Driver Assistance Systems (ADAS) and autonomous driving, offers rich semantic and geometric information for environmental perception tasks [7]. These tasks include object detection [8], [9], multi-object tracking [10], [11], ego localization [3], and points registration for mapping [12], [13], [14]. However, single LiDAR modality is not able to comprehensively understand surrounding environment.

Recent works show that fusing LiDAR and camera data at the raw stream level [15] can significantly improve the robustness of autonomous driving perception systems [16], [17]. However, the performance of either camera or LiDAR drops severely in adverse weather, i.e., rain, snow, and fog [18], [19]. For instance, cameras become blurred, LiDAR’s operating range reduces significantly, and the number of noise points of LiDAR increases due to the impact of heavy snow or fog. As illustrated in Fig. 1(b), the number of noisy points around the ego vehicle increases while the far end LiDAR points disappear, compared to the LiDAR point cloud measured in clear weather, as shown in Fig. 1(a).

In contrast with cameras and LiDAR, radar can naturally avoid the influence of adverse weather because the millimeter wave is able to penetrate the rain, snow, and fog [20]. Such characteristic makes radar a powerful complementary sensor in ADAS and autonomous driving [21], and even in the cooperative perception [22]. The typical Frequency Modulated Continuous Wave (FMCW) radar, widely used in automotive systems, have the ability to accurately separate objects in range and speed dimension but having a low angular resolution due to the limited number of antennas in a compact space [23], [24], [25], [26]. Thus, radar can only provide many sparser detection points with ambiguity in angular information compared to the dense LiDAR point cloud. Such a drawback

Received 24 September 2024; revised 3 December 2024; accepted 14 December 2024. Date of publication 23 December 2024; date of current version 7 May 2025. The work of Yanlong Yang and Gang Ma was supported by Beijing Municipal Science and Technology Commission under Grant Z201100003920003. The work of Tao Huang was supported in part by Australian Government through Australian Research Council’s Discovery Projects Funding Scheme under Project DP220101634. This article was recommended by Associate Editor Y. S. Rawat. (Yanlong Yang and Jianan Liu are co-first authors.) (Corresponding authors: Jianan Liu; Tao Huang.)

Yanlong Yang is with the College of Science and Engineering, James Cook University, Smithfield Cairns, QLD 4878, Australia, and also with Vanjee Technology, Beijing 100193, China (e-mail: yanlong.yang@my.jcu.edu.au).

Jianan Liu is with Momoni AI, 43000 Gothenburg, Sweden (e-mail: jianan.liu@momoni.ai.com).

Tao Huang is with the College of Science and Engineering, James Cook University, Smithfield Cairns, QLD 4878, Australia (e-mail: tao.huang1@jcu.edu.au).

Qing-Long Han is with the School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Melbourne, VIC 3122, Australia (e-mail: qhan@swin.edu.au).

Gang Ma is with Vanjee Technology, Beijing 100193, China (e-mail: magang@wanji.net.cn).

Bing Zhu is with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China (e-mail: zhubing@buaa.edu.cn).

Digital Object Identifier 10.1109/TCSVT.2024.3521375

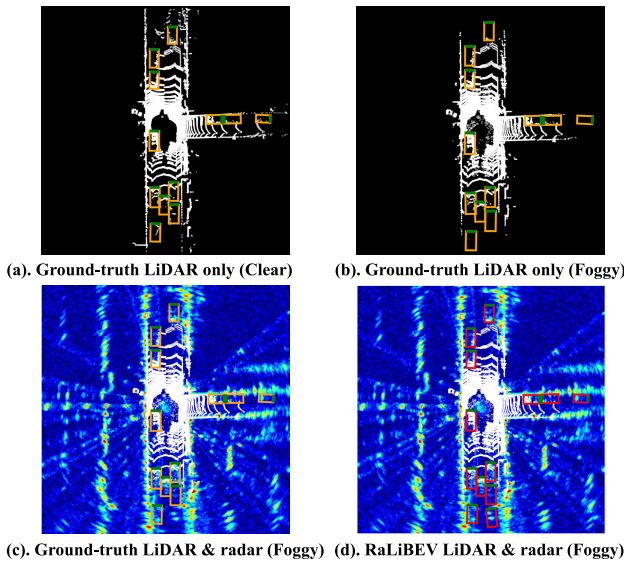


Fig. 1. Performance of the RaLiBEV in clear and foggy weather. The LiDAR and radar visualization results are combined by radar range azimuth heatmap with jet pseudo-color in the background and LiDAR with white point cloud and 2D object bounding boxes. The ground-truth boxes are orange, and the predicted boxes by RaLiBEV are red. All the boxes use green lines to indicate the heading direction.

limits its performance when being applied in the perception system.

To leverage radar's robustness in adverse weather, recent methods [27], [28] use radar range-azimuth heatmaps instead of detection points to fuse with LiDAR point clouds for vehicle detection. However, these anchor box-based methods face generalization issues due to the need for predefined anchor boxes tailored to specific datasets, resulting in unsatisfactory high-accuracy performance.

To address these challenges, we introduce RaLiBEV, a fusion model that integrates radar and LiDAR data in Bird's Eye View (BEV) for enhanced object detection without relying on anchor boxes. Unlike [3], which focuses on global localization using LiDAR-only approaches, our work emphasizes object detection by fusing radar and LiDAR, particularly in adverse weather conditions. While [16] explores camera-LiDAR fusion for 3D semantic segmentation, we leverage radar-LiDAR fusion to enhance detection accuracy in foggy conditions. In contrast to anchor box-based methods [64], which emphasize balancing samples of varying quality anchor boxes or different classes, our proposed method prioritizes achieving consistency between classification and box regression.

RaLiBEV transforms the LiDAR point cloud into a structured BEV format (pillarization) and uses radar range-azimuth heatmaps for discrete BEV sampling. Independent feature extraction from radar and LiDAR data in BEV is followed by fusion to create a detailed BEV feature. This feature is processed through a neural network backbone and a multi-scale detection head that estimates vehicle bounding boxes without anchor boxes.

In this work, two key innovative methods are proposed. First, an innovative label assignment strategy is proposed to reduce classification-regression inconsistencies. Second,

an interactive transformer-based BEV fusion technique is proposed to enhance radar and LiDAR feature map interaction. Experimental results show that RaLiBEV outperforms state-of-the-art methods by 13.1% and 19.0% at IoU 0.8 in clear and foggy weather tests, respectively.

The main contributions of this paper are concluded as below:

- A radar and LiDAR BEV fusion-based anchor box free object detector, RaLiBEV, is proposed. The RaLiBEV is able to generate accurate 2D bounding boxes in BEV by fusion of features extracted from the radar range-azimuth heatmap and the LiDAR point cloud.
- A novel label assignment strategy, Gaussian Area-based Consistent Heatmap and IoU Cost Positive Sample Assignment (GACHIPS), is applied to combat the inconsistency between the classification of key points and the corresponding bounding box regression tasks in the radar and LiDAR fusion-based anchor box free object detector.
- A novel interactive transformer-based fusion module, Dense Query Map-based Interactive BEV Fusion (DQMITBF), is proposed to fuse radar and LiDAR feature maps. With the introduction of dense, learnable queries and a symmetric fusion architecture, the fusion module provides an information-sensitive fusion mechanism to aggregate suitable features from both maps under various weather conditions.
- The aforementioned object detector can achieve fairly accurate detection performance even in adverse weather, as shown in of Fig. 1(d) compared to the ground-truth bounding boxes illustrated in Fig. 1(c) which outperforms the state-of-the-art radar and LiDAR fusion-based object detector with a large margin in the popular Oxford Radar RobotCar (ORR) dataset [1].

The rest of this paper is organized as follows. Section II discusses the related works, including LiDAR-based perception methods, radar-based perception methods, and the methods which fuse radar and LiDAR for object detection. In Section III, the proposed radar and LiDAR fusion-based object detection approach, RaLiBEV, is introduced. In addition, the proposed label assignment strategies and interactive transformer-based fusion schemes are also described. Then, the experiment process and corresponding results on ORR datasets are explained and analyzed in Section IV. Finally, the conclusion and further work are drawn in Section V.

II. RELATED WORKS

A. LiDAR-Based Object Detection

LiDAR, a pivotal sensor in environmental perception, plays a critical role in object detection within the surrounding environment. Deep learning techniques have enabled various approaches to LiDAR-based object detection, which can be categorized into four main methods based on their processing of input point clouds.

The first method, point-based detection, employs linear layers to extract features from unordered point clouds [29], [30]. In contrast, the grid-based approach discretizes point clouds into 2D grids or 3D voxels, utilizing convolutional

neural networks (CNNs) for feature extraction [8], [9], [31], [32]. Graph-based methods transform point clouds into graph data structures, leveraging Graph Neural Networks (GNNs) to extract features and detect objects [33]. Here, points are represented as vertices, and the relationships between points are defined as edges. The fourth category, the combination method, integrates features extracted from different representations of point clouds or multiple views of 2D grids, enhancing the overall detection performance [34], [35].

Despite the precise and dense 3D spatial information provided by LiDAR, its effectiveness can be limited by adverse weather conditions. Nonetheless, ongoing research is dedicated to overcoming these challenges and enhancing LiDAR-based object detection methods.

B. Radar-Based Object Detection

Radar, operating within the millimeter wave band, remains effective in object detection under adverse weather conditions. Conventional methods like CFAR [36] and clustering techniques [37] dynamically adjust detection thresholds and group points into objects but face challenges in complex environments and with overlapping objects.

Deep learning approaches, including semantic segmentation and clustering [38], [39], offer detailed object detection but require significant computational resources. Recent advancements such as PointNet [40], [41] and Graph Neural Networks (GNNs) [42], [43] directly process point cloud data, improving detection accuracy. These methods demand substantial training data and computational resources.

Range-azimuth heatmaps are also used for radar object detection, either through clustering and classification [44], [45] or deep learning models predicting object presence and location [46], [47], [48], [49], [50]. Despite radar's robustness in adverse weather, limitations like multi-path propagation and lower resolution hinder detailed object recognition, emphasizing the importance of sensor fusion. Integrating radar with LiDAR, which provides high-resolution point cloud data, can significantly enhance object detection and recognition capabilities.

C. Radar and LiDAR Fusion-Based Object Detection

Radar and LiDAR both have their strength and weaknesses. Fusing the information from the two types of sensors can achieve a better surrounding perception [27], [28], [51], [52], [53]. Generally, sensor fusion can be performed at the result or feature levels.

In result-level fusion, each sensor is characterized as an individual detection union. Then the detected objects are aggregated to generate the final result. This method consumes less transmission bandwidth and has a fast operation speed. However, its performance is severely limited by each single sensor detector.

In feature-level fusion, it is essential to aggregate feature information from both sensors, providing rich and reliable data for end-to-end networks [54], [55]. General fusion strategies, such as concatenating or adding data from different sensors after CNN-based feature extraction, offer faster computational

speeds but limited capacity in terms of weighted combination of feature information. Utilizing the attention mechanism from transformers [56] for information scraping has proven more effective [57]. However, previous efforts employing cross-attention, whether across different data channels, various temporal data [27], [28], [58] or different modality data [59] encounter effectiveness challenges. This is attributed to the limitation of relying solely on attention to capture information without query constraints, which restricts the model's ability to adapt to complex samples. Bi-LRFusion [60] proposes a bidirectional LiDAR and radar feature fusion scheme, which uses a query-based approach guided by both elevation-view and bird's-eye-view perspectives for fusing LiDAR and radar features. However, its utilization of sparse queries severely restricts its performance when handling dense radar range-azimuth heatmap data. To tackle this problem, we propose a dense, learnable query map-based transformer block that fusing features from each sensor. This approach offers dense and adaptable attention weights for enhanced sensor fusion.

D. Label Assignment in Object Detection

The traditional label assignment approach was first introduced in both anchor-based and anchor-free object detection frameworks. These frameworks use fixed thresholds to classify an anchor box or point as a positive sample, based on criteria such as the IoU between anchor boxes and ground-truth boxes, or whether an anchor point falls within a ground-truth box. However, these heuristic designs overlook the consistency between classification and box regression, leading to suboptimal box selection during inference and reduced detection performance [61].

In anchor-based frameworks [62], [63], all anchor boxes that meet the IoU or IoU plus boxes centers distance [64] above a specified threshold with the ground-truth box compute a combined loss of classification and box regression. However, This parallel loss design causes inconsistency issues. Similarly, in anchor-free frameworks [65], [66], [67], fixed threshold label assignment identifies certain anchor points within the ground-truth box as positive samples. The separate calculation and combination of classification and box regression losses result in a numerical gap, perpetuating inconsistency.

To address this, methods like freeAnchor [68], ATSS [69], AutoAssign [70], PAA [71], and OTA [72] align predicted outcomes with ground-truth, reducing the consistency gap. However, these methods focus on image-based object detection, addressing varying object sizes and occlusion by assigning multiple positive samples to individual targets. This leads to competition among positive samples during training, exacerbating the consistency gap during inference.

III. PROPOSED METHOD

In this section, the framework of the proposed anchor box-free radar and LiDAR fusion-based object detector are first introduced. Then, two enhancements of RaLiBEV, one with label-assignment strategies and the other with interactive transformer fusion strategies, are described.

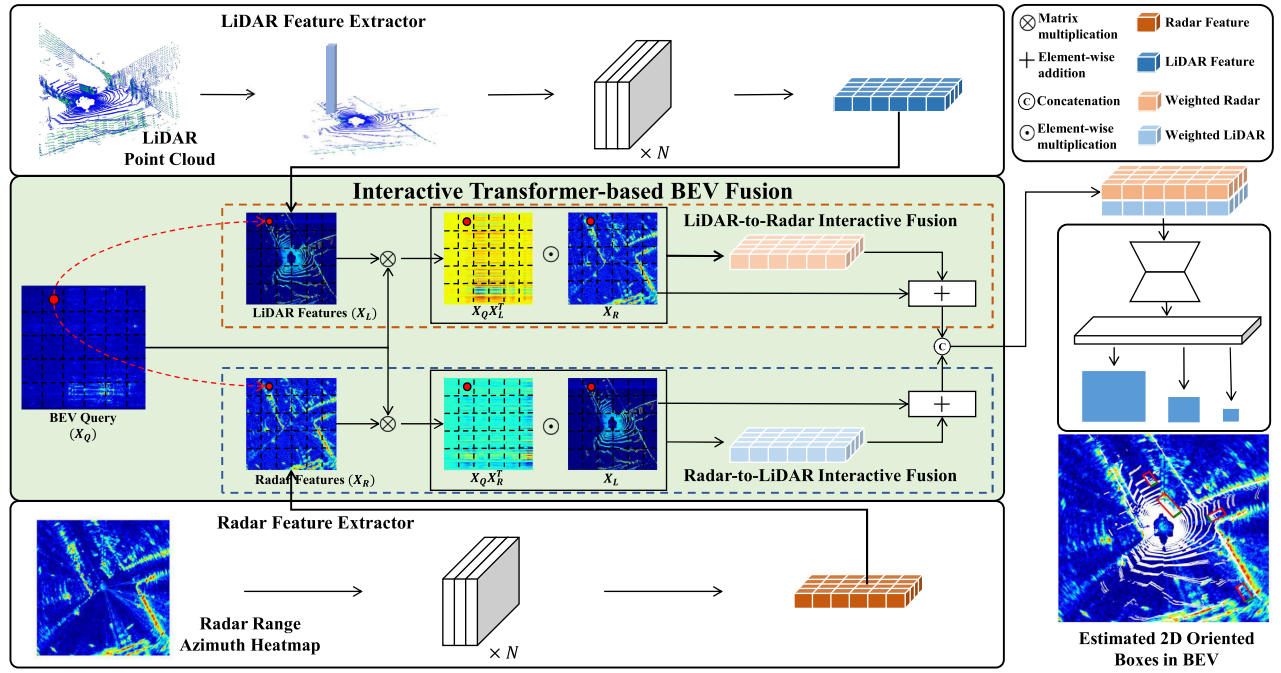


Fig. 2. The pipeline of the proposed radar and LiDAR fusion-based anchor box free object detector, RaLiBEV.

A. Anchor Box Free Object Detection

Aggregating features from each sensor under BEV representation is straightforward. As shown in Fig. 2, the raw LiDAR point cloud and radar range azimuth heatmap undergo pre-processing via their respective feature extractors. These features are then fused and processed by a YOLOV4-based [73] feature extractor and a multi-scale anchor-free detection head to decode 2D BEV rotating bounding boxes.

Radar data is converted from polar to Cartesian coordinates using bilinear interpolation, resulting in a shape of $[W_r, H_r, C_r]$. LiDAR data is structured using PointPillars [9] into a fixed shape $[W_l, H_l, C_l]$ under BEV.

Advanced feature extraction techniques are applied to each sensor branch using N convolution layers. The extracted features are integrated using an enhanced fusion module with an interactive transformer mechanism. The fused features are then processed by the YOLOv4-based backbone for further refinement.

A multi-scale anchor-free detection head decodes the detection results. The first output channel generates a heatmap representing the object's existence probability. Subsequent channels predict Δx , Δy , width, length, angle classification, angle offset, and object class.

Using the anchor-free detection approach, positive anchor points are identified on the feature map, and regression begins from those points. For each ground-truth bounding box in the BEV, a Gaussian distribution map is generated as a supervision signal for heatmap prediction:

$$G(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi^{\frac{d}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \quad (1)$$

where $\mathbf{x} = [i, j]$ is the BEV spatial vector, $\boldsymbol{\mu}$ is the center coordinates of ground-truth boxes, and $\boldsymbol{\Sigma}$ is the covariance

matrix:

$$\boldsymbol{\Sigma} = E(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^T, \quad (2)$$

using the four corners' 2D coordinates of a box as $\mathbf{x}$.

B. Label Assignment Strategies

For object detection, the total loss is the sum of classification loss ($\mathcal{L}_{cls}$) and box regression loss ($\mathcal{L}_{box}$):

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{box}, \quad (3)$$

The classification loss ($\mathcal{L}_{cls}$) uses focal loss modified with Gaussian mixture distribution weight $G_{i,j}$:

$$\mathcal{L}_{cls} = \begin{cases} - \sum_{i,j=0}^M G_{i,j} (1 - p_{i,j})^\gamma \log(p_{i,j}) & \text{if } G_{i,j} = 1, \\ - \sum_{i,j=0}^M G_{i,j} p_{i,j}^\gamma \log(1 - p_{i,j}) & \text{otherwise.} \end{cases} \quad (4)$$

The box loss ($\mathcal{L}_{box}$) includes Δx , Δy , width, length, and angle offset $\Delta \theta$ loss, formulated as smooth $L1$ loss, and angle classification loss ($l_{\Delta \theta}^{ce}$):

$$\mathcal{L}_{box} = \sum_{i,j=0}^M \alpha_{i,j} (l_{\Delta xy}^{reg} + l_{wl}^{reg} + l_{\theta}^{ce} + l_{\Delta \theta}^{reg}). \quad (5)$$

The box regression loss occurs only on positive samples ($\alpha_{i,j} = 1$), while classification loss is computed on all anchor points. This can cause inconsistency in label assignments, affecting detection accuracy.

Three label assignment strategies are typically used: bipartite matching, multi-positive, and single-positive label assignment. "Multi/single" indicates the number of anchor points chosen as positive samples for one object.

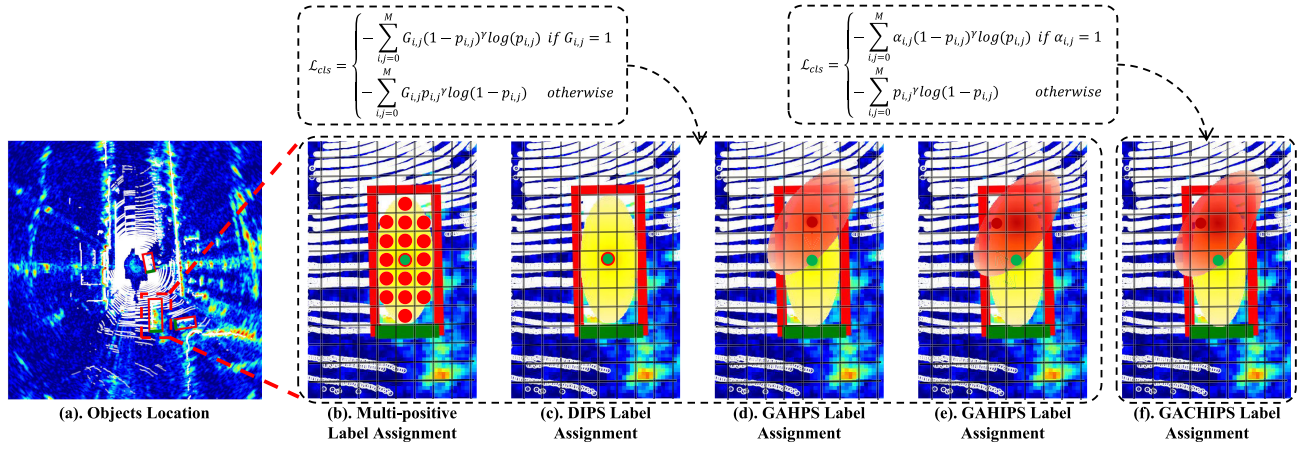


Fig. 3. Overview of label assignment strategies for object detection. (a) Identifies the object with a red-bordered ground-truth bounding box and a green headline. Yellow and red ellipses represent ground-truth and predicted Gaussian distributions, respectively. Red dots are positive sample points, and green dots mark the ground-truth Gaussian centers. Strategies (b) to (e) show different methods for selecting positive samples for box loss calculation, ranging from (b) using all anchor points within the ground-truth area, to (c) selecting the center, (d) the point with the highest foreground-background classification score, and (e) the point with the highest “score plus IoU”. Strategy (f) integrates the approach from (e) with an alternative loss function as Eq. (14).

1) *Bipartite Matching Label Assignment*: The typical bipartite matching label assignment is based on the Hungarian algorithm. This set-based label assignment builds a cost matrix of IoU between predicted and ground-truth boxes, where the predicted boxes are decoded from a predefined location. Then it finds the best matching positive samples by the Hungarian algorithm. The traditional bipartite matching process is based on anchors [75], while DETR [76] leveraging object query changes this matching art to a total end-to-end style. Using the Hungarian algorithm, DETR computes loss directly between predefined object queries and ground-truth object boxes. Then after training, the converged queries can be directly decoded into the final predicted boxes without Non-Maximum Suppression (NMS). However, the size of the query map in our proposed method is set to be the same as the size of the radar and LiDAR feature map. Such a large number of queries leads to massive computational consumption, which is not feasible when applying the Hungarian algorithm. Thus, such bipartite matching label assignment is not considered for our proposed method.

2) *Multi-Positives Label Assignment*: Following the idea that multiple positive anchor points may alleviate positive and negative sample imbalance problems in anchor-free detection architecture, a Gaussian area-based multi-positive label assignment is designed. Based on the ground-truth Gaussian mixture distribution heatmap $\mathbf{G}$ created in Section III-A, a static threshold is used to extract the specific region as the positive anchor points. Foreground-background classification and box regression are then applied together on each positive anchors points. As Fig. 3(b) shows, all dots with red colors are the selected positive samples. They are all located in the ground-truth Gaussian distribution yellow-colored area, thus this approach is named as *Gaussian Area-based Multi-Positives Sample assignment*.

Nonetheless, the introduction of multiple positive samples inevitably leads to competition in training procedure among them because they all devote to the gratitude, exacerbating the

consistency gap during the interference phase. The experiment result is shown later in Section IV-C1.

3) *Single Positive Label Assignment*: In order to prevent numerical regression conflicts caused by multiple positive samples, single positive sample allocation strategy has been designed. The fundamental characteristic of a single positive sample is that, for each ground-truth object, there is only one positive anchor point corresponding to it, which contributes to the gradient of the loss in each training iteration. Four methods have been proposed successively to alleviate the inconsistency between foreground-background classification and box regression mentioned in the previous discussion. The experiment shows that abandoning the center prior can most effectively alleviate the consistency issue between object detection and regression. The proposed four methods share the same training framework as below:

- ① Generate candidate anchor points: For each ground-truth object, create a set of candidate anchor points.
- ② Design unique costs: Assign a unique cost to each candidate anchor point.
- ③ Select positive samples: Choose the most suitable candidate anchor points as positive samples.
- ④ Compute loss: Calculate the loss result based on the selected positive anchor points.

The four label assignment strategies contribute different innovations in the pipeline. The details are elaborated below.

The first step is generating candidate anchor points. This step is the common procedure for all four label assignment strategies. According to Eq. (1), a Gaussian distribution $\{\mathbf{G}_k, k = 1, \dots, n\}$ is generated for each ground-truth object, and the multiple objects form a Gaussian mixture distribution $\mathbf{G}$. By setting a fixed threshold, the anchor points inside each Gaussian distribution can be obtained, which are called candidate anchor points. This step can be described as:

$$\mathbf{A}_k^{\mathbf{G}} = \mathcal{A} \cap \mathbf{G}_k, \quad (6)$$

where A_k^G represents the set of candidate anchor points in the k -th ground-truth Gaussian distribution, $\mathcal{A}$ represents the set of all anchor points on the BEV grid, and $\cap$ represents the intersection operation.

The second and third steps are to design unique cost for these candidate anchor points and then select one as positive sample. In these two steps, the first three label assignment strategies make improvements rely on different design:

- The first design is *Direct Index based Positive Sample (DIPS) assignment*. As Fig. 3(c) shows, DIPS treats the label center locations as the positive anchor points, then applies foreground-background classification and box regression on these anchor points. The DIPS sample assignment is a standard design of CenterPoint [8] whose cost and selection are described as:

$$a_k = \frac{1}{n} \sum_{i=1}^n A_k^G. \quad (7)$$

The process begins by determining if a candidate anchor point is positioned at the center of A_k^G . Here n represents number of anchor points of k^{th} object. The location of center anchor a_k is calculated from the mean location of all the candidate anchor positions. This is calculated and represented as a_k according to Eq. (7). Following this, a binary mask, symbolized as α , is used to indicate the selection of positive samples. This mask is assigned a value of 1 in accordance with:

$$\alpha(a_k) = 1. \quad (8)$$

After the establishment of α , the location for the box regression loss is pinpointed. Subsequently, the final loss is then calculated based on Eq. (3), which is summation of two values obtained from Eq. (4) and Eq. (5) for all the positive samples.

- The second approach is termed as the *Gaussian Area-based Heatmap cost Positive Sample (GAHPS) assignment*. In this method, we substitute the cost calculation from determining the centers of ground-truth boxes to assessing the maximum predicted heatmap value for each anchor point. The remaining steps, including the first and the fourth, are kept unchanged. Firstly, the model infers predicted heatmap P_k^{cls} and boxes P_k^{box} follows Eq. (9). As illustrated in Fig. 3(d), the red fading ellipse represents the predicted Gaussian distribution heatmap. This step is described as:

$$P_k^{cls}, P_k^{box} = \text{Decode}(F, A_k^G). \quad (9)$$

The decoding function, denoted as $\text{Decode}(\cdot)$, takes two inputs: the feature map F and the anchor points in k -th Gaussian distribution A_k^G . Then, different from DIPS, the cost is designed as predicted heatmap as shown in:

$$C_k^a = P_k^{cls}. \quad (10)$$

The anchor point with highest predicted heatmap value are chosen as the positive sample a_k as shown in:

$$a_k = \text{argmax}(C_k^a). \quad (11)$$

The Eqs. (9), (10), and (11) are utilized to substitute Eq. (7). Following this, the binary mask α is computed using Eq. (8). Similar to the first DIPS strategy, the final loss is also calculated using Eq. (3), (4) and (5). However, the final loss is determined based on the established α here in this design.

- The third approach is termed as the *Gaussian Area-based Heatmap and IoU cost Positive Sample (GAHIPS) assignment*, as shown in Fig. 3(e). In this method, we go beyond merely substituting the cost with the predicted heatmap value. Instead, we replace it with the sum of the predicted heatmap value and the IoU value between the predicted box and the ground-truth box. Therefore, the first step is to calculate the IoU score of predicted bounding box and ground-truth box. The decoded P_k^{box} derived from Eq. (9) is utilized to compute the IoU scores S_k^{IoU} in relation to the ground-truth bounding boxes $\mathcal{G}_k^{box}$:

$$S_k^{IoU} = \text{computeIoU}(P_k^{box}, \mathcal{G}_k^{box}). \quad (12)$$

Following this, the sum of the predicted heatmap values and IoU scores, represented as C_k^a , is employed as the new cost, as indicated in:

$$C_k^a = P_k^{cls} + S_k^{IoU}. \quad (13)$$

The positive anchor point a_k is then selected using Eq. (11), and the binary mask α is also determined by Eq. (8). Lastly, same as aforementioned strategies, the final loss is still calculated using Eq. (3), which is a combination of Eq. (5) and Eq. (4).

To be noted that, although the processes of determining α are different in DIPS, GAHPS and GAHIPS, the ways of computing final loss are same. However, because the foreground-background classification loss and box regression loss are independent from each other, such definition of final loss leads to a consistency gap between the two loss above. To address this issue, the fourth design is proposed.

- The fourth design is *Gaussian Area-based Consistent Heatmap and IoU cost Positive Sample (GACHIPS) assignment*, which is a divergence of the GAHIPS. GACHIPS follows the previous three steps as the GAHIPS, but modifies the final step of loss computation as indicated in Fig. 3(f). By doing it, the foreground-background classification loss and box regression loss are both controlled by α . Specifically, the former foreground-background classification loss function which is shown in Eq. (4), should be replaced by:

$$L_{cls} = \begin{cases} - \sum_{i,j=0}^M \alpha_{i,j} (1 - p_{i,j})^\gamma \log(p_{i,j}) & \text{if } \alpha_{i,j} = 1, \\ - \sum_{i,j=0}^M p_{i,j}^\gamma \log(1 - p_{i,j}) & \text{otherwise.} \end{cases} \quad (14)$$

During inference procedure, the key step is to extract peaks from the predicted heatmap. We set a static threshold to pick every anchor point to generate the corresponding box, then

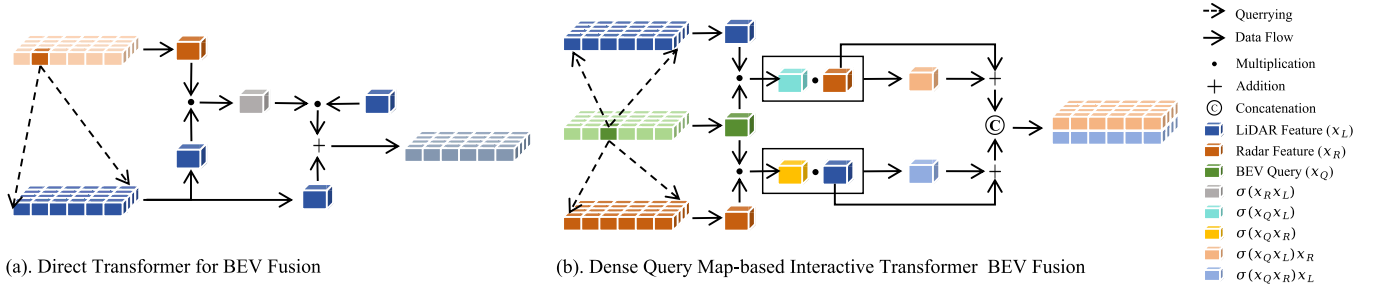


Fig. 4. Comparison of direct transformer for BEV fusion with dense query map-based interactive transformer for BEV fusion.

NMS is applied to discard other boxes that overlap with the higher confidence box by a certain size. Beware that NMS can still filter out a certain amount of overlapping bounding boxes estimated from three branches in multiple scales. However, the estimated bounding boxes rarely overlap in a single scale branch since the anchor box-free approach with a single positive label assignment strategy is employed.

C. Interactive Transformer for Fusion of Radar and LiDAR in BEV

In the pursuit of enhancing the performance of sensor fusion in autonomous driving systems, we propose a novel approach that transcends the conventional method of simply concatenating features from each sensor. The standard transformer [57] based traditional approach, while straightforward, is inherently limited in its ability to effectively weigh the features from each sensor. To this end, we have designed an innovative interactive transformer-based BEV Fusion block, to fuse data from the two sensor branches more effectively. In the following, we will describe the traditional transformer-based fusion method for comparison purposes, followed by our proposed interactive transformer fusion method. Our method builds on the traditional method but includes significant innovation.

The traditional transformer-based fusion method, termed Direct Transformer for BEV Fusion, serves as a foundational approach employing a straight-forward transformer mechanism for BEV fusion. Illustrated in Fig. 4(a), it utilizes the radar feature map as the query and the LiDAR feature map as both the key and value. The fusion process involves computing a weight matrix through the matrix multiplication of the query and key, normalizing it with a softmax function, and then applying the dot product of the weight and value, with an added residual process involving the value itself. This can be mathematically represented as:

$$X_{out} = \sigma(X_R X_L^T) \odot X_L + X_L. \quad (16)$$

Here, X denotes the feature map, $\sigma(\cdot)$ denotes the softmax function, $\odot$ denotes element-wise multiplication, and X_R , X_L represent the radar and LiDAR feature maps, respectively. This baseline method demonstrates the initial effectiveness of transformer-based BEV fusion.

The approach proposed in this paper, Dense Query Map-based Interactive Transformer for BEV Fusion (DQMITBF), is based on the core concept of the conventional Direct Transformer. However, a significant variation in the query mechanism is introduced in the DQMITBF. Shown in Fig. 4(b), this method employs a randomly initialized, learnable BEV query of the same size as the input key feature map. It facilitates interactive attention between this BEV query and both radar and LiDAR feature maps, treated as key and value, respectively. The fusion process involves three steps: First, calculating the correlation between the BEV query and radar feature map to enable radar-to-LiDAR interaction. Second, using the softmax output as a weighting factor for the LiDAR feature map. Third, adding the weighted LiDAR feature map to the original LiDAR feature. This process is mirrored for LiDAR-to-radar feature fusion, culminating in the concatenation of the fused radar and LiDAR feature maps for further processing. The fusion can be succinctly represented as Eq. (15), shown at the bottom of the page. In this equation, X_Q is the learnable query, X_R and X_L are the radar and LiDAR feature maps, respectively, and $\odot$ denotes the concatenation process.

The DQMITBF architecture interacts symmetrically with radar and LiDAR features through a BEV query and integrates them in a unified phase. This interactive query allows the model to prioritize feature-rich branches, especially in varying weather conditions. Thus, this novel approach optimizes the weighting and combination of heterogeneous sensor features in an adaptive manner, enhancing the effectiveness of autonomous driving systems in different scenarios.

D. Possible Extension to Other Perceptual Tasks

RaLiBEV, while primarily focused on object detection, has the potential to generalize to tasks like instance and semantic segmentation due to its flexible transformer-based fusion module and dense BEV queries. For instance segmentation, the detection head could be extended with a mask prediction branch, and for semantic segmentation, the fusion mechanism could be refined for per-pixel classification. The BEV-based representation in RaLiBEV effectively captures spatial and

$$X_{out} = (\sigma(X_Q X_R^T) \odot X_L + X_L) \odot (\sigma(X_Q X_L^T) \odot X_R + X_R) \quad (15)$$

TABLE I
EXAMPLE LIST OF PR CURVE AT IoU = 0.5

Scores	0.98	0.97	0.96	0.95	0.94	0.92	0.91
Prediction Objects	TP	TP	FP	TP	TP	TP	FP
Precision	1	1	2/3	3/4	4/5	5/6	5/7
Recall	1/5	2/5	2/5	3/5	4/5	1	1

geometric relationships, making it suitable for these dense prediction tasks.

IV. IMPLEMENTATION AND RESULTS

A. Dataset and Evaluation Metrics

The ORR dataset [1] is used for its synchronized LiDAR point cloud, radar range azimuth heatmap, and car motion pose parameters. LiDAR data is captured by two Velodyne HDL-32E LiDARs, and radar data by a NavTech CTS350-X radar with a 360° field of view. Radar angle resolution is 0.9° every 400ms, while LiDAR is 0.33° every 20ms.

Synchronization between LiDAR and radar data, and annotation generation, is achieved by interpolation using ego vehicle pose [27]. Pose parameters correct misalignment, and annotations are generated every 20 frames, with intermediate frames interpolated. For foggy scenarios, MVDNet [27] relocates LiDAR points using DEF [51], setting fog probability to 0.05 to adjust visibility. The dataset includes 8862 frames, split into 7071 for training and 1791 for testing.

Evaluation uses 2D rotated bounding box average precision (AP) under different IoU thresholds. AP measures precision and recall, with higher AP indicating better performance. Prediction boxes are classified as True Positive (TP) if they overlap with ground-truth objects above the IoU threshold, or False Positive (FP) if below. Precision and recall are defined:

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP + FP}, \\ \text{Recall} &= \frac{TP}{TP + FN}. \end{aligned} \quad (17)$$

Table I shows AP calculation at IoU = 0.5. Precision decreases and recall increases with lower object scores. AP@IoU = 0.5 is the area under the precision-recall curve, with higher values indicating better model performance.

B. Implementing Details

- **Architecture details:** The preprocessing of LiDAR data is based on the PointPillars framework, where a BEV spatial resolution of 0.2 m×0.2 m is used. This transforms the unstructured LiDAR point cloud into a structured matrix with dimensions 320 × 320 × 9. Similarly, the radar data is naturally represented in a BEV format with dimensions 320 × 320 × 1.

As illustrated in Fig. 2, the subsequent convolutional block includes a 3 × 3 convolution layer, a batch normalization layer, and a leaky ReLU activation layer. This block is replicated $N = 3$ times. After processing through the convolutional network, both the LiDAR and radar features are transformed into a shape of 320 × 320 × 16.

TABLE II
PARAMETERS OF PROPOSED NETWORK

Parameters	Value
BEV Resolution (m)	0.2 × 0.2
Pre-process Feature Shape of Both Branch	320 × 320
LiDAR Pillar Feature Channel	9
Radar Range-azimuth Feature Channel	1
Fusion Module Input Feature Channel of LiDAR/Radar	16/16
Fusion Module Output Feature Channel	32
Detection Head Scale Number	3
Detection Head Channel Number	8

Following this preprocessing, the LiDAR and radar features are input into the Interactive Transformer-based BEV Fusion module. In this module, the features are weighted and fused using a pre-set BEV query, which also has a shape of 320 × 320 × 16. The fused feature, now with dimensions 320 × 320 × 32, is passed into a YOLO-V4-based feature extraction and detection network.

The YOLO-V4 backbone comprises Darknet-53 and a three-scale detection head. During inference, the predicted boxes are decoded from the anchor-free heads of the three-scale output. A static threshold of 0.1 is applied to extract peaks, which represent key points among all anchor points in the predicted heatmap. After box generation from the three-scale output feature maps, non-maximum suppression (NMS) is employed to eliminate redundant boxes, with a threshold set at 0.2.

- **Parameter details:** The details of the parameters for our proposed network are provided in Table II.
- **Training process details:** During training, a batch size of 32 is used, with four Nvidia Tesla V100S GPUs. The first two epochs serve as a warm-up phase, during which the learning rate linearly increases from 0 to 0.01. Afterward, the learning rate decreases by a factor of 0.1 every 10 epochs. The training process is optimized using the ADAM optimizer.

C. Experimental Results and Analysis

1) **Ablation Study on Label Assignment Strategies:** Experiments on the Oxford Radar RobotCar dataset evaluate the proposed strategies, focusing on AP for cars at IoU 0.5, 0.65, and 0.8.

Table III shows results for different label assignment strategies and interactive transformer fusion architectures in RaLiBEV. The **Gaussian Area-based Multi-Positive Sample Assignment** strategy, using the YOLOv4 backbone and neck, serves as a baseline. It achieves APs of 95.4% at IoU 0.5 and 94.2% at IoU 0.65, both improving by 2.2%, but drops 5.8% at IoU 0.8 to 75.7%, compared to ST-MVDNet++ [58], which has APs of 93.2%, 92.0%, and 81.5% at IoUs of 0.5, 0.65, and 0.8, respectively.

The **DIPS Assignment** strategy shows minor improvements of 0.7% - 0.9% at IoUs of 0.5 and 0.65, but a 6% improvement at IoU 0.8. The **GAHPS** strategy further improves high precision AP by 5%, reaching 89.5% at IoU 0.8. The **GAHPS** strategy maintains similar performance at IoU 0.5 but significantly improves AP at IoU 0.8 to 92.0%, a 10.5% increase over ST-MVDNet++. The **GACHIPS** strategy, using heatmap

TABLE III
ABLATION STUDY OF LABEL ASSIGNMENT METHODS

Methods	AP@IoU=0.5	AP@IoU=0.65	AP@IoU=0.8
Gaussian Area-based Multi-Positives Sample Assignment	95.4	94.2	75.7
Direct Index-based Positive Sample assignment (DIPS)	96.3	94.9	84.5
Gaussian Area-based Heatmap cost Positive Sample assignment (GAHPS)	96.7	95.4	89.5
Gaussian Area-based Heatmap and IoU cost Positive Sample assignment (GAHIPS)	96.8	96.7	92.0
Gaussian Area-based Consistent Heatmap and IoU cost Positive Sample assignment (GACHIPS)	97.4	96.7	93.9

TABLE IV
ABLATION STUDY OF FUSION METHODS. (w/O: WITHOUT)

Method	Query	Key	Value	AP@IoU=0.5	AP@IoU=0.65	AP@IoU=0.8
DQMITBF	BEV Query	L R	R L	98.8	97.7	94.8
One Side 1	BEV Query	L	R	97.9	97.8	93.9
One Side 2	BEV Query	R	L	97.8	97.5	92.3
w/o BEV Query	R	L	L	96.3	94.7	84.5
w/o Transformer	/	/	/	96.2	94.0	82.0

and IoU as cost, further improves AP to 97.4% at IoU 0.5 and 93.9% at IoU 0.8.

Experiments indicate that a **multi-positive sample** strategy can adversely affect high-precision performance due to inter-sample competition. This strategy, which involves calculating multiple losses for the same object, introduces inconsistencies, particularly in the Δx and Δy values, thereby hindering optimal convergence.

In traditional image detection, multi-positive sample strategies enhance detection rates under overlapping conditions. However, this approach proves unsuitable for BEV perception in autonomous driving, where targets typically do not overlap. By assuming a single positive sample point, inter-sample competition is eliminated, allowing the model to focus on a singular convergence point. Nevertheless, this can lead to mismatches between foreground-background classification and bounding box regression, limiting peak performance. The **DIPS** strategy, which uses a fixed target center for box regression loss calculation, may not align with the optimal classification point during inference, resulting in accuracy loss due to classification-regression inconsistency.

The **GAHPS** strategy addresses this issue by utilizing peak points from the inference heatmap as box regression anchors during training, thereby enhancing performance. Compared to DIPS, GAHPS improved precision at IoU 0.8 by 5%.

Further refinement of the cost design by integrating IoU with the heatmap led to additional accuracy gains. **GAHIPS** surpassed GAHPS by 2.5% at IoU 0.8. Building on this, **GACHIPS** allows the selected positive sample to manage both classification and regression losses, maximizing consistency. It employs the optimal anchor point from inference for loss calculations during training and for decoding the optimal detection box during inference, thereby unifying classification and regression while aligning training and inference phases. This approach achieved the highest precision, reaching 93.9% at IoU 0.8.

2) *Ablation Study on Interactive Transformer-Based BEV Fusion Approaches*: To verify the effectiveness of our proposed method from other fusion methods. Several control groups of different fusion methods are presented in Table IV.

As depicted in Eq. (15) and Fig. 4, our proposed method incorporates a symmetric data interactive fusion module. This

module consists of two symmetric parts: one side uses LiDAR as the key and radar as the value, while the other uses radar as the key and LiDAR as the value. Therefore, ablation studies were designed to evaluate the functionality of each fusion side, the effectiveness of the BEV Query, and the impact of the transformer architecture.

Based on the results presented in Table IV, the findings can be summarized as follows: the symmetric interactive sensor fusion algorithm based on the transformer achieves the highest performance in object detection.

- **Symmetric Fusion Architecture**: In the first two ablation studies, *One Side 1* and *One Side 2*, the effectiveness of the symmetric interaction fusion design is validated. When only one side of the interactive fusion is used, the model's performance fails to reach its optimal level, demonstrating the necessity of the symmetric architecture.
- **BEV Query**: In the third experiment, the BEV Query is removed, resulting in a noticeable drop in performance on $AP@IoU = 0.8$. This indicates that the BEV Query significantly contributes to the model's ability to meet higher precision requirements.
- **Transformer Module**: In the final experiment, the transformer fusion module is replaced with a simple concatenation process. This substitution leads to a marked deterioration in the model's overall performance, highlighting the critical role of the transformer architecture in the proposed design.

These ablation studies collectively demonstrate the importance of the symmetric interactive fusion module, the BEV Query, and the transformer architecture in achieving superior object detection performance.

To demonstrate the effectiveness of DQMITBF in a more comprehensible manner, the feature maps of two selected two frames in the interactive transformer fusion module are visualized as depicted in Fig. 5. In Fig. 5, The top two rows correspond to Frame 2799 and the bottom two rows correspond to Frame 3299, which are selected randomly. Each frame includes selections from clear and foggy weather conditions for analysis. For each column, the feature maps X_Q , X_R , X_L , $X_Q X_R^T$, and $X_Q X_L^T$ are illustrated for visual

TABLE V
COMPARISON OF LiDAR ONLY APPROACHES AND RaLiBEV WHICH EMPLOYS FUSION OF RADAR AND LiDAR.
THE NUMBERS IN BOLD ARE THE BEST OF ALL METHODS

Method	Train	Modality	Clear+Foggy						Clear-Only					
	Test		Clear			Foggy			Clear			Foggy		
	IoU		0.5	0.65	0.8	0.5	0.65	0.8	0.5	0.65	0.8	0.5	0.65	0.8
PointPillars [9]		L	85.8	83.0	58.3	72.8	70.3	48.6	85.8	82.9	60.6	71.3	68.3	47.8
RaLiBEV LiDAR-Only		L	96.3	93.8	78.3	92.2	88.3	74.4	97.7	96.4	87.5	68.2	62.7	40.2
RaLiBEV with GACHIPS (Ours)		L+R	97.7	97.7	93.9	97.8	96.7	93.7	97.4	96.7	93.9	84.6	82.4	73.8
RaLiBEV with DQMITBF (Ours)		L+R	98.8	97.7	95.1	97.9	97.8	94.1	98.8	97.7	94.8	83.3	80.3	70.6

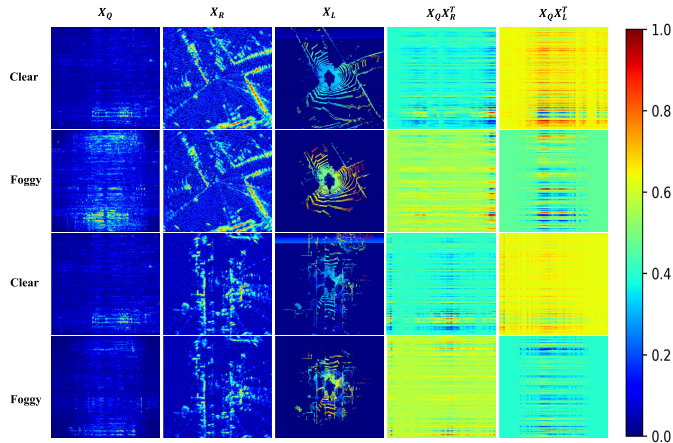


Fig. 5. Interactive transformer feature map visualization results. Above two rows and below two rows represent model performance in different scenarios.

comprehension. Furthermore, all feature maps are normalized to the range $[0, 1]$ for a consistent representation.

Utilizing a unified BEV query X_Q , the input radar feature map X_R and LiDAR feature map X_L are weighted to $X_Q X_R^T$ and $X_Q X_L^T$ respectively, subsequently producing a fused feature for additional processing. It can be observed that X_Q manages radar and LiDAR features differently under varying weather conditions. In clear weather, $X_Q X_L^T$ is noticeably larger in numerical value than $X_Q X_R^T$. However, in foggy conditions, the input X_L is significantly impacted by noise, resulting in $X_Q X_R^T$ being larger than $X_Q X_L^T$. From these observations, we can draw the following conclusions:

- The approach utilizes an interactive transformer to aggregate data from LiDAR and radar branches in a unified manner.
- The BEV query in the interactive transformer can focus on more information-rich signals under different weather conditions.

Additionally, as depicted in Fig. 5, the inputs X_L and X_R resemble the raw source data, and X_Q does not focus on specific object areas. This is because the interactive transformer fusion module operates early in the encoder process, where data features exist in a raw energy distribution space rather than a box location feature space.

The experiment results show that a reasonable fusion logic significantly enhances RaLiBEV's performance. While direct transformer fusion among feature maps can improve performance, interactive transformer-based concepts yield superior

results due to the innovative querying fusion method, which introduces symmetry and unifying properties into the model.

3) *Ablation Study of Multi-Modality Data Fusion:* To highlight the benefits of multi-modality fusion, Table V compares performance metrics between the LiDAR-only approach and the fusion of radar and LiDAR. The comparison shows that neither RaLiBEV with LiDAR only nor PointPillars match the performance levels achieved by radar and LiDAR fusion. This enhancement is particularly noticeable in foggy scenarios. When trained on a "Clear+Foggy" dataset, RaLiBEV in the radar and LiDAR fusion modality (denoted as "R+L") exhibits exceptionally high performance, especially under the stringent requirements of a 0.8 IoU in "Foggy" test scenarios.

The "Clear-only" training experiment assesses model robustness Table V demonstrates that GACHIPS slightly outperforms DQMITBF in "Foggy" conditions after "Clear-only" training, with IoU improvements observed at various thresholds. Nonetheless, both models exhibit significant performance degradation in "Foggy" conditions absent from the training data. This decline arises from their limited capacity to handle novel foggy data, with DQMITBF being especially impacted due to its dependence on LiDAR data, rendering it vulnerable to atmospheric noise. These findings highlight the need for training under diverse conditions, as demonstrated in experiments including both "Clear" and "Foggy" data, to enhance model robustness and adaptability.

4) *Performance Comparison With State-of-the-Art Methods on Different Weather Conditions:* This study, following MVDNet [27], evaluates model performance across different weather conditions, detailed in Table VI. We use MVDNet's method to simulate foggy LiDAR data for experiments comparing training with "Clear-only" and "Clear+Foggy" data against "Clear" and "Foggy" testing scenarios. Two top-performing models, RaLiBEV with GACHIPS and DQMITBF, are highlighted. GACHIPS outperforms DQMITBF in "Clear-only" training and "Foggy" testing, showing a 10% improvement post data enhancement. Specifically, GACHIPS's performance jumps to 97.8%, 96.7%, and 93.7% in "Foggy" tests, surpassing ST-MVDNet [28] by over 20% and Bi-LRFusion [60] by 20.7% in IoU 0.8. Moreover, RaLiBEV shows advancements of 4.4%, 7.8%, and 19.0% in IoU 0.5, 0.65, and 0.8 respectively, under "Clear+Foggy" training and "Foggy" testing, compared to ST-MVDNet++ [58]. These results demonstrate RaLiBEV's superior accuracy in adverse weather, benefiting from its fusion logic and label assignment strategy.

TABLE VI

COMPARISON OF AP OF ORIENTED BOUNDING BOXES IN BIRD'S EYE VIEW BETWEEN OUR PROPOSED RaLiBEV AND OTHER STATE-OF-THE-ART RADAR AND LiDAR FUSION-BASED METHODS. THE NUMBERS IN BOLD ARE THE BEST OF ALL METHODS

Method	Train	Modality	Clear+Foggy						Clear-Only					
	Test		Clear			Foggy			Clear			Foggy		
	IoU		0.5	0.65	0.8	0.5	0.65	0.8	0.5	0.65	0.8	0.5	0.65	0.8
DEF (CVPR 2020 [51])		L+R	86.6	78.2	46.2	81.4	72.5	41.1	85.9	78.1	44.2	71.8	63.7	32.4
MVDNet (CVPR 2021 [27])		L+R	90.9	88.8	74.6	87.4	84.6	68.9	87.2	86.1	72.6	78.0	75.9	61.6
TransFusion (T-ITS 2024 [53])		L+R	91.8	89.6	75.4	88.6	85.3	70.0	88.2	87.2	74.0	79.1	76.3	63.2
Bi-LRFusion (CVPR 2023 [60])		L+R	92.2	-	74.4	-	-	-	-	-	-	-	-	-
ST-MVDNet (CVPR 2022 [28])		L+R	94.7	93.5	80.7	91.8	88.3	73.6	91.4	89.9	78.4	81.2	80.8	64.9
ST-MVDNet++ (ICASSP 2023 [58])		L+R	96.0	94.7	82.0	93.4	90.0	75.1	93.2	92.0	81.5	83.7	83.2	67.5
RaLiBEV with GACHIPS (Ours)		L+R	97.7	97.7	93.9	97.8	96.7	93.7	97.4	96.7	93.9	84.6	82.4	73.8
RaLiBEV with DQMITBF (Ours)		L+R	98.8	97.7	95.1	97.9	97.8	94.1	98.8	97.7	94.8	83.3	80.3	70.6

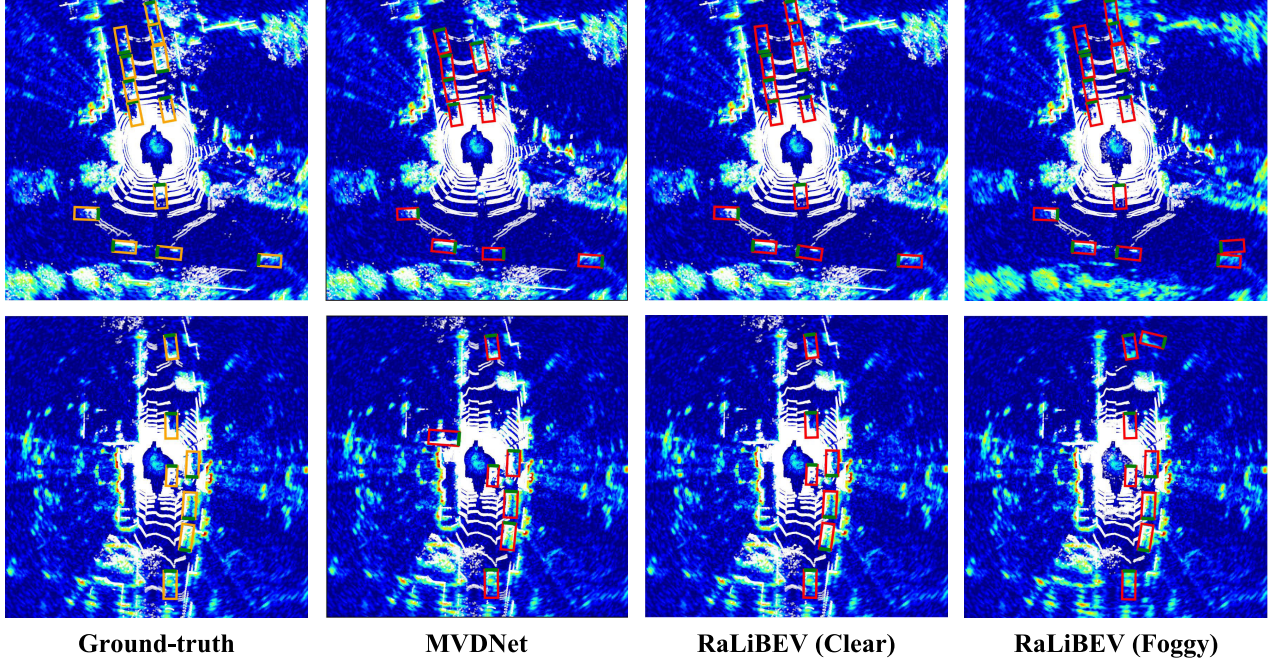


Fig. 6. Visualization of detection results of MVDNet and RaLiBEV.

Fig. 6 compares ground-truth and predictions from MVDNet and RaLiBEV under clear and foggy conditions at the 3332th and 4103th frames. Ground-truth boxes are in yellow, predictions in red, with green lines for object headings. MVDNet's predictions show 3 misses and 4 incorrect headings in the first row, and 1 false alarm with 2 heading errors in the second row, even in clear weather. Conversely, RaLiBEV achieves perfect detection in clear conditions and only one false detection without heading mistakes in fog, for both frames.

V. CONCLUSION

In this paper, we introduce RaLiBEV, an innovative method for fusing LiDAR and radar data. According to the experiment results, this method effectively showcases the significant role that radar can play in fusing autonomous driving perception technology, especially in adverse weather. To enhance the performance of the model, especially for high-precision detection, novel label assignment and transformer-based fusion strategies are proposed. Specifically, the GACHIPS strategy is proposed

to dynamically adjust the classification and regression positive sample position, to solve the inconsistency problem in the loss procedure. Meanwhile, the DQMITBF is proposed to dynamically aggregate weighted features from different sensor data. Thanks to the interaction of the dense learnable BEV query with radar and LiDAR in a symmetrical manner, RaLiBEV can switch the focus on different sensor modalities adaptively, according to the change of the environment and weather conditions. Notably, incorporating the GACHIPS label assignment into a simpler fusion architecture enables the model to achieve performance comparable to more complex transformer-based fusion methods. Experiments on the ORR dataset demonstrate that RaLiBEV significantly outperforms existing state-of-the-art approaches.

While our method performs exceptionally well on the given dataset with simulated foggy environments, it is important to acknowledge the gap between simulated and real-world weather conditions, which are often more complex and challenging. To further advance such algorithms, collecting and annotating real-world data under adverse weather conditions

is crucial. Additionally, future research could focus on integrating multimodal sensor information, such as 3D spatial structure, geometric layout, and semantic information, to enhance driving safety in poor weather. To improve model efficiency, sparse computing techniques can be employed. Replacing the dense BEV queries in the current fusion architecture with sparse queries implemented through a deformable transformer process can significantly reduce computational overhead while maintaining focus on critical information during fusion. Dynamic label assignment also offers potential for refinement by progressively reducing the number of candidate anchor points during training. Furthermore, the robustness of the fusion module can be enhanced through closed-loop testing. Similarly, the robustness of the label assignment module can be improved by incorporating multiple object classes, thereby enhancing performance across diverse scenarios. These aspects represent important directions for future exploration.

REFERENCES

- [1] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The Oxford radar RobotCar dataset: A radar extension to the Oxford RobotCar dataset," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Aug. 2020, pp. 6433–6438.
- [2] A. Li, S. Chen, L. Sun, N. Zheng, M. Tomizuka, and W. Zhan, "SceGene: Bio-inspired traffic scenario generation for autonomous driving testing," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14859–14874, Sep. 2022.
- [3] Z. Wang, L. Zhang, S. Zhao, and Y. Zhou, "Global localization in large-scale point clouds via roll-pitch-yaw invariant place recognition and low-overlap global registration," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 5, pp. 3846–3859, May 2024.
- [4] D. J. Fremont et al., "Formal scenario-based testing of autonomous vehicles: From simulation to the real world," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–8.
- [5] X. Zhu et al., "Interaction-aware cut-in trajectory prediction and risk assessment in mixed traffic," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 10, pp. 1752–1762, Oct. 2022.
- [6] L. Chen, X. Hu, W. Tian, H. Wang, D. Cao, and F.-Y. Wang, "Parallel planning: A new motion planning framework for autonomous driving," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 1, pp. 236–246, Jan. 2019.
- [7] S. Chen, B. Liu, C. Feng, C. Vallespi-Gonzalez, and C. Wellington, "3D point cloud processing and learning for autonomous driving: Impacting map creation, localization, and perception," *IEEE Signal Process. Mag.*, vol. 38, no. 1, pp. 68–86, Jan. 2021.
- [8] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11784–11793.
- [9] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12697–12705.
- [10] J. Liu, L. Bai, Y. Xia, T. Huang, B. Zhu, and Q.-L. Han, "GNN-PMB: A simple but effective online 3D multi-object tracker without bells and whistles," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 2, pp. 1176–1189, Feb. 2023.
- [11] G. Ding, J. Liu, Y. Xia, T. Huang, B. Zhu, and J. Sun, "LiDAR point cloud-based multiple vehicle tracking with probabilistic measurement-region association," 2024, *arXiv:2403.06423*.
- [12] Y. Yuan et al., "Learning discriminative features via multi-hierarchical mutual information for unsupervised point cloud registration," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 9, pp. 8343–8354, Sep. 2024.
- [13] Y. Yuan et al., "Learning compact transformation based on dual quaternion for point cloud registration," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–12, 2024.
- [14] Y. Yuan, Y. Wu, X. Fan, M. Gong, W. Ma, and Q. Miao, "EGST: Enhanced geometric structure transformer for point cloud registration," *IEEE Trans. Vis. Comput. Graphics*, vol. 30, no. 9, pp. 6222–6234, Sep. 2024.
- [15] L. Wang et al., "Multi-modal 3D object detection in autonomous driving: A survey and taxonomy," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 7, pp. 3781–3798, Jul. 2023.
- [16] X. Chang, H. Pan, W. Sun, and H. Gao, "A multi-phase camera-LiDAR fusion network for 3D semantic segmentation with weak supervision," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3737–3746, Aug. 2023.
- [17] Z. Lu, B. Cao, and Q. Hu, "LiDAR-camera continuous fusion in voxelized grid for semantic scene completion," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Jul. 29, 2024, doi: [10.1109/TCSVT.2024.3435045](https://doi.org/10.1109/TCSVT.2024.3435045).
- [18] A. Singha and M. K. Bhowmik, "Salient features for moving object detection in adverse weather conditions during night time," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3317–3331, Oct. 2020.
- [19] M. Dreissig, D. Scheuble, F. Piewak, and J. Boedecker, "Survey on LiDAR perception in adverse weather conditions," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2023, pp. 1–8.
- [20] A. Chen, X. Wang, K. Shi, Y. Huo, J. Chen, and Q. Ye, "Towards weather-robust 3D human body reconstruction: Millimeter-wave radar-based dataset, benchmark, and multi-modal fusion," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Sep. 16, 2024, doi: [10.1109/TCSVT.2024.3461960](https://doi.org/10.1109/TCSVT.2024.3461960).
- [21] S. Sun, A. P. Petropulu, and H. V. Poor, "MIMO radar for advanced driver-assistance systems and autonomous driving: Advantages and challenges," *IEEE Signal Process. Mag.*, vol. 37, no. 4, pp. 98–117, Jul. 2020.
- [22] T. Huang et al., "V2X cooperative perception for autonomous driving: Recent advances and challenges," 2023, *arXiv:2310.03525*.
- [23] J. Liu, Q. Zhao, W. Xiong, T. Huang, Q.-L. Han, and B. Zhu, "SMURF: Spatial multi-representation fusion for 3D object detection with 4D imaging radar," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 799–812, Jan. 2024.
- [24] L. Zheng et al., "RCFusion: Fusing 4-D radar and camera with bird's-eye view features for 3-D object detection," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–14, 2023.
- [25] W. Xiong, J. Liu, T. Huang, Q.-L. Han, Y. Xia, and B. Zhu, "LXL: LiDAR excluded lean 3D object detection with 4D imaging radar and camera fusion," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 79–92, Jan. 2024.
- [26] J. Liu et al., "Which framework is suitable for online 3D multi-object tracking for autonomous driving with automotive 4D imaging radar?" 2023, *arXiv:2309.06036*.
- [27] K. Qian, S. Zhu, X. Zhang, and L. E. Li, "Robust multimodal vehicle detection in foggy weather using complementary LiDAR and radar signals," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 444–453.
- [28] Y.-J. Li, J. Park, M. O'Toole, and K. Kitani, "Modality-agnostic learning for radar-LiDAR fusion in vehicle detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 918–927.
- [29] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.
- [30] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5099–5108.
- [31] D. Maturana and S. Scherer, "VoxNet: A 3D convolutional neural network for real-time object recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 922–928.
- [32] C. He, R. Li, S. Li, and L. Zhang, "Voxel set transformer: A set-to-set approach to 3D object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8417–8427.
- [33] W. Shi and R. Rajkumar, "Point-GNN: Graph neural network for 3D object detection in a point cloud," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1711–1719.
- [34] J. S. K. Hu, T. Kuai, and S. L. Waslander, "Point density-aware voxels for LiDAR 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8469–8478.
- [35] A. Laddha, S. Gautam, S. Palombo, S. Pandey, and C. Vallespi-Gonzalez, "MVFuseNet: Improving end-to-end object detection and motion forecasting through multi-view fusion of LiDAR data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2859–2868.

- [36] M. Kronauge and H. Rohling, "Fast two-dimensional CFAR procedure," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 49, no. 3, pp. 1817–1823, Jul. 2013.
- [37] S. Lim, S. Lee, and S.-C. Kim, "Clustering of detected targets using DBSCAN in automotive radar systems," in *Proc. 19th Int. Radar Symp. (IRS)*, Jun. 2018, pp. 1–7.
- [38] J. Liu et al., "Deep instance segmentation with automotive radar detection points," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 1, pp. 84–94, Jan. 2023.
- [39] W. Xiong, J. Liu, Y. Xia, T. Huang, B. Zhu, and W. Xiang, "Contrastive learning for automotive mmWave radar detection points based instance segmentation," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2022, pp. 1255–1261.
- [40] A. Danzer, T. Griebel, M. Bach, and K. Dietmayer, "2D car detection in radar data with PointNets," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 61–66.
- [41] M. Dreher, E. Erçelik, T. Bänziger, and A. Knoll, "Radar-based 2D car detection using deep neural networks," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–8.
- [42] P. Svenningsson, F. Fioranelli, and A. Yaroyov, "Radar-PointGNN: Graph based object recognition for unstructured radar point-cloud data," in *Proc. IEEE Radar Conf.*, May 2021, pp. 1–6.
- [43] M. Meyer, G. Kusch, and S. Tomforde, "Graph convolutional networks for 3D object detection on radar data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3053–3062.
- [44] K. Patel, K. Rambach, T. Visentin, D. Rusev, M. Pfeiffer, and A. B. Yang, "Deep learning-based object classification on automotive radar spectra," in *Proc. IEEE Radar Conf.*, Apr. 2019, pp. 1–6.
- [45] A.-E. Cozma, L. Morgan, M. Stolz, D. Stoeckel, and K. Rambach, "DeepHybrid: Deep learning on automotive radar spectra and reflections for object classification," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 2682–2687.
- [46] B. Major et al., "Vehicle detection with automotive radar using deep learning on range-azimuth-Doppler tensors," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 924–932.
- [47] C. Decourt, R. VanRullen, D. Salle, and T. Oberlin, "DAROD: A deep automotive radar object detector on range-Doppler maps," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2022, pp. 112–118.
- [48] C. Decourt, R. VanRullen, D. Salle, and T. Oberlin, "A recurrent CNN for online object detection on raw radar frames," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 13432–13441, Oct. 2024, doi: [10.1109/TITS.2024.3404076](https://doi.org/10.1109/TITS.2024.3404076).
- [49] P. Li, P. Wang, K. Berntorp, and H. Liu, "Exploiting temporal relations on radar perception for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17071–17080.
- [50] T.-Y. Huang, M.-C. Lee, C.-H. Yang, and T.-S. Lee, "YOLO-ORE: A deep learning-aided object recognition approach for radar systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 5, pp. 5715–5731, May 2023.
- [51] M. Bijelic et al., "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11679–11689.
- [52] Y. Chae, H. Kim, and K.-J. Yoon, "Towards robust 3D object detection with LiDAR and 4D radar fusion in various weather conditions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 15162–15172.
- [53] C. Zhang, H. Wang, Y. Cai, L. Chen, and Y. Li, "TransFusion: Multi-modal robust fusion for 3D object detection in foggy weather based on spatial vision transformer," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 9, pp. 10652–10666, Sep. 2024, doi: [10.1109/TITS.2024.3420432](https://doi.org/10.1109/TITS.2024.3420432).
- [54] L. Wang et al., "InterFusion: Interaction-based 4D radar and LiDAR fusion for 3D object detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2022, pp. 12247–12253.
- [55] L. Wang et al., "Multi-modal and multi-scale fusion 3D object detection of 4D radar and LiDAR for autonomous driving," *IEEE Trans. Veh. Technol.*, vol. 72, no. 5, pp. 5628–5641, May 2023.
- [56] S. Chen, L. Zhang, and L. Zhang, "Cyclic cross-modality interaction for hyperspectral and multispectral image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Sep. 16, 2024, doi: [10.1109/TCSVT.2024.3461829](https://doi.org/10.1109/TCSVT.2024.3461829).
- [57] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.
- [58] Y.-J. Li, M. O'Toole, and K. Kitani, "ST-MVDNet++: Improve vehicle detection with LiDAR-radar geometrical augmentation via self-training," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
- [59] M. Chen, Z. Zheng, and Y. Yang, "Transferring to real-world layouts: A depth-aware framework for scene adaptation," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 399–408.
- [60] Y. Wang et al., "Bi-LRFusion: Bi-directional LiDAR-RADAR fusion for 3D dynamic object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 13394–13403.
- [61] S. Li, C. He, R. Li, and L. Zhang, "A dual weighting label assignment scheme for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9387–9396.
- [62] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
- [63] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
- [64] W. Xiao, Y. Peng, C. Liu, J. Gao, Y. Wu, and X. Li, "Balanced sample assignment and objective for single-model multi-class 3D object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 5036–5048, Sep. 2023.
- [65] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9626–9635.
- [66] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6568–6577.
- [67] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 642–656.
- [68] X. Zhang, F. Wan, C. Liu, R. Ji, and Q. Ye, "FreeAnchor: Learning to match anchors for visual object detection," in *Proc. Neural Inf. Process. Syst. (NIPS)*, Dec. 2019, pp. 147–155.
- [69] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9759–9768.
- [70] B. Zhu et al., "AutoAssign: Differentiable label assignment for dense object detection," 2020, *arXiv:2007.03496*.
- [71] C. H. Nguyen, T. C. Nguyen, T. N. Tang, and N. L. H. Phan, "Improving object detection by label assignment distillation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1322–1331.
- [72] Z. Ge, S. Liu, Z. Li, O. Yoshie, and J. Sun, "OTA: Optimal transport assignment for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 303–312.
- [73] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
- [74] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
- [75] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2155–2162.
- [76] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.



Yanlong Yang (Member, IEEE) received the B.Eng. degree in electronic science and technology (opto-electronics) from Tianjin University, Tianjin, China, in 2017, and the M.Eng. degree in optical engineering from Beijing Institute of Technology, China, in 2020. He is currently pursuing the Ph.D. degree with the College of Science and Engineering, James Cook University, Cairns, Australia. Since July 2020, he has been employed as an Algorithm Engineer with Beijing Vanjee Technology. His research interests include perception, radar/LiDAR sensing, and artificial intelligence.



Jianan Liu received the B.Eng. degree in electronics and information engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2007, the M.Eng. degree in telecommunication engineering from the University of Melbourne, Australia, in 2009, and the M.Sc. degree in communication systems from Lund University, Sweden, in 2012. He has over ten years of experience in software and algorithm design and development. He has held senior research and development roles in the AI consulting, automotive,

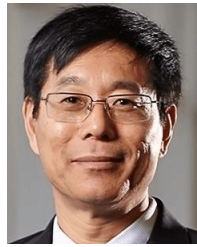
and telecommunication industries. His research interests include applying statistical signal processing and deep learning for medical image processing, wireless communications, the IoT networks, indoor sensing, and outdoor perception using a variety of sensor modalities, such as radar, camera, LiDAR, and WiFi.



Tao Huang (Senior Member, IEEE) received the B.Eng. degree in electronics and information engineering from the Huazhong University of Science and Technology, China, in 2003, the M.Eng. degree in sensor system signal processing from The University of Adelaide, Australia, in 2007, and the Ph.D. degree in electrical engineering from The University of New South Wales, Australia, in 2016.

He was an Endeavour Australia Cheung Kong Research Fellow, a Visiting Scholar with The Chinese University of Hong Kong, a Research Associate with The University of New South Wales, and a Post-Doctoral Research Fellow with James Cook University, Cairns, Australia. Before academia, he worked in the industry and held positions, such as a Senior Software Engineer, a Senior Data Scientist, the Project Team Lead, and the Technical Lead. He is a Senior Lecturer with James Cook University. He is the co-inventor of an international patent on MIMO systems. His research interests include electronics systems, deep learning, intelligent sensing, computer vision, pattern recognition, wireless communications, and IoT security.

Dr. Huang is a member of the IEEE Industrial Electronics Society, the IEEE Communication Society, the IEEE Vehicular Technology Society, and the IEEE Computational Intelligence Society. He received the Australian Postgraduate Award, the Engineering Research Award at The University of New South Wales, the Best Paper Award from IEEE WCNC, the IEEE Outstanding Leadership Award, and the JCU Citation for Outstanding Contribution to Student Learning. He serves as the MTT-S/ComSoc Chapter Chair and the Vice Chair for the IEEE Northern Australia Section. He served as the TPC chair, the program vice chair, and the symposium chair at international conferences. He is an Associate Editor of IEEE OPEN JOURNAL OF COMMUNICATIONS SOCIETY, IEEE ACCESS, and *IET Communications*. He is an IEEE Access Outstanding Associate Editor of 2023.



Qing-Long Han (Fellow, IEEE) received the B.Sc. degree in mathematics from Shandong Normal University, Jinan, China, in 1983, and the M.Sc. and Ph.D. degrees in control engineering from the East China University of Science and Technology, Shanghai, China, in 1992 and 1997, respectively.

He is currently the Pro Vice-Chancellor (Research Quality) and a Distinguished Professor with the Swinburne University of Technology, Melbourne, Australia. He held various academic and management positions with Griffith University and Central Queensland University, Australia. His research interests include networked control systems, multi-agent systems, time-delay systems, smart grids, unmanned surface vehicles, and neural networks.

Prof. Han is a member of the Academia Europaea (The Academy of Europe). He is a fellow of The International Federation of Automatic Control (IFAC), an Honorary Fellow of The Institution of Engineers Australia (HonFIEAust), and a fellow of Chinese Association of Automation (FCAA). He was awarded the 2024 IEEE Dr.-Ing. Eugene Mittelmann Achievement Award (the Highest Award in Industrial Electronics), the 2021 Norbert Wiener Award (the Highest Award in systems science and engineering, and cybernetics), and the 2021 M. A. Sargent Medal (the Highest Award of the Electrical College Board of Engineers Australia). He was a recipient of the IEEE Systems, Man, and Cybernetics Society Andrew P. Sage Best Transactions Paper Award in 2022, 2020, and 2019, the IEEE/CAA Journal of Automatica Sinica Norbert Wiener Review Award in 2020, and the IEEE Transactions on Industrial Informatics Outstanding Paper Award in 2020. He is a Highly Cited Researcher in both Engineering and Computer Science (Clarivate). He has served as an AdCom Member for the IEEE Industrial Electronics Society (IES), a member for the IEEE IES Fellows Committee, a member for the IEEE IES Publications Committee, the Chair for the IEEE IES Technical Committee on Network-Based Control Systems and Applications, and the Co-Editor-in-Chief for IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS. He is the Editor-in-Chief of IEEE/CAA JOURNAL OF AUTOMATICA SINICA and the Co-Editor of *Australian Journal of Electrical and Electronic Engineering*.



Gang Ma received the B.Eng. degree in materials science and engineering and the M.Eng. degree in computer measurement and control of material processing from Beihang University, Beijing, China, in 2005 and 2008, respectively. He is currently a Senior Hardware Engineer with rich experience in LiDAR product development.



Bing Zhu (Senior Member, IEEE) was born in Wuhan, China. He received the B.S. and Ph.D. degrees in control theory and applications from Beihang University, Beijing, China, in 2007 and 2013, respectively. He was with the University of Pretoria, Pretoria, South Africa, as a Postdoctoral Fellow supported by Vice-Chancellor Postdoctoral Fellowship from 2013 to 2015. He was with Nanyang Technological University, Singapore, as a Research Fellow, from 2015 to 2016. He joined Beihang University, Beijing, as an Associate Professor, in 2016. His research interests include model predictive control, smart sensing for UAV and UGV, and demand-side management for new energy systems. He serves as an Associate Editor for *Acta Automatica Sinica*.