

Article

RCRFNet: Enhancing Object Detection with Self-Supervised Radar–Camera Fusion and Open-Set Recognition

Minwei Chen ¹, Yajun Liu ¹, Zenghui Zhang ^{1,*}  and Weiwei Guo ² ¹ Shanghai Key Laboratory of Intelligent Sensing and Recognition, Shanghai Jiao Tong University, Shanghai 200240, China; chenminwei2022@sjtu.edu.cn (M.C.); liuyajun@sjtu.edu.cn (Y.L.)² Center of Digital Innovation, Tongji University, Shanghai 200092, China; weiweiguo@tongji.edu.cn

* Correspondence: zenghui.zhang@sjtu.edu.cn

Abstract: Robust object detection in complex environments, poor visual conditions, and open scenarios presents significant technical challenges in autonomous driving. These challenges necessitate the development of advanced fusion methods for millimeter-wave (mmWave) radar point cloud data and visual images. To address these issues, this paper proposes a radar–camera robust fusion network (RCRFNet), which leverages self-supervised learning and open-set recognition to effectively utilise the complementary information from both sensors. Specifically, the network uses matched radar–camera data through a frustum association approach to generate self-supervised signals, enhancing network training. The integration of global and local depth consistencies between radar point clouds and visual images, along with image features, helps construct object class confidence levels for detecting unknown targets. Additionally, these techniques are combined with a multi-layer feature extraction backbone and a multimodal feature detection head to achieve robust object detection. Experiments on the nuScenes public dataset demonstrate that RCRFNet outperforms state-of-the-art (SOTA) methods, particularly in conditions of low visual visibility and when detecting unknown class objects.

Keywords: radar–camera fusion; target detection; self-supervised learning; open-set recognition; autonomous driving



Citation: Chen, M.; Liu, Y.; Zhang, Z.; Guo, W. RCRFNet: Enhancing Object Detection with Self-Supervised Radar–Camera Fusion and Open-Set Recognition. *Sensors* **2024**, *24*, 4803. <https://doi.org/10.3390/s24154803>

Academic Editor: Bogdan Smolka

Received: 24 June 2024

Revised: 18 July 2024

Accepted: 22 July 2024

Published: 24 July 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Autonomous driving technology has seen rapid advancements in recent years, focusing mainly on three critical components: environment perception, decision-making, and control [1]. Environment perception is the foundation of autonomous driving, performing functions such as 2D and 3D object detection, depth estimation, and prediction, similar to how human eyes work. This perception relies on real-time data gathered by onboard sensors. However, the data obtained from a single sensor are often insufficient for the complex tasks required for environment perception.

Common onboard sensors include vision cameras, LiDAR, and millimeter-wave (mmWave) radar. Vision cameras, while providing rich geometric information, are significantly impacted by lighting conditions and have a reliable detection distance limited to 50 m. LiDAR sensors provide high-resolution 3D mapping and precise distance measurements, making them excellent for detailed environmental modelling, but they are costly and less effective in bad weather [2]. In contrast, mmWave radar emits active signals to measure the reflective information of objects. Although it offers less geometric detail, it provides a reliable detection distance of up to 250 m and has strong penetrability, making it ideal for harsh environments. Consequently, mmWave radar has become an indispensable choice for current automated driving applications.

Adverse weather conditions, including heavy rain and fog, present significant challenges to visual detection technology, leading to reduced object detection performance in

models such as SSD [3], DETR [4], and Faster R-CNN [5]. Additionally, the diversity of objects in real-world scenes often surpasses the categories covered by detection model training data, resulting in a model mismatch problem [6]. Recently, there has been increasing interest in open-set detection and recognition methodologies [7,8]. In [9], a deep learning approach is proposed for open-set recognition, which involves fitting extreme value distributions specific to each category. This method utilizes the class probabilities derived from these distributions to identify open-set images. In the realm of video action recognition, ref. [10] introduces deep evidence learning grounded in evidence theory, effectively addressing classification uncertainty in video actions. Moreover, ref. [11] introduces a reconstruction learning algorithm aimed at enhancing the robustness of detecting unknown classes.

Radar can reliably detect a wide range of object categories, including those not present in the training set, making it useful for detecting obstacles outside the training categories. As a result, the fusion of radar and camera technologies has gained considerable attention in recent years, aiming to leverage the strengths of both sensors. Millimeter-wave radar point clouds offer precise measurements of object distance and velocity [2]. However, due to their low angular resolution and sparsity, they cannot independently perform complex object detection tasks or accurately identify an object's category and geometric information. Traditional mmWave radar point cloud processing methods, such as distance-FFT, Doppler-FFT, incoherent accumulation, and angle estimation, have limitations in achieving comprehensive object detection. Recent advancements include Radar-pointGNN [12], which uses graph neural networks for object detection, and methods that utilise raw radar tensor data for 3D object detection through graph neural networks [13]. However, in order to extract features, these algorithms largely rely on large amounts of training data and prior knowledge. This poses difficulties when fusing with image data, which results in reduced robustness and increased complexity. Consequently, radar and vision machine learning fusion detection methods have gained increasing attention from researchers. These approaches aim to integrate the strengths of both radar and vision systems to enhance detection capabilities and overcome the limitations of individual sensor modalities.

Currently, there are three primary approaches for multimodal fusion [14]: early fusion (data-level) methods [15], intermediate fusion (feature-level) methods [16–19], and late fusion (decision-level) methods [20,21]. In data-level fusion, the region of interest (ROI) is first generated based on radar points [22]. Subsequently, the corresponding region in the visual image is extracted using this ROI. Finally, object detection is conducted on these images using a feature extractor and classifier. For instance, in [23], a fusion of the mmWave radar and camera vision is proposed for pedestrian tracking, whereby the size of the initial ROI is determined by the distance between the obstacle and mmWave radar. In [24], Kadow et al. applied the Haar-like model for feature extraction, which is a classic feature extraction algorithm for face detection. Data-level fusion operates at the most basic level of information, making it vulnerable to uncertainties, incompleteness, and instability in sensor data [25]. Feature-level fusion methods, offering greater flexibility, typically involve layer-by-layer feature extraction and fusion. For instance, in [26], mmWave radar is integrated into visual detection models using frustum correlation to associate heterogeneous data sources. In [27], the authors proposed a transformer-based approach for LiDAR–camera fusion, enhancing robustness against degraded images and sensor misalignment through soft correlation strategies. In [18], to tackle issues such as low angular resolution in mmWave radar, which complicates distinguishing radial objects and leads to false-positive ghost points due to multi-path interference, radar data are augmented with image data containing semantic features. These augmented radar data are then fused by transforming the prediction frame of image modalities into polar coordinate systems. The paper [28] proposes a Spatial Attention Fusion (SAF) module for sensor feature fusion, demonstrating that fusing mmWave radar point clouds significantly enhances detection robustness and improves performance in various weather conditions. Decision-level fusion involves each sensor processing its data independently and making decisions, followed by transmitting the results to a fusion centre for final decision-making; this includes two

steps: sensing information processing [29] and decision fusion [30]. For instance, some researches have utilised lists of radar detection targets to validate the results of vision-based detection [31]. Additionally, reference [30] introduced the motion stereo algorithm to further adjust and refine the final detection outcomes. Mainstream traditional methods for decision-level fusion include D-S evidence theory, Bayesian reasoning, and fuzzy inference theory, among others. Although most fusion detection models demonstrate improved accuracy compared to single-sensor models, they still encounter difficulties in complex environments, poor visual conditions, and open scenarios.

To address the aforementioned challenges, this paper presents a novel radar–camera fusion detection method. The key contributions are as follows:

(1) Development of a new fusion detection framework with parallel inputs: This framework enhances feature extraction from visual images and radar point clouds through multiple modules. It aims to generate effective data features for robust detection, even in adverse weather conditions like heavy rain and fog.

(2) Introduction of a novel data correlation and self-supervised learning technique: This method effectively integrates radar point clouds and visual images, leveraging their complementary strengths to enhance the model’s robustness. It incorporates an attention mechanism and self-supervised learning for efficient fusion.

(3) Implementation of an unknown category recognition method with confidence assessment: This approach utilizes depth estimation from radar point clouds and visual images to improve the detection and recognition of open-set categories. By leveraging depth information, the model enhances its capability to identify unknown categories while also improving the classification of known categories.

The rest of this paper is organised as follows. Section 2 presents the fusion detection framework, emphasizing robust feature extraction, radar–camera correlation, self-supervised learning, and open-set recognition. Section 3 conducts experimental validation and comparative analyses. Section 4 provides the ablation study of each proposed module. Finally, Section 5 concludes the entire paper.

2. Materials and Methods

2.1. Overall Network Structure

To enhance object detection performance in low-visibility and open conditions, we propose a novel deep learning-based radar camera robust feature fusion network (RCRFNet). This network effectively leverages the complementary information from both millimeter-wave radar data and visual camera images.

The overall architecture of the RCRFNet, depicted in Figure 1, consists of two parallel sub-networks for processing visual images and radar cloud data, respectively. Additionally, it features the following modules: the feature extraction backbone, the feature association and self-supervised learning module, the radar–camera open-set recognition module, and the multimodal feature detection head.

The feature extraction backbone: It extracts multilevel features from both radar point cloud data and visual images by projecting radar data onto pixel coordinates to obtain multidimensional data. Subsequently, a multi-layer fusion structure, consisting of upsampling and downsampling layers, is utilised to extract high-dimensional features and align data at various resolutions.

The feature association and self-supervised learning module: It innovatively selects high-fidelity radar–camera data pairs to generate self-supervised signals to boost the network’s performance by means of a frustum association process applied to the heatmaps of both radar data and visual images. Furthermore, spatial attention and channel attention mechanisms are introduced to enhance the heatmap association.

The radar–camera open-set recognition module: It is designed to improve object classification accuracy, especially for unknown categories.

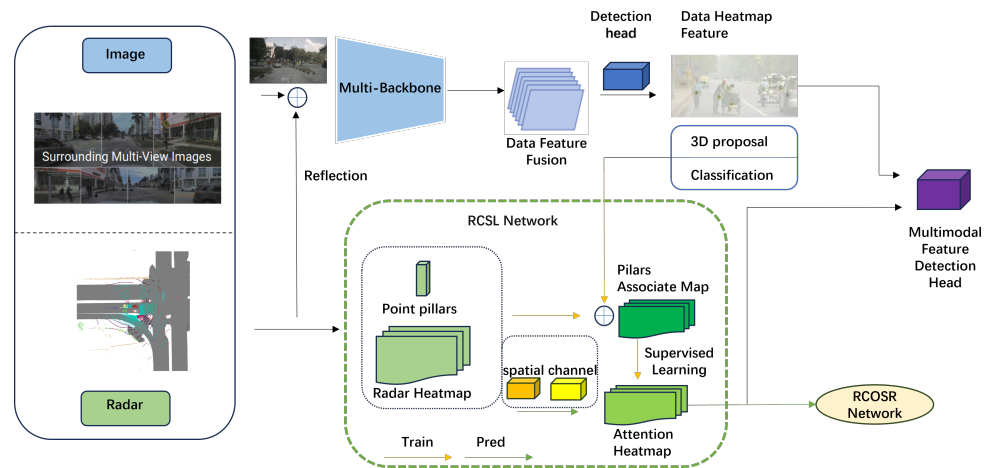


Figure 1. The overall architecture of the proposed RCRF network.

The multimodal feature detection head: Finally, the fused multilevel features are linearly transformed to the same size for splicing, generating the final detection results through the constructed multimodal feature detector.

2.2. Feature Extraction Backbone and Heatmap Generation

We introduce a multi-layer fusion structure to merge different layers of radar and visual features, aiming to enhance the detector’s capability to detect targets across various scales. It is based on the deep layer aggregation (DLA) backbone from [32] and incorporates deformable convolution layers to capture the geometric deformations of targets. The designed Deformable DLA (DDLA) module is depicted in Figure 2.

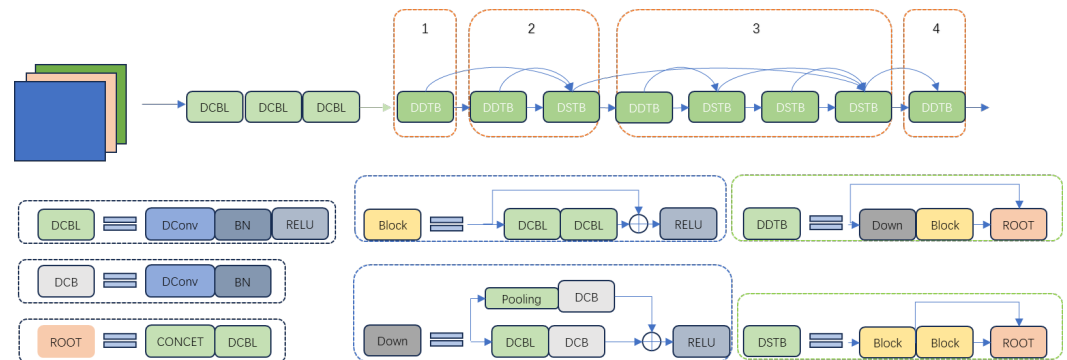


Figure 2. The designed DDLA feature extraction backbone. DConv denotes the deformable convolution layer.

The DLA34 network is a deep residual network [33] that utilizes a multi-layer hierarchical fusion structure. Its two-dimensional structure is made up of the hierarchical deep aggregation module, which preserves and combines feature channels by joining blocks from different stages into a tree. As a result, the network can integrate both deep and shallow feature information, producing more feature hierarchies and richer combinations. The adopted DLA34 in Figure 2 first goes through a base layer, and then the four tree hierarchies from level 1 to level 4. However, the existing DLA34 relies on standard $n \times n$ convolution kernels for feature extraction, which may not effectively capture geometric deformations of objects [34]. To address this deficiency, we incorporate deformable convolution layers into the DLA34 backbone, introducing additional parameter directions and expanding the learning range.

Deformable convolution layers integrate extra offsets, enhancing the extraction of features for irregular objects. Figure 3 demonstrates the contrast between deformable convolution and traditional convolution, with arrows indicating the additional offsets

introduced. This shows that deformable convolution can represent object features more efficiently. The deformable convolution can be expressed as follows:

$$\beta(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in R} \mathbf{w}(\mathbf{p}_n) \cdot \alpha(\mathbf{p}_0 + \mathbf{p}_n + \Delta \mathbf{p}_n) \quad (1)$$

where $\mathbf{w}(\mathbf{p}_n)$ is the convolution layer weight, α is the feature map, $\mathbf{p}_0$ is any point in the feature map, $\mathbf{p}_n$ is the position vector of each point in the convolution kernel with respect to the centre, and $\Delta \mathbf{p}_n$ denotes the positional offset between the feature map and the convolution kernel. A new offset has been introduced at each point, which is generated from the input feature map and another convolution kernel.

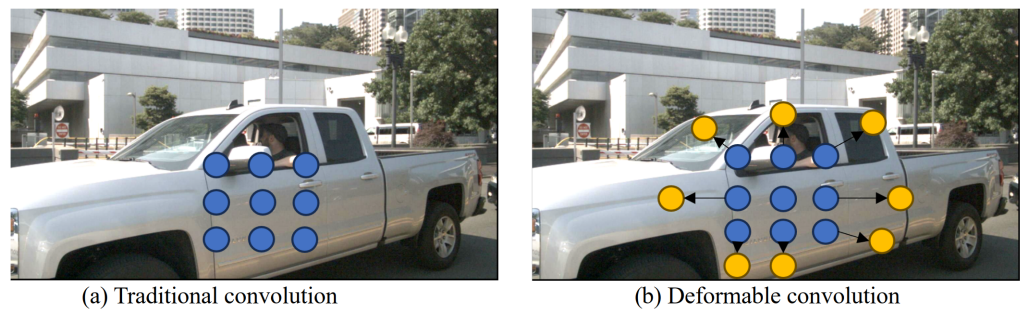


Figure 3. Contrast between (a) traditional convolution and (b) deformable convolution. The arrows in (b) indicate the additional parameter offsets introduced by deformable convolution.

The image features extracted by the DDLA module are then fed into a tentative detection head to generate visual heatmaps for subsequent processing. Suppose the tentative detector obtains a series of target candidates belonging to C categories; then, the target candidates of the same category are utilised to generate a heatmap. Thus, we have

$$H_i(m(\mathcal{P}_i)) \in R^{\frac{H}{16} \times \frac{W}{16} \times 3}, \quad i = 1, 2, \dots, C \quad (2)$$

where $\mathcal{P}_i$ denotes the set of target candidate centres corresponding to category i , and $m(\mathcal{P}_i)$ is a Gaussian kernel function applied to the set $\mathcal{P}_i$. Let $\mathcal{P}_i = \{\mathbf{p}_{i,1}, \mathbf{p}_{i,2}, \dots, \mathbf{p}_{i,i_N}\}$, where i_N is the number of target candidates belonging to category i . Then, $m(\mathcal{P}_i)$ can be expressed as follows:

$$m(\mathcal{P}_i) \triangleq m(\mathbf{x}) = \sum_{j=1}^{i_N} \exp\left(-\frac{D^2(\mathbf{x}, \mathbf{p}_{i,j})}{2\sigma_{i,j}^2}\right) \quad (3)$$

where $\mathbf{x}$ is the coordinate of any point on the i -th heatmap, $D(\mathbf{x}, \mathbf{p}_{i,j})$ is the Euclidean distance between the two points $\mathbf{x}$ and $\mathbf{p}_{i,j}$, and $\sigma_{i,j}$ is the adaptive variance set by the target radius of $\mathbf{p}_{i,j}$.

Next, we will introduce radar point cloud data filtering and heatmap generation. To address the issues of sparsity and low-resolution in radar point cloud data, we apply data filtering to enhance data density and suppress noise. Considering the targets' manoeuvring movement, we utilise the extended Kalman filter (EKF) algorithm to preprocess radar point cloud data.

Let (P_x, P_y) and (V_x, V_y) denote the position and velocity of a target measured by the mmWave radar. The measurement model can be characterised by

$$h(x) = \begin{bmatrix} R \\ \theta \\ V \end{bmatrix} = \begin{bmatrix} \sqrt{P_x^2 + P_y^2} \\ \arctan(P_y/P_x) \\ \frac{P_x V_x + P_y V_y}{\sqrt{P_x^2 + P_y^2}} \end{bmatrix} \quad (4)$$

The Jacobian matrix for the measurement model is [35]

$$J = \begin{bmatrix} \frac{P_x}{\sqrt{P_x^2 + P_y^2}} & \frac{P_y}{\sqrt{P_x^2 + P_y^2}} & 0 & 0 \\ -\frac{P_y}{P_x^2 + P_y^2} & \frac{P_x}{P_x^2 + P_y^2} & 0 & 0 \\ \frac{P_y(V_x P_y - V_y P_x)}{(P_x^2 + P_y^2)^{3/2}} & \frac{P_x(V_y P_x - V_x P_y)}{(P_x^2 + P_y^2)^{3/2}} & \frac{P_x}{\sqrt{P_x^2 + P_y^2}} & \frac{P_y}{\sqrt{P_x^2 + P_y^2}} \end{bmatrix} \quad (5)$$

In practice, we take the weighted sum of the current frame data and the Kalman filter data as the preprocessing results. Furthermore, the radar heatmap is generated using Gaussian kernels with the preprocessed point cloud as the centre and a fixed value as the radius. Each point cloud has three channels corresponding to range, azimuth angle, and radial velocity, respectively. Additionally, the point cloud is extended into a pillar with a preset value to address the association deficiency caused by missing height information.

2.3. Radar–Camera Association and Self-Supervised Learning

The visual camera and mmWave radar sensors can both offer complimentary data about the target. Visual images provide high resolution and can capture detailed features such as geometry, contours, and texture, but they are sensitive to variations in lighting conditions. Conversely, despite lower data resolution and angular accuracy, radar point cloud data can provide physical attribute information about the target, such as speed, angle, and distance, regardless of the lighting.

To leverage the complementary advantages of both sensors, this paper adopts a process where high-resolution visual image features serve as a reference, and point cloud data with high spatial location matching are then selected for correlation and fusion. In essence, correlated data obtained under favourable lighting conditions are chosen to generate self-supervised signals. Our method aims to enhance the network’s feature learning and representation capabilities by leveraging the inherent strengths of both sensor modalities.

Specifically, the frustum association approach is utilised to correlate the radar point cloud with visual images, generating the self-supervised signal for training, as depicted in Figure 4. The target’s 3D bounding box is transformed into a frustum coordinate system, with its centre and orientation angle denoted as (O_X, O_Y, O_Z, ϕ) . The orientation of the target in the bird’s eye view is also indicated by an arrow in Figure 4. The association between the radar point cloud (P_X, P_Y, P_Z) and the target’s bounding box is achieved by calculating the following Euclidean distance:

$$\rho_i = \sqrt{(P_X^2 - O_X)^2 + (P_Y^2 - O_Y)^2 + (P_Z^2 - O_Z)^2} \quad (6)$$

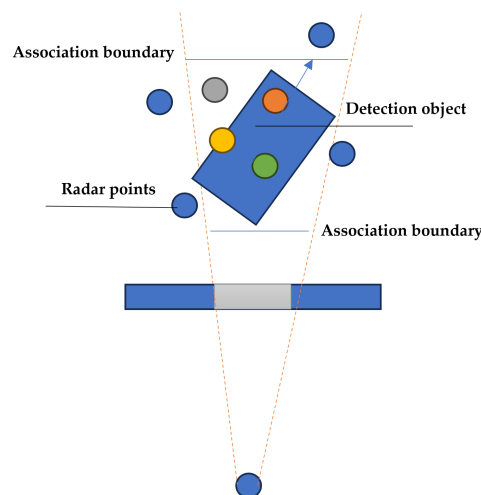


Figure 4. The frustum association approach in the bird’s eye view.

The point cloud with the lowest Euclidean distance is selected, and its heatmap will be concatenated with the image heatmap to generate the final detection results. An example of the frustum association result is shown in Figure 5. We can see that the radar point cloud is well matched with the target in the image. Here, the radar point cloud is shown in the pillar format.

It is important to emphasise that the frustum association approach heavily depends on visual detection results. When visual information is unreliable, it struggles to accurately correlate millimeter-wave radar point clouds.



Figure 5. An example of the frustum association result.

Next, the mean squared error (MSE) loss between the radar heatmap and the frustum-associated heatmap is utilised for self-supervised learning to accelerate the convergence speed of the model:

$$L_H = \frac{\sum_{x=1}^H \sum_{y=1}^W (H_o(x, y) - H_a(x, y))^2}{H \times W} \quad (7)$$

where H_o and H_a are the radar heatmaps generated by all point clouds and the associated point cloud, respectively. In the prediction phase, the associated radar heatmap is directly concatenated with the image heatmap and fed into the detection head to generate the final results.

Additionally, to enhance the quality of the associated radar heatmap, spatial attention and channel attention mechanisms are introduced, as shown in Figure 6. First, features are compressed in the spatial dimension using global average pooling to gain a larger receptive field. Subsequently, normalised weights are obtained through a fully connected layer, activation function, and normalisation operation. Finally, these weights are used to emphasise the importance of each feature channel, resulting in the attention-enhanced associated point cloud heatmap. Moreover, the radar point cloud features are used to filter noise and outliers, thereby reducing false positives during detection.

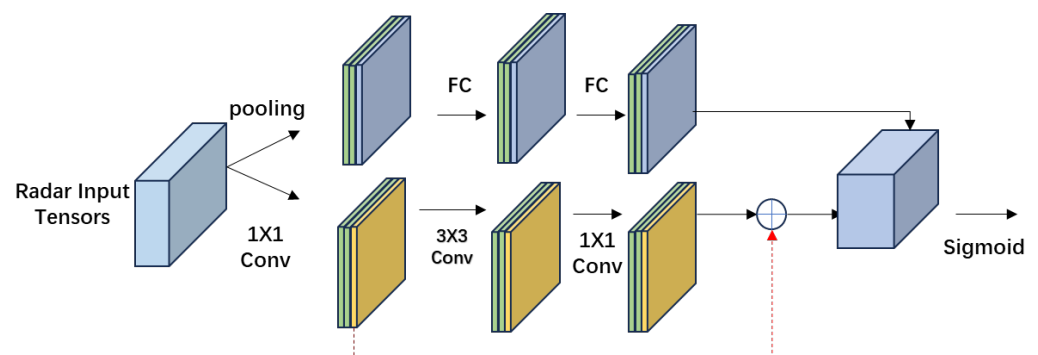


Figure 6. The attention module for enhancing the quality of the association heatmap.

2.4. Radar–Camera Open-Set Recognition Module

For the visual heatmap generation in Equation (2), it is assumed that the targets to be detected belong to C predefined categories. However, targets in real-world scenarios are diverse, and environments are complex. Thus, it is challenging to enumerate all possible target categories in advance. When a target to be detected does not fall within one of the C predefined categories, target detection performance may significantly decrease. To address this issue and improve the approach's performance in real-world applications, recently developed open-set techniques are introduced, and a radar–camera open-set recognition (RCOSR) module is designed.

The RCOSR leverages the complementary strengths of both sensors to improve the recognition accuracy of unknown class targets. Visual images, with their high resolution and detailed target information, enable highly confident target category predictions based on visual features. However, they are vulnerable to weather and lighting conditions, and the depth information they provide is often ambiguous. In contrast, radar provides more accurate depth information than visual images [36]. Therefore, we propose a novel depth-information-constrained feature aggregation method for open-set recognition, as depicted in Figure 7. Specifically, we calculate the global and local depth consistencies between the radar point cloud and visual images, which are then multiplied with visual image features to construct known target confidence levels, thus achieving unknown target detection.

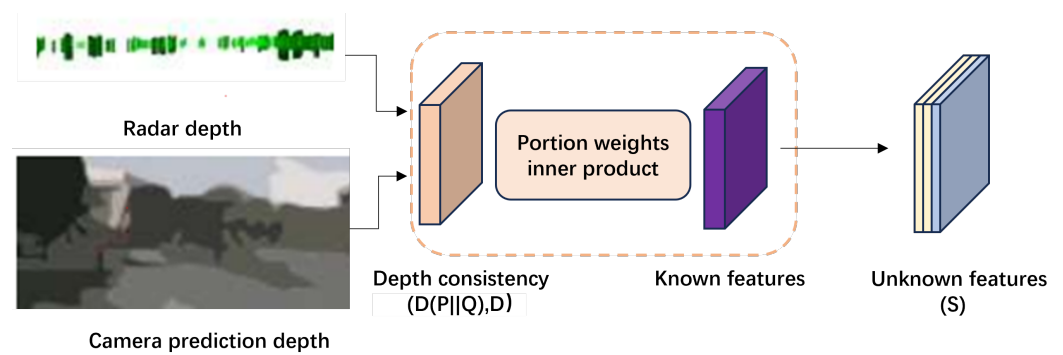


Figure 7. Radar–camera open-set recognition module.

The global depth consistency is quantified using the Kullback–Leibler (KL) divergence between the radar depth image $d_R(x, y)$ and the camera depth image $d_C(x, y)$, as shown below:

$$D(P||Q) \triangleq D(d_R||d_C) = \sum d_R(x, y) \log_2 \frac{d_R(x, y)}{d_C(x, y)}. \quad (8)$$

The local depth consistency $D(x, y)$ at the coordinate (x, y) is calculated using the following formula:

$$D(x, y) = 1 - \text{softmax}\left(\text{abs}(d_R(x, y) - d_C(x, y))\right). \quad (9)$$

Additionally, considering that the occurrence probabilities of each target category in real-world scenarios may vary significantly and affect the discrimination of unknown classes, a new weight σ_i measuring the occurrence probability is introduced. Finally, the weighted confidence level of known class targets is determined:

$$S(x, y) = \frac{1}{|C|} \cdot D(P||Q) \cdot \sum_{i=1}^C \sigma_i \cdot D(x, y). \quad (10)$$

A target candidate with $S(x, y) \leq S_t$ is classified as an unknown class target, where S_t is a threshold set according to the confidence probability.

2.5. Multimodal Feature Detection Head

The multimodal feature detection head comprises two stages. The first stage, comprising a convolution layer and an activation function, is designed to generate target proposals for radar–camera association. The second stage generates the final fusion features through point-to-point splicing matching between the fusion heatmap and the camera heatmap. The structure of the detection head is illustrated in Figure 8. This two-stage design ensures more reliable detection.

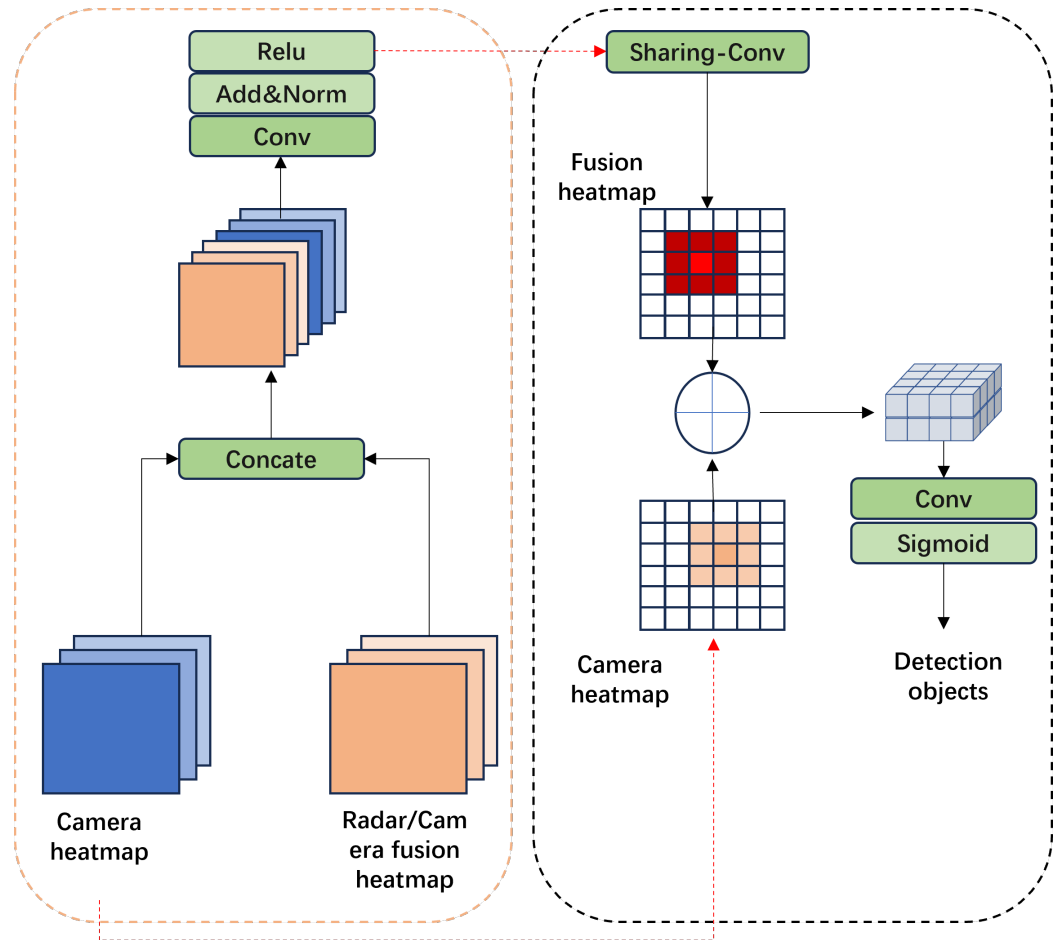


Figure 8. Multimodal feature detection head.

The loss function of the proposed model has two parts: the focal loss and the binary cross-entropy (BCE) loss, which can be expressed as

$$\begin{aligned} L_{total} &= L_{pre} + L_{fusion} \\ &= (L_{pre,cls} + L_{pre,reg}) + (L_{fusion,cls} + L_{fusion,reg}) \end{aligned} \quad (11)$$

The focal loss $L_{pre,cls}$ and $L_{fusion,cls}$ are associated with category information, while the BCE loss $L_{pre,reg}$ and $L_{fusion,reg}$ pertain to size information. This combination allows the model to effectively learn and optimise for both category and size information.

3. Experiments and Results

The experiments are divided into three sections: Section 3.1 presents the overall detection performance of the proposed approach, Section 3.2 focuses on robustness analysis, and Section 3.3 primarily verifies the open-set performance.

For comparison, three state-of-the-art (SOTA) detection algorithms for visual images—Mono3D [37], CenterNet [38], and FCOS3D [39]—as well as the latest fusion detec-

tion method, CenterFusion [18], are selected as benchmarks. For the open-set recognition module, the traditional Openmax method is chosen for comparison because current mainstream radar–camera fusion detection methods rarely address the open-set issue.

We used the nuScenes dataset [40] for the experiments. This dataset comprises data from mmWave radar, cameras, and other sensors, as detailed in Table 1. We selected data from over 1000 scenes in various environments and lighting conditions. Each scene contains 40 keyframes and 20 s of data.

Table 1. Dataset details.

Tensors	Information
Camera	50 m, 12 Hz
Radar	250 m, 13 Hz
LiDAR	70 m, 20 Hz
GPS and IMU	RTK ap: 20 mm, 1000 Hz

The experiments were conducted on an Ubuntu 18.4 system, equipped with four GeForce RTX 4080 GPUs (NVIDIA, Santa Clara, CA, USA), each with 24 GB of RAM. The Adam optimizer was used to iteratively update the network weights, with a weight decay rate set to 0.0001. Following a step learning strategy, the learning rate was halved every 20 epochs. A total of 50 epochs of training were performed. The training parameter settings are summarised in Table 2. Preprocessing provided by the nuScenes official website was applied to the dataset during the experiments.

Table 2. Training parameter settings.

Parameter	Epoch	Batch	Momenta	Learning Rate
Value	50	64	0.922	0.0001

To evaluate detection accuracy, we use the average precision (AP) metric, where a match is determined by thresholding the two-dimensional centre distance on the ground plane. The mean average precision (mAP) is calculated as follows:

$$\text{mAP} = \frac{1}{|C||M|} \sum_{c \in C} \sum_{m \in M} \text{AP}_{c,m} \quad (12)$$

where C denotes the number of predicted classes and M represents the distance threshold between the predicted bounding box and the ground truth. In this study, we set the threshold values to $M = 0.5, 1, 2, 4$ m.

Additionally, the experiments calculate five true-positive (TP) metrics: average translation error (ATE), average scale error (ASE), average orientation error (AOE), average velocity error (AVE), and average attribute error (AAE). For each of these metrics, the mean true positive (mTP) for all classes can be derived as follows:

$$\text{mTP}(x) = \frac{1}{|C|} \sum_{c \in C} (1 - \min\{1, x\}), x \in \{\text{ATE}, \text{ASE}, \text{AOE}, \text{AVE}, \text{AAE}\} \quad (13)$$

Following this, the nuScenes detection score (NDS) is computed as

$$\text{NDS} = \frac{1}{10} \left[5 \times \text{mAP} + \sum_x (1 - \min\{1, \text{mTP}(x)\}) \right] \quad (14)$$

This metric effectively encapsulates various aspects of the detection task, including velocity and attribute estimation.

3.1. Overall Detection Performance

The comparison results of our approach with SOTA algorithms are presented in Table 3, where all the aforementioned metrics are evaluated. Among these metrics, NDS and mAP are the two most representative, with higher values indicating better performance. As shown in Table 3, our approach achieves the highest NDS and mAP scores. Specifically, the NDS improves to 0.46, representing a 2.5% relative improvement over CenterFusion. Additionally, the AOE and AVE show significant relative improvements of 12.6% and 9.8%, respectively. Our approach incorporates radar information through the RCSL module, leading to a substantial enhancement in the accuracy of target speed estimation, as reflected in the AVE metrics. In summary, the proposed method effectively fuses mmWave radar point cloud data with visual images, resulting in improved overall detection performance.

Table 3. Comparison results with SOTA algorithms on the nuScenes dataset. $\uparrow$ indicates that higher is better and $\downarrow$ indicates that lower is better.

Method	Modalities	NDS $\uparrow$	mAP $\uparrow$	mATE	mASE	Error ($\downarrow$)		
						mAOE	mAVE	mAAE
Mono3D	C	0.429	0.366	0.642	0.252	0.523	1.591	0.119
CenterNet	C	0.400	0.338	0.658	0.255	0.629	1.629	0.142
FCOS3D	C	0.428	0.358	0.690	0.249	0.452	1.434	0.124
CenterFusion	R + C	0.449	0.326	0.631	0.261	0.516	0.614	0.115
RCRNet	R + C	0.460	0.335	0.658	0.270	0.451	0.554	0.145

CF: CenterFusion; RF: RCRNet.

Figure 9 illustrates several visualisation results of our proposed approach. We selected three representative frames from the test data, which include scenarios with pedestrian interference, target occlusion, and multi-scale small objects. The target centre, ground truth, and detection results for these frames are displayed from left to right. Our method demonstrates notable proficiency in detecting and identifying distant objects (marked in green), occluded objects (marked in yellow), and obstacles (marked in orange).



Figure 9. Results of our approach under harsh environmental conditions.

For autonomous driving, target detection results are occasionally converted into bird's eye view (BEV) to aid in downstream planning and control tasks. Figure 10 shows several BEV plots of detection results in typical scenarios. These plots demonstrate that RCRFNet delivers exceptional detection results by fusing radar and camera data.

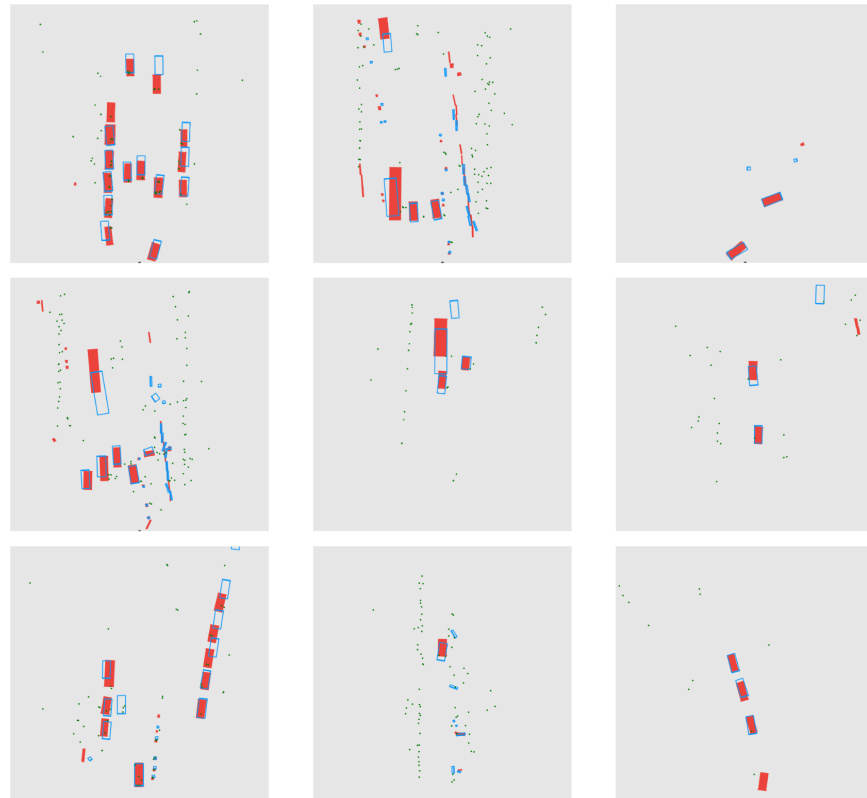


Figure 10. BEV plots of detection results.

3.2. Robustness Performance Analysis

To validate the robustness of the proposed method, we conducted a comparative analysis under various visibility conditions. In this section, experiments were performed with images of four different visibility levels, while keeping other parameter settings unchanged. This setup simulates different weather conditions. Table 4 presents the experimental results for CenterFusion and our RCRFNet. In this table, the maximum loss value, calculated as the difference between the performance index of the highest visibility image and that of the lowest visibility image, is used as the robustness criterion. It can be observed that the RCRFNet exhibits a lower maximum loss value for NDS, mAOE, mAVE, and mAP, indicating higher robustness under varying weather conditions.

Table 4. Comparison of the maximum loss value under varying image visibility levels.

Metrics	CenterFusion	RCRFNet
NDS	0.0207	0.0168
mAOE	0.0379	0.0152
mAVE	0.0998	0.0820
mAP	0.0296	0.0094

Table 5 details the performance of CenterFusion and the RCRFNet method across four visibility levels: 0–40%, 40–60%, 60–80%, and 80–100%. In these fine-grained comparative experiments, the training data volume is significantly reduced due to data screening, posing additional challenges for detection models. Despite this, our method outperforms

CenterFusion, thanks to the implementation of various new processing modules that effectively integrate the complementary strengths of radar point cloud data and camera image data. Figure 11 displays several detection results of the RCRFNet method under harsh environmental conditions. From top to bottom, the images are presented as the original image, ground truth, detection results, and image with the object's depth information.

Table 5. Performance for four different image visibility levels. $\uparrow$ indicates that higher is better and $\downarrow$ indicates that lower is better.

Visibility	0–40%	40–60%	60–80%	80–100%
NDS $\uparrow$ (CF)	0.3639	0.3765	0.3795	0.3846
NDS $\uparrow$ (RF)	0.4414	0.4419	0.4490	0.4582
mAOE $\downarrow$ (CF)	0.5108	0.4824	0.4691	0.4729
mAOE $\downarrow$ (RF)	0.4459	0.4560	0.4475	0.4307
mAVE $\downarrow$ (CF)	1.3009	1.2472	1.2598	1.2011
mAVE $\downarrow$ (RF)	0.6395	0.5933	0.5797	0.5575
mAP $\uparrow$ (CF)	0.2707	0.2904	0.2940	0.3003
mAP $\uparrow$ (RF)	0.3191	0.3130	0.3227	0.3285

CF: CenterFusion; RF: RCRFNet.



Figure 11. Results of our approach under harsh environmental conditions.

3.3. Open-Set Performance Analysis

In this subsection, we aim to verify the effectiveness of the RCOSR module proposed in this paper. We divide the dataset categories into known and unknown categories. During training, only the known categories are used, and the trained model is then applied in open-set recognition experiments. A comparison with the Openmax method under two different visibility levels is shown in Table 6. To ensure a fair comparison, the number of training iterations is set to 30, while other parameters remain unchanged. Table 7 presents detailed performance results for the closed-set categories of Openmax and the RCRFNet method. The experimental results on the nuScenes dataset demonstrate that the radar–vision fusion open-set recognition method proposed in this paper achieves better detection accuracy. These findings confirm the effectiveness of the RCOSR module to a certain extent.

Table 6. Open-set accuracy comparison.

Method	mAP	
Visibility	40–70%	70–100%
Openmax	0.387	0.415
RCRFNet	0.405	0.427

Table 7. Comparison details for closed-set categories.

Method	Car	Truck	Bus	Trailer	Vehicle	Ped.	Moto	Bicycle	Traffic
Openmax	0.455	0.209	0.295	0.089	0.020	0.338	0.269	0.172	0.522
RCRFNet	0.466	0.223	0.314	0.105	0.021	0.346	0.259	0.175	0.520

4. Ablation Study

This section describes the ablation experiments on the NuScenes dataset, using CenterFusion as our baseline to validate the effectiveness of each module: deformable convolution layers, RCSL, and RCOSR.

In the experiments, we incrementally included the modules of deformable convolution layers, RCSL, and RCOSR into our model. The results are summarised in Table 8. The findings indicate that the deformable convolution layers module improves NDS compared to baseline. The RCSL module, which primarily enhances robustness, improves the mAAE and mAOE metrics. The RCOSR module, by computing a contrastive loss between radar depth information and the depth predicted from images, corrects the depth prediction derived from images, significantly enhancing both mAP and NDS. Notably, the NDS index shows an improvement of 8.9%, demonstrating the effectiveness of incorporating radar depth information to enhance detection performance.

Furthermore, the comparisons of the model parameter size and the NDS metric between our approach and SOTA methods are listed in Table 9. It can be seen that the proposed RCRFNet has only a 4.1% model size increase compared to CenterFusion but has achieved a 2.5% improvement in the NDS metric.

Table 8. Ablation experiments for each module. ↑ indicates that higher is better and ↓ indicates that lower is better.

Method	Rad	DC	RL	RR	NDS↑	mAP↑	Error (↓)				
							mATE	mASE	mAOE	mAVE	mAAE
Ours	✓	✓	-	-	0.451	0.326	0.674	0.267	0.445	0.555	0.154
Ours	✓	✓	✓	-	0.456	0.328	0.672	0.266	0.435	0.570	0.142
Ours	✓	✓	✓	✓	0.460	0.335	0.658	0.270	0.451	0.554	0.145

Rad: radar; DC: deformable convolution; RL: RCSL; RR: RCOSR.

Table 9. Comparisons of model size and NDS metric.

Method	Parameters (M)	NDS (%)
FCOS3D	127.6	41.5
DETR3D	613.2	42.2
CenterNet (Hourglass)	730.4	40.0
CenterNet (DLA)	241.6	32.8
CenterFusion (DLA)	248.9	44.9
RCRFNet (DLA)	258.7	46.0

To demonstrate the superior performance of our approach compared to the CenterFusion method under various scene conditions, Figure 12 presents detection results for scenarios including occluded obstacles, small objects in low-light conditions, long-distance observations, and lens-contaminated scenes. It is evident that our approach is more effective in detecting distant objects and handling scenes with occlusions and lens contamination.



Figure 12. Qualitative comparison between our approach and CenterFusion under various scene conditions. For each row, the images from left to right are the raw image, the detection result of CenterFusion, and the result of our approach.

5. Conclusions

To address the limitations of current mainstream methods in robust target detection under low visual visibility and the challenge of effectively handling unknown class targets, this paper proposes a new fusion method that combines mmWave radar point cloud data with visual images. The method introduces an adaptive correlation module for enhanced robustness and incorporates uncertainty calculation for the recognition of unknown category targets. Experimental results on the nuScenes dataset show that the proposed RCRFNet method outperforms SOTA methods, particularly in low-visibility scenes. Although the RCRFNet method is currently designed for data within the field of view, future research will explore its application in bird's eye view, potentially further improving target detection and classification performance.

Author Contributions: Conceptualisation, M.C. and Z.Z.; methodology, M.C. and W.G.; validation, M.C. and Y.L.; writing—original draft preparation, M.C.; writing—review and editing, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under grant 62271311 and grant 62071333.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in [nuScenes] at <https://doi.org/10.1109/CVPR42600.2020.01164>, reference number [40].

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

- Chen, L.; Teng, S.; Li, B.; Na, X.; Li, Y.; Li, Z.; Wang, J.; Cao, D.; Zheng, N.; Wang, F.Y. Milestones in Autonomous Driving and Intelligent Vehicles—Part II: Perception and Planning. *IEEE Trans. Syst. Man. Cybern. Syst.* **2023**, *53*, 6401–6415. [\[CrossRef\]](#)
- Wei, Z.; Zhang, F.; Chang, S.; Liu, Y.; Wu, H.; Feng, Z. MmWave Radar and Vision Fusion for Object Detection in Autonomous Driving: A Review. *Sensors* **2022**, *22*, 2542. [\[CrossRef\]](#) [\[PubMed\]](#)
- Huo, B.; Li, C.; Zhang, J.; Xue, Y.; Lin, Z. SAFF-SSD: Self-Attention Combined Feature Fusion-Based SSD for Small Object Detection in Remote Sensing. *Remote Sens.* **2023**, *15*, 3027. [\[CrossRef\]](#)
- Carion, N.; Massa, F. End-to-End Object Detection with Transformers. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; pp. 213–229. [\[CrossRef\]](#)
- Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [\[CrossRef\]](#)
- Li, H.; Sima, C.; Dai, J.; Wang, W.; Lu, L.; Wang, H.; Zeng, J.; Li, Z.; Yang, J.; Deng, H.; et al. Delving Into the Devils of Bird’s-Eye-View Perception: A Review, Evaluation and Recipe. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 2151–2170. [\[CrossRef\]](#) [\[PubMed\]](#)
- Zhao, S.; Zhang, Z.; Zhang, T.; Guo, W.; Luo, Y. Transferable SAR Image Classification Crossing Different Satellites under Open Set Condition. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [\[CrossRef\]](#)
- Zhao, S.; Zhang, Z.; Guo, W.; Luo, Y. An Automatic Ship Detection Method Adapting to Different Satellites SAR Images with Feature Alignment and Compensation Loss. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–17. [\[CrossRef\]](#)
- Zhou, H.; Dong, C.; Wu, R.; Xu, X.; Guo, Z. Feature Fusion Based on Bayesian Decision Theory for Radar Deception Jamming Recognition. *IEEE Access* **2021**, *9*, 16296–16304. [\[CrossRef\]](#)
- Bendale, A.; Boulton, T.E. Towards Open Set Deep Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27 June–1 July 2016; pp. 1563–1572. [\[CrossRef\]](#)
- Bao, W.; Yu, Q.; Kong, Y. Evidential Deep Learning for Open Set Action Recognition. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 13329–13338. [\[CrossRef\]](#)
- Svenningsson, P.; Fioranelli, F.; Yarovsky, A. Radar-PointGNN: Graph Based Object Recognition for Unstructured Radar Point-cloud Data. In Proceedings of the 2021 IEEE Radar Conference (RadarConf21), Atlanta, Georgia, USA, 8–14 May 2021; pp. 1–6. [\[CrossRef\]](#)
- Meyer, M.; Kusch, G.; Tomforde, S. Graph Convolutional Networks for 3D Object Detection on Radar Data. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 3053–3062. [\[CrossRef\]](#)
- Zhang, Z.; Zhang, L.; Wu, J.; Guo, W. Optical and Synthetic Aperture Radar Image Fusion for Ship Detection and Recognition: Current state, challenges, and future prospects. *IEEE Geosci. Remote Sens. Mag.* **2024**, 2–38. [\[CrossRef\]](#)
- Bi, X.; Tan, B.; Xu, Z.; Huang, L. *A New Method of Target Detection Based on Autonomous Radar and Camera Data Fusion*; SAE Technical Paper; SAE International: Warrendale, PA, USA, 1977; 2017-01-1977. [\[CrossRef\]](#)
- Chang, S.; Zhang, Y.; Zhang, F.; Zhao, X.; Huang, S.; Feng, Z.; Wei, Z. Spatial Attention Fusion for Obstacle Detection Using MmWave Radar and Vision Sensor. *Sensors* **2020**, *20*, 956. [\[CrossRef\]](#) [\[PubMed\]](#)
- Drewe, F.; Feng, D.; Faion, F.; Rosenbaum, L.; Ulrich, M.; Gläser, C. DeepFusion: A Robust and Modular 3D Object Detector for Lidars, Cameras and Radars. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022; pp. 560–567. [\[CrossRef\]](#)
- Nabati, R.; Qi, H. CenterFusion: Center-based Radar and Camera Fusion for 3D Object Detection. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 5–9 January 2021; pp. 1526–1535. [\[CrossRef\]](#)
- Kim, J.; Kim, Y.; Kum, D. Low-level sensor fusion network for 3d vehicle detection using radar range-azimuth heatmap and monocular image. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020; pp. 388–402. [\[CrossRef\]](#)
- Chavez-Garcia, R.O.; Burt, J.; Vu, T.D.; Aycard, O. Frontal object perception using radar and mono-vision. In Proceedings of the 2012 IEEE Intelligent Vehicles Symposium, Alcalá de Henares, Spain, 3–7 June 2012; pp. 159–164. [\[CrossRef\]](#)
- Česić, J.; Marković, I.; Cvišić, I.; Petrović, I. Radar and stereo vision fusion for multitarget tracking on the special Euclidean group. *Robot. Auton. Syst.* **2016**, *83*, 338–348. [\[CrossRef\]](#)

22. Wang, X.; Xu, L.; Sun, H.; Xin, J.; Zheng, N. Bionic vision inspired on-road obstacle detection and tracking using radar and visual information. In Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), Qingdao, China, 8–11 October 2014; pp. 39–44. [\[CrossRef\]](#)
23. Guo, X.; Du, J.; Gao, J.Y.; Wang, W. Pedestrian Detection Based on Fusion of Millimeter Wave Radar and Vision. In Proceedings of the 2018 International Conference on Artificial Intelligence and Pattern Recognition, Beijing, China, 18–20 August 2018; pp. 38–42. [\[CrossRef\]](#)
24. Kadow, U.; Schneider, G.; Vukotich, A. Radar-Vision Based Vehicle Recognition with Evolutionary Optimized and Boosted Features. In Proceedings of the 2007 IEEE Intelligent Vehicles Symposium, Istanbul, Turkey, 13–15 June 2007; pp. 749–754. [\[CrossRef\]](#)
25. Han, Y.; Zhang, H.; Li, H.; Jin, Y.; Lang, C.; Li, Y. Collaborative Perception in Autonomous Driving: Methods, Datasets, and Challenges. *IEEE Intell. Transp. Syst. Mag.* **2023**, *15*, 131–151. [\[CrossRef\]](#)
26. Nobis, F.; Geisslinger, M.; Weber, M.; Betz, J.; Lienkamp, M. A Deep Learning-based Radar and Camera Sensor Fusion Architecture for Object Detection. In Proceedings of the 2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF), Bonn, Germany, 15–17 October 2019; pp. 1–7. [\[CrossRef\]](#)
27. Maiti, A.; Elberink, S.O.; Vosselman, G. TransFusion: Multi-modal Fusion Network for Semantic Segmentation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 17–24 June 2023; pp. 6537–6547. [\[CrossRef\]](#)
28. Aziz, K.; De Greef, E.; Rykunov, M.; Bourdoux, A.; Sahli, H. Radar-camera Fusion for Road Target Classification. In Proceedings of the 2020 IEEE Radar Conference (RadarConf20), Florence, Italy, 21–25 September 2020; pp. 1–6. [\[CrossRef\]](#)
29. Kim, J.; Emeršič, Ž.; Han, D.S. Vehicle Path Prediction based on Radar and Vision Sensor Fusion for Safe Lane Changing. In Proceedings of the 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC), Okinawa, Japan, 11–13 February 2019; pp. 267–271. [\[CrossRef\]](#)
30. Bertozzi, M.; Bombini, L.; Cerri, P.; Medici, P.; Antonello, P.C.; Miglietta, M. Obstacle detection and classification fusing radar and vision. In Proceedings of the 2008 IEEE Intelligent Vehicles Symposium, Eindhoven, The Netherlands, 4–6 June 2008; pp. 608–613. [\[CrossRef\]](#)
31. Obrvan, M.; Cesic, J.; Petrović, I. Appearance Based Vehicle Detection by Radar-Stereo Vision Integration. In Proceedings of the ROBOT, Lisbon, Portugal, 19–21 November 2015; pp. 437–449. [\[CrossRef\]](#)
32. Yu, F.; Wang, D.; Shelhamer, E.; Darrell, T. Deep Layer Aggregation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2403–2412. [\[CrossRef\]](#)
33. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [\[CrossRef\]](#)
34. Yoshihashi, R.; Shao, W.; Kawakami, R.; You, S.; Iida, M.; Naemura, T. Classification-Reconstruction Learning for Open-Set Recognition. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4011–4020. [\[CrossRef\]](#)
35. Zhang, R.; Cao, S. Extending Reliability of mmWave Radar Tracking and Detection via Fusion with Camera. *IEEE Access* **2019**, *7*, 137065–137079. [\[CrossRef\]](#)
36. Zhao, Z.; Dong, M. Channel-Spatial Dynamic Convolution : An Exquisite Omni-dimensional Dynamic Convolution. In Proceedings of the 2023 8th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 21–23 April 2023; pp. 1707–1711. [\[CrossRef\]](#)
37. Chen, X.; Kundu, K.; Zhang, Z.; Ma, H.; Fidler, S.; Urtasun, R. Monocular 3D Object Detection for Autonomous Driving. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2147–2156. [\[CrossRef\]](#)
38. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6568–6577. [\[CrossRef\]](#)
39. Wang, T.; Zhu, X.; Pang, J.; Lin, D. FCOS3D: Fully Convolutional One-Stage Monocular 3D Object Detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 913–922. [\[CrossRef\]](#)
40. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuScenes: A Multimodal Dataset for Autonomous Driving. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11618–11628. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.