引用格式:任坤, 李盼, 韩红桂. 挑战性环境下基于双尺度CBAM的毫米波雷达与视觉特征融合目标检测[J]. 北京工业大学学报, 2025, 51(3): 284-294.
REN K, LI P, HAN H G. Object detection in challenging environments via dual-scale CBAM feature fusion of mmWave radar and vision[J]. Journal of Beijing University of Technology, 2025, 51(3): 284-294. (in Chinese)

挑战性环境下基于双尺度CBAM的毫米波雷达与视觉特征融合目标检测

任 坤[1,2,3], 李 盼[1,2,3], 韩红桂[1,2,3]
(1. 北京工业大学信息学部, 北京 100124; 2. 数字社区教育部工程研究中心, 北京 100124;
3. 城市轨道交通北京实验室, 北京 100124)

摘 要: 针对恶劣天气和低光照对基于深度学习的视觉目标检测算法带来的挑战,提出一种基于双尺度卷积注意力模块(convolutional block attention module,CBAM)的双模态目标检测算法,旨在通过视觉与毫米波雷达数据的特征融合,提高目标检测算法在挑战性环境下的鲁棒性和准确性。该算法采用双分支的一阶段检测结构,图像分支采用预训练的CSPDarkNet53骨干网络提取图像特征,雷达分支采用基于体素的雷达特征生成网络提取雷达特征。然后,分别在颈部网络前后利用提出的基于双尺度CBAM的特征融合模块进行雷达-视觉特征融合。最后,使用解耦检测头实现目标的分类和定位。在nuScenes数据集上,利用对比实验和消融实验验证了该特征融合检测算法在挑战性环境下的有效性和优越性。

关键词: 深度学习; 目标检测; 毫米波雷达; 特征融合; 多模态; 注意力机制
中图分类号: TP 311 **文献标志码**: A **文章编号**: 0254-0037(2025)03-0284-11
doi: 10.11936/bjutxb2023070003

Object Detection in Challenging Environments via Dual-scale CBAM Feature Fusion of mmWave Radar and Vision

REN Kun[1,2,3], LI Pan[1,2,3], HAN Honggui[1,2,3]
(1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;
2. Engineering Research Center of Digital Community, Ministry of Education, Beijing 100124, China;
3. Beijing Laboratory for Urban Mass Transit, Beijing 100124, China)

Abstract: A dual-modality object detection algorithm, based on the dual-scale convolutional block attention module (CBAM), is addressed to tackle challenges posed by adverse weather conditions and low lighting for visual object detection algorithms based on deep learning. The algorithm aims to improve the robustness and accuracy of object detection in challenging environments by fusing features from vision and millimeter wave (mmWave) radar. It utilized a dual-branch one-stage architecture, with the image branch using a pre-trained CSPDarkNet53 backbone network to extract image features and the radar branch employing a voxel-based radar feature generation network to extract radar features. The proposed dual-scale CBAM feature fusion module integrated radar and visual features before and after the neck

收稿日期: 2023-07-05; 修回日期: 2024-06-05
基金项目: 国家重点研发计划资助项目(2023YFC3904605); 国家自然科学基金资助项目(62125301, 62203022)
作者简介: 任 坤(1973—), 女, 副教授, 主要从事计算机视觉方面的研究, E-mail: renkun@bjut.edu.cn

network. Finally, a decoupled detection head was deployed to classify and locate objects. The effectiveness and superiority of the proposed fusion detection algorithm were validated by comparative and ablation experiments conducted on the nuScenes dataset in challenging environments.

Key words: deep learning; object detection; millimeter wave (mmWave) radar; feature fusion; multi-modality; attention mechanism

鲁棒的高精度环境感知系统在自动驾驶(automatic driving,AD)中至关重要,其性能直接决定了 AD 车辆和所有交通参与者的安全。然而,基于视觉的环境感知系统受光照和天气影响非常大。在弱光或恶劣天气环境(如大雨、雪和雾)中,即使是最先进的基于深度学习的视觉目标检测方法也面临失效的挑战[1]。在 AD 领域,多传感器融合的环境感知系统是对抗自然界挑战性环境的必然方案[1-3]。多传感器融合不仅可以有效地结合多种模态的优点以挖掘更多场景语义信息,而且可以在部分模态失效时尽可能减小对模型性能的影响[4]。

在 AD 系统中,有 3 种流行的多传感器融合方案:激光雷达-视觉(LiDAR-vision,LV)、雷达-视觉(radar-vision,RV)和激光雷达-雷达-视觉(LiDAR-radar-vision,LRV)[2]。传感器的工作电磁频段决定了其能够探测感知环境信息的有效范围。对于工作在可见光频段的视觉相机,尽管其可以捕捉丰富的视觉结构和纹理信息,但低光照和雨、雾等恶劣天气会显著降低其成像性能。同样,激光雷达在暴雪、大雨和大雾等不利天气下探测范围缩小,视野和视距受阻[1-5]。与激光雷达相比,毫米波雷达具有更强的穿透力和更长的探测距离,能够适应各种不同的恶劣天气以及复杂环境,可以全天时全天候工作[6-7]。因此,RV 融合是挑战性环境下的优选方案[8-9]。

RV 融合通常分为数据融合、特征融合和决策融合[9-10]。数据融合是指利用雷达数据预测感兴趣区域(region of interest,RoI),并与图像融合形成图像 RoI 用于检测。数据融合在图像检测模型的输入端完成。尽管雷达数据被充分用于融合预测,但这种级联处理使得数据融合系统的精度受到雷达预测精度的限制。决策融合方法是对不同传感器的检测结果进行融合,通过联合决策模型生成最终检测结果。然而,在决策融合中,多传感器的联合密度函数难以建模[9],而且整个决策融合系统非常复杂,计算成本高昂。特征融合系统是将雷达数据特征与图像特征融合进行预测[10]。与数据融合和决策融合相比,特征融合不仅能够突破单一传感器的信息限制,实现更好的检测性能[9],而且其复杂度低于决策融合,更为经济。因此,近年来特征融合已成为相关研究的热点方向。

最近,基于 RV 特征融合的 2D 目标检测利用深度学习强大的特征表示能力,在 AD 领域取得了一些令人瞩目的成果[11]。Chadwick 等[12]首次将毫米波雷达特征与图像特征融合,实现了基于深度学习的 RV 融合目标检测。在这项工作中,雷达数据被投影到图像平面上,在提取特征后与深度网络提取的图像特征融合成融合特征层,并通过输入深度视觉检测模型进行检测,有效提高了远距离小目标的检测精度。RVNet[13]将雷达数据映射至相机坐标系下,使用卷积神经网络(convolution neural network,CNN)分别对雷达数据与图像进行特征提取,将雷达和图像特征在检测网络中进行融合。CRF-Net[14]将毫米波雷达数据映射到图像平面,分别在骨干网络和特征金字塔网络(feature pyramid network,FPN)[15]的不同阶段进行特征融合,并提出了 BlackIn 训练策略,使网络更依赖雷达数据进行训练,在挑战性环境下的检测性能得到有效提升。Chang 等[16]提出空间注意力融合(spatial attention fusion,SAF)方法,利用多尺度空间注意力将雷达特征编码为权重矩阵,并应用于图像特征,实现加权融合。该方法提高了模型对关键信息的感知能力,取得了优于特征拼接和特征相加等融合方式的性能。

然而,目前基于 RV 特征融合的深度目标检测算法还处于研究阶段,仍面临很多挑战。首先,雷达数据是一种稀疏的 2D 点云数据,其数据分布与图像不同[3],通用卷积网络并不适用于提取雷达数据特征[5]。如何有效提取适用于检测任务的雷达数据特征仍然是一个开放问题。其次,雷达特征和视觉特征的融合方式也有待深入研究。目前,RV 特征融合操作主要为元素相加[13]、元素相乘[17]或特征拼接[18]。这些操作没有考虑在不同环境中雷达和图像提供的有效信息及其重要性是不同的,而基于空间注意力的融合方式没有关注通道间的相关性,存在性能瓶颈。

针对上述问题,本文提出一种基于双尺度卷积注意力模块(convolutional block attention module,CBAM)的 RV 特征融合检测模型。该模型采用双分支的一阶段检测框架,由雷达数据特征生成、图像特征提取、雷达数据与图像的特征融合和检测头 4 个部分构成。在雷达特征生成部分,借鉴了激光点云检测模型 VoxelNet[19] 来抑制数据噪声,减少冗余信息干扰,生成更加适合目标检测任务的有效雷达特征。在特征融合部分,本文提出基于双尺度 CBAM 的 RV 特征融合模块(RV feature fusion module,RVFFM),分别在空间和通道上改进模型对目标尺度变化的适应能力,并增强重要特征的表达能力,从而提升目标检测的精度和鲁棒性。最后,在检测头部分,考虑到分类和定位任务存在空间错位的问题[20],引入了解耦检测头以改进检测网络,进一步提高模型检测精度。为了验证模型的有效性,本文在 nuScenes 公共数据集[21]上进行了模型训练和评估。与最先进的(state-of-the-art, SOTA)视觉检测模型 YOLOv7-x[22]、YOLOX-x[23] 以及 YOLOv8-x[24] 相比,本文方法在夜间环境和雨、雾天气条件下的检测精度均有显著提高。同时,本文方法的检测精度也优于代表性的 RV 特征融合检测模型。此外,消融实验的结果也验证了本文方法的有效性。

1 基于双尺度 CBAM 的 RV 特征融合检测模型

RV 特征融合检测模型采用双分支结构,整体结构如图 1 所示。图中 A 为锚框数。在雷达分支中,雷达特征生成网络对经预处理后的雷达数据进行特征提取,生成用于特征融合的雷达特征。视觉分支采用一阶段目标检测框架。首先,采用 YOLOv5[25] 骨干网络 CSPDarkNet53 提取图像特征;然后,分别在颈部网络前后,利用基于双尺度 CBAM 的 RVFFM 在 3 个尺度上进行雷达和图像特征融合;最后,在检测头部分,设计简化解耦检测头,缓解分类和定位回归之间的相互影响,进一步提升检测模型的定位精度。

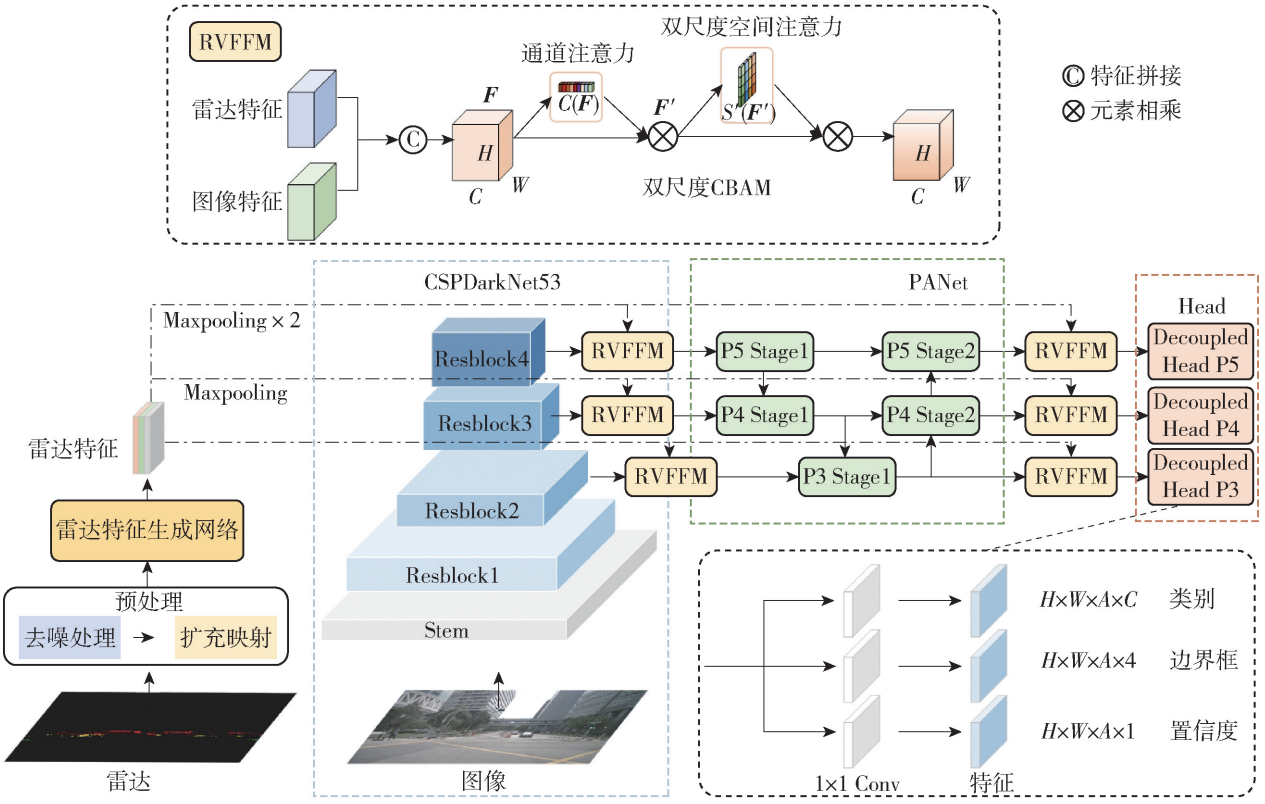


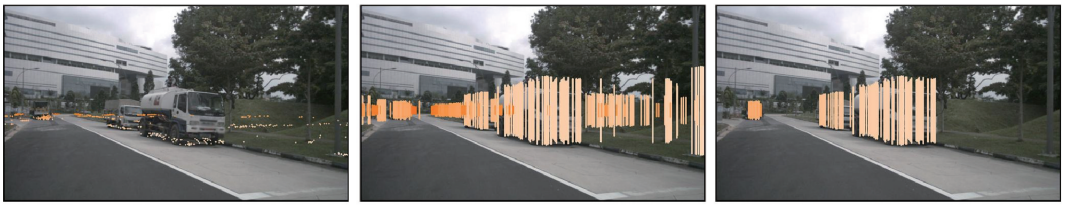
图 1 基于双尺度 CBAM 的 RV 特征融合检测模型结构

Fig. 1 Structure of RV feature fusion detection model based on dual-scale CBAM

1.1 基于体素的雷达特征生成网络

由于毫米波雷达数据与视觉图像是完全异构的,在特征提取前需要对雷达数据进行坐标变换和采样周期同步等预处理。本文采用 CRF-Net 的预处理方法。首先,通过合并 13 个周期的雷达数据增加数据的信息密度,同时实现采样同步。合并后雷达

数据经坐标变换映射到图像坐标系，如图 2(a)所示。其次，为弥补雷达数据无高度信息的问题，设雷达探测目标高度为 3m，经坐标变换后不同径向距离的目标点在图像的映射如图 2(b)所示。最后，通过 3D 边界框与雷达点的关联性滤除未被雷达检测到的目标以及 3D 边界框以外的雷达数据，以此获取较理想的雷达数据。滤波后的雷达数据如图 2(c)所示。由于毫米波雷达数据是点云数据，直接使用基于图像预训练的卷积骨干网络提取雷达特征是不合适的。本文借鉴处理激光雷达数据的 VoxelNet[19] 特征提取方法构建了生成毫米波雷达特征的雷达特征生成网络(radar feature generation network，RFGNet)。RFGNet 通过空间编码和 3D 卷积层将预处理后的雷达数据映射到更高维的特征空间，通过学习生成更适合目标检测任务的雷达特征。RFGNet 的网络结构如图 3 所示。



(a) 周期合并　　(b) 高度扩展　　(c) 噪声过滤

图 2　雷达数据在图像上的映射

Fig. 2　Radar data mapping on images

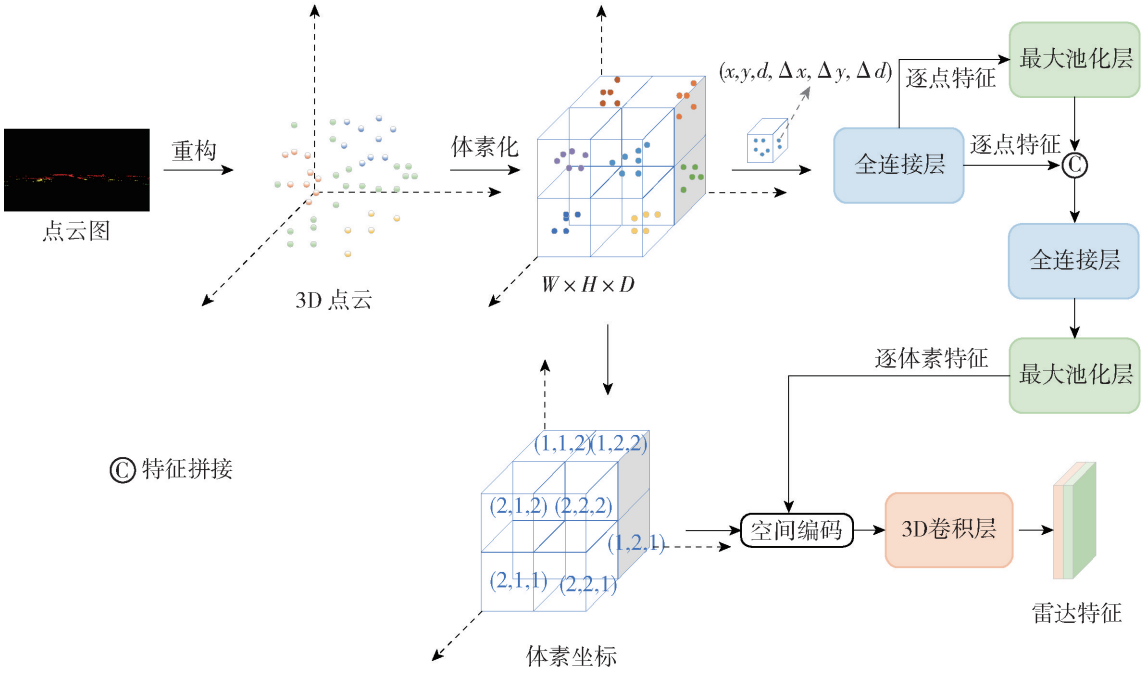


图 3　雷达点云体素处理和 RFGNet 结构

Fig. 3　Voxel processing of radar point cloud and RFGNet structure

在空间编码阶段，首先将三维点云划分为设定大小的体素网格 $\boldsymbol{V} \in \mathbb{R}^{W \times H \times D}$，$\mathbb{R}$ 为实数集，W、H 和 D 为体素网格的 3 个维度，如图 3 所示。然后，通过全连接层和最大池化操作生成体素特征。

如图 3 所示，经预处理的每一个三维雷达点可定义为 $\boldsymbol{r}_i = (x_i, y_i, d_i, \Delta x_i, \Delta y_i, \Delta d_i)^{\mathrm{T}} \in \mathbb{R}^6$，其中：$(x_i, y_i, d_i)$ 为第 i 个雷达点的空间位置坐标；$(\Delta x_i, \Delta y_i, \Delta d_i)$ 为第 i 个雷达点坐标相对于其所属体素质心坐标的偏移量，其定义为

$$(\Delta x_i, \Delta y_i, \Delta d_i) = (x_i, y_i, d_i) - \frac{1}{J}\sum_{j=1}^{J}(x_j, y_j, d_j), \quad j = 1, 2, \cdots, J \tag{1}$$

式中 J 为该雷达点所在体素网格中雷达点的数量。雷达点 $\boldsymbol{r}_i$ 通过全连接层映射得到点特征 $\boldsymbol{r}_i'$，再对相同体素内所有点特征做最大池化以得到该体素局部聚合特征。然后，通过将每一个点特征 $\boldsymbol{r}_i'$ 与体素局部聚合特征进行拼接得到拼接特征，再经一个全连

接映射和最大池化操作,生成该体素的聚合特征$\boldsymbol{v} \in \mathbb{R}^C$,其中 C 表示特征通道数。

在此基础上,第 g 个体素特征 $\boldsymbol{v}_g$ 结合其对应的体素空间位置(W_g, H_g, D_g)进行特征编码。所有体素编码特征构成了基于体素的雷达特征张量 $\boldsymbol{T} \in \mathbb{R}^{W \times H \times D \times C}$。最后,将 $\boldsymbol{T}$ 输入 3D 卷积层,生成最终用于融合的雷达特征 $\boldsymbol{F}_{\mathrm{R}} \in \mathbb{R}^{W' \times H' \times C'}$。

1.2 基于双尺度 CBAM 的 RVFFM

RVFFM 的具体结构设计如图 4 所示。雷达和图像特征拼接形成拼接特征 $\boldsymbol{F}$,然后通过双尺度 CBAM 在通道上增强特征融合的同时,在空间上利用双尺度感受野进一步增强不同尺度特征,以优化融合特征。其中通道注意力(channel attention,CA)模块使用全局平均池化(global average pooling,GAP)和全局最大池化(global max pooling,GMP)聚合 $\boldsymbol{F}$ 的空间信息,分别得到 GAP 特征和 GMP 特征,两者分别用于捕捉特征的整体信息和最显著的特征。

随后,分别经过全连接层和非线性激活函数的处理,得到 2 个不同的 CA 特征向量。最后,将 2 个 CA 特征向量相加,并通过 Sigmoid 激活函数进行归一化,得到最终的 CA 矩阵图,计算过程表示为

$$\mathrm{CA}(\boldsymbol{F}) = \mathrm{Sigmoid}(\mathrm{MLP}(\mathrm{GAP}(\boldsymbol{F})) \oplus \mathrm{MLP}(\mathrm{GMP}(\boldsymbol{F}))) \tag{2}$$

式中:GAP(·)表示全局平均池化操作;GMP(·)表示全局最大池化操作;⊕表示相加操作;Sigmoid(·)表示 Sigmoid 激活函数。CA 模块通过对特征 $\boldsymbol{F}$ 的不同通道信息分配权重,使得最具有区分度和代表性的特征能够更好地被保留和强化,而对于噪声或不相关的特征则进行削弱。

为提升模型对于不同尺度目标的感知能力,本文对 CBAM[26] 中的空间注意力进行了改进,提出了双尺度空间注意力(dual-scale spatial attention,DSSA)模块,其结构如图 4 所示。首先,使用 GAP 和 GMP 聚合特征图的通道特征,分别得到全局平均池化特征和全局最大池化特征,并将两者进行拼接。然后,在标准卷积层中分别进行 3×3 和 7×7 这2 个尺度的卷积操作,得到 2 个二维特征张量,它们分别对应不同尺度下的空间信息,以捕获不同大小目标的特征。最后,将 2 个具有不同感受野大小的特征张量进行拼接后,在通道维度做全局平均池化操作,再经过 Sigmoid 激活函数进行归一化,得到最终的 DSSA 矩阵图。DSSA 的计算过程可表示为

$$\mathrm{DSSA}(\boldsymbol{F}') = \mathrm{Sigmoid}(\mathrm{GAP}([\mathrm{Conv3}(\boldsymbol{F}_{\mathrm{m}}), \mathrm{Conv7}(\boldsymbol{F}_{\mathrm{m}})])) \tag{3}$$

$$\boldsymbol{F}_{\mathrm{m}} = [\mathrm{GAP}(\boldsymbol{F}'), \mathrm{GMP}(\boldsymbol{F}')]$$

$$\boldsymbol{F}' = \mathrm{CA}(\boldsymbol{F}) \otimes \boldsymbol{F}$$

式中:$\boldsymbol{F}'$为 DSSA 模块输入特征;$\boldsymbol{F}_{\mathrm{m}}$为中间拼接特征;[·,·]表示拼接操作;Conv∗(·)表示卷积核大小为∗×∗的卷积操作;⊗表示相乘操作。双尺度的改进使得生成的空间注意力矩阵具有双尺度的感受野,在一定程度上能够提升小目标的检测精度。

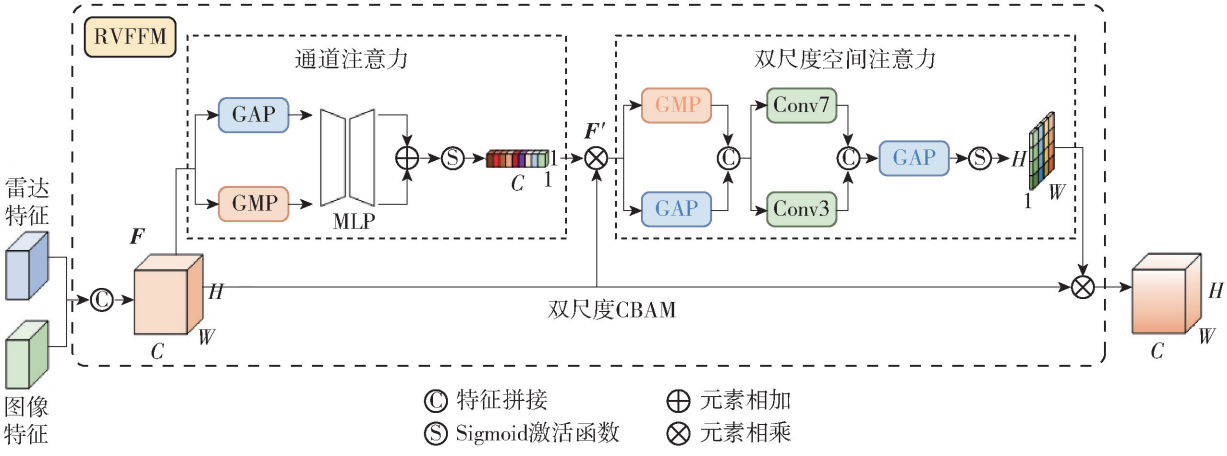


图 4 双尺度 CBAM 和 RVFFM 结构

Fig. 4 Dual-scale CBAM and RVFFM structure

RVFFM 能够自适应预测潜在的关键特征,从通道上和空间上对雷达特征和图像特征进行加权学习,增强对检测任务更具有意义的特征,进而提升特征融合的效果和模型的检测性能。本文分别使用

3×3、5×5 和 7×7 任意 2 个尺度以及 3 个尺度组合的卷积操作进行测试实验。

1.3　解耦检测头

在模型预测阶段，分类任务与定位任务所关注的关键目标或感兴趣的内容信息不同。其中分类任务更多地关注目标的局部纹理信息，而定位任务更多地关注目标的关键位置信息[17]。因此，为减小不同任务之间空间错位的影响，本文借鉴 YOLOX[23] 的解耦检测头，设计了 3 个 1×1 卷积分别用于预测目标类别、边界框和置信度，具体结构如图 1 所示。

模型总损失函数 L_{total} 定义为

$$L_{\text{total}} = \sum_{l=3}^{5} \left(\lambda_1 L_{\text{box},P_l} + \lambda_2 L_{\text{obj},P_l} + \lambda_3 L_{\text{cls},P_l} \right) \quad (4)$$

式中：P_l 为第 l 个检测层；L_{box}、L_{obj} 和 L_{cls} 分别为边界框损失、置信度损失和分类损失；λ_1、λ_2 和 λ_3 为平衡因子，分别取值为 0.05、0.60 和 0.05。

2　实验分析

为验证本文方法的有效性，分别开展了与 SOTA 纯视觉检测方法和代表性 RV 特征融合方法的对比实验及消融实验。最后，通过本文方法与 YOLOv8-x 的检测特征和结果的可视化分析，进一步说明本文方法在挑战性环境下具有优越性。

2.1　数据集和实验设置

2.1.1　数据集和评价指标

nuScenes 数据集[21] 是由 Motional 团队开发的 AD 公共大规模数据集，它记录了波士顿和新加坡共 1 000 个不同场景和不同条件（如雨、雾天气和夜间）下的图像和雷达数据，其中，700 个场景为训练集，150 个场景为验证集，包含了汽车、人、卡车、摩托车等 23 个对象类别的边界框标注，150 个场景为测试集且无标注。

本文对 nuScenes 包含标注的 850 个场景数据集进行划分，按照 6∶2∶2 的比例分别划分为训练集、验证集和测试集。经过划分后的训练集包含 20 480 对图像和雷达数据，验证集包含 6 839 对图像和雷达数据，测试集包含 6 830 对图像和雷达数据。划分后的各数据集均包含白天、雨天和夜间场景数据。这里，将整体测试集作为通用测试集 Test。为进一步区分雨天和夜间环境下的检测性能，将通用测试集 Test 中所有雨天场景数据（共1 215 对图像和雷达数据）作为雨天测试集 Rain，所有夜间场景数据（共 804 对图像和雷达数据）作为夜间测试集 Night。

在本文中，通过使用 CRF-Net[14] 中的方法将数据集中的 3D 边界框投影至图像平面来获得其中汽车、公共汽车、人、自行车、摩托车、卡车、拖车这 7 个类别的 2D 边界框标注，并将其作为待检测的目标类别。

这里，使用交并比（intersection of union，IoU）为 0.5 条件下的平均精度（average precision，AP）和平均精度均值（mean average precision，mAP）[27] 来评价模型在各类别目标上的以及总体的检测性能。

2.1.2　实验平台及参数设置

本算法基于 Pytorch 框架实现，在 NVIDIA GeForce RTX 2080Ti GPU 上进行训练和测试，将输入图像和雷达点云的分辨率设置为 384×640。在实验中，所有网络通过加载预训练权重对模型进行参数初始化，同时采用冻结训练的方式训练 100 Epoch。在前 50 个 Epoch 中冻结模型骨干网络的训练参数。在训练过程中，优化器采用随机梯度下降（stochastic gradient descent，SGD）优化器，初始学习率设置为 0.01，并使用 Cosine Annealing[28] 学习率衰减策略。

2.2　与 SOTA 视觉检测方法对比

为了验证本文方法在不同场景下的优势，选取基线模型 YOLOv5-x 和 SOTA 视觉目标检测模型 YOLOv7-x、YOLOX-x 和 YOLOv8-x 进行对比实验。所有实验模型在 nuScenes 公共数据集训练后，分别在 Test、Rain 和 Night 测试集上测试，实验结果见表 1。

在 Test 上，本文方法的 mAP 不仅比 YOLOv5-x 提高了 6.99 个百分点，而且比 YOLOv8-x 提高了 6.24 个百分点。在 Test 中以正常光照和能见度图像为主，挑战性环境的测试图像较少。YOLOv8-x 的检测精度比 YOLOv5-x 仅提高了 0.75 个百分点。本文方法的检测性能在通用环境中大大优于 SOTA 视觉模型。

在 Rain 上，本文方法的 mAP 达 51.21%，比 YOLOv8-x 提高了 8.09 个百分点，比 YOLOX-x 提高了 6.58 个百分点。对比 Rain 和 Test 的实验结果，可以观察到在雨天环境下，视觉模型的检测性能大幅度降低，而本文方法能够有效补偿雨天环境对视觉检测模型性能的影响。

在 Night 上，本文方法的 mAP 达 52.27%，不仅比 YOLOv5-x 提高了 13.09 个百分点，而且比 YOLOv7-x 提高了 5.76 个百分点。基线模型 YOLOv5-x 在通用环境下的 mAP 为 50.64%，在

Rain 上下降了 5.83 个百分点,在夜间环境下下降了 11.46 个百分点。本文方法在夜间基线模型性能大幅降低的条件下,实现了优于在 Test 上的 SOTA 视觉模型的检测性能。

对比雨天和夜间 2 种挑战性环境,SOTA 视觉模型整体在夜间检测精度更低,低光照对视觉性能影响更大,而本文方法的夜间检测性能优于雨天,mAP 提高了 1.06 个百分点。这是由于毫米波雷达探测受光照和天气等恶劣环境影响较小,能够提供有效信息,而且相对于低光照,降雨造成的多重后向散射效应会影响毫米波雷达的探测性能[29]。

综上分析,本文方法在挑战性环境下优于 SOTA 视觉模型的检测性能。

2.3 与现有 RV 特征融合方法对比

为验证本文方法的优越性,与代表性 RV 特征融合方法,如 RVNet[13]、CRF-Net[14]等,在 nuScenes 数据集上的 mAP 进行对比分析,实验结果见表 2。

表 1 RV 特征融合模型与视觉模型的对比实验结果

Table 1 Comparative experimental results of RV feature fusion model and visual model %

测试集	模型	自行车	公共汽车	汽车	人	摩托车	拖车	卡车	mAP
Test	YOLOv5-x	28.93	66.24	74.32	43.80	40.49	49.61	51.09	50.64
	YOLOv7-x	31.2	65.87	74.00	42.57	39.72	50.97	52.63	51.01
	YOLOX-x	32.76	65.63	75.07	42.04	38.52	46.38	**57.74**	51.16
	YOLOv8-x	28.75	**68.31**	73.57	43.97	40.94	49.80	54.38	51.39
	本文	**35.14**	61.89	**82.05**	**63.87**	**52.84**	**52.15**	55.50	**57.63**
Rain	YOLOv5-x	16.41	**83.86**	74.79	29.73	4.19	55.73	48.96	44.81
	YOLOv7-x	14.17	74.95	74.04	33.50	2.84	**57.93**	50.67	44.01
	YOLOX-x	13.53	80.20	74.50	32.96	3.76	51.36	**56.07**	44.63
	YOLOv8-x	15.02	73.36	73.18	29.13	6.35	54.37	50.40	43.12
	本文	**20.69**	74.76	**79.75**	**63.53**	**14.17**	56.17	49.42	**51.21**
Night	YOLOv5-x	0.42	54.54	78.83	39.48	39.97		21.85	39.18
	YOLOv7-x	**5.93**	**62.56**	79.04	40.52	**44.30**		46.71	46.51
	YOLOX-x	1.27	60.23	79.82	35.40	34.36		40.94	42.00
	YOLOv8-x	1.19	51.04	78.92	40.52	33.10		40.62	40.90
	本文	0.00	53.20	**89.01**	**52.44**	42.97		**75.98**	**52.27**

注:加粗数字表示在对应场景和类别下最高的 mAP 值。

表 2 不同 RV 特征融合方法的性能对比

Table 2 Performance comparison of different RV feature fusion methods %

模型	mAP
RVNet[13]	56.00
CRF-Net[14]	55.99
Lukas & Philipp[18]	36.78
Li & Xie[17]	48.40
本文	57.63

注:表 2 中各方法实验结果引自相应文献。

从表 2 中的数据可以看到,本文方法显著优于对比的代表性 RV 特征融合方法。与 CRF-Net[14]不同的是,在计算 mAP 时,本文没有对各类别 AP 进行样本比例加权。尽管如此,本文方法的 mAP 仍然比其高 2.02 个百分点,这一结果更加充分说明本文方法具有优越性。

2.4 消融实验

本文通过在通用、雨天、夜间环境下的 5 个消融实验验证 RFGNet、双尺度 CBAM 和解耦检测头的有效性,实验结果见表 3。表中:实验方案 1 为纯视觉模型 YOLOv5-x;方案 2 为基于特征拼接(Concat)的 RV 特征融合模型 RFGNet-YOLOv5-x;方案 3 为基于空间注意力 RV 特征融合模型 RFGNet-YOLOv5-x-SAF;方案 4 为基于 Concat、双尺度 CBAM 和 YOLOv5 检测头的 RV 特征融合模型 RFGNet-YOLOv5-x;方案 5 为基于 Concat、双尺度 CBAM 和解耦检测头的 RV

特征融合模型 RFGNet-YOLOv5-x。

1）RFGNet 的有效性

对比表 3 中方案 1 和方案 2 的实验结果可知，当采用 RFGNet 作为扩展的雷达分支获取到雷达特征，通过 Concat 的方式与图像进行特征融合后，模型在 Test 上的 mAP 相较于基线模型 YOLOv5-x 提高了 4.75 个百分点，而在 Rain 和 Night 上分别提高了 0.93 和 6.89 个百分点。结果表明，引入 RFGNet 生成的雷达特征能够有效提升模型的检测性能。

2）双尺度 CBAM 的有效性

双尺度 CBAM 在通过 Concat 得到融合特征的基础上，进一步自适应预测其潜在的关键特征，从通道上和空间上对融合特征进行重新整合。对比表 3 中方案 2 和方案 4 的实验结果可以看出，基于双尺度 CBAM 的融合方式相比于 Concat 融合，在 Test 上 mAP 提高了 2.62 个百分点，在 Rain 和 Night 上分别提高了 5.05 和 3.13 个百分点。同时，与第 3 组 SAF 模块[16]进行 RV 特征融合的方法相比，在 3 个测试集上 mAP 分别提高了 1.20、1.34 和 2.57 个百分点。实验结果表明，基于双尺度 CBAM 的 RV 特征融合方法优于 Concat 方法和 SAF 方法。

3）解耦检测头的有效性

表 3 中方案 5 采用解耦检测头。与方案 4 采用 YOLOv5 检测头相比，解耦检测模型在 Rain 和 Night 上的检测性能分别提高了 0.42 和 3.07 个百分点。尽管在 Test 上检测性能稍有下降，但也仍保持在相当的水平。结果表明，解耦检测头更适合挑战性环境下的检测。

表 3　所提方法的消融实验结果

Table 3　Ablation experimental results of the proposed method

方案	模型	Concat	双尺度 CBAM(3,7)①	DH②	mAP/%		
					Test	Rain	Night
1	YOLOv5-x				50.64	44.81	39.18
2	RFGNet-YOLOv5-x	✓			55.39	45.74	46.07
3	RFGNet-YOLOv5-x-SAF				56.81	49.45	46.63
4	RFGNet-YOLOv5-x	✓	✓		**58.01**	50.79	49.20
5	RFGNet-YOLOv5-x	✓	✓	✓	57.63	**51.21**	**52.27**

注：✓表示有对应模块或操作。

① (3,7)表示其空间注意力中卷积层的卷积核大小。

②DH 为解耦检测头(decoupled head)。

4）双尺度 CBAM 结构设计的有效性

针对双尺度 CBAM 及卷积核大小的设计进行验证实验，实验结果见表 4。实验分别给出单尺度、双尺度和多尺度 CBAM(multi-scale CBAM, MSCBAM)融合模块在通用、雨天和夜间条件下的检测结果。实验结果表明，双尺度 CBAM 的检测性能优于单尺度以及多尺度模块。在双尺度 CBAM 中卷积核大小分别取(3,5)、(5,7)和(3,7)的对比实验的结果表明，在空间注意力卷积层中双尺度 CBAM(3,7)总体上性能最优。

由以上消融实验结果可知，本文提出的雷达特征生成模块、基于双尺度 CBAM 的特征融合模块和解耦检测头在雨天和夜晚这样的挑战环境中都能有效增强检测性能。

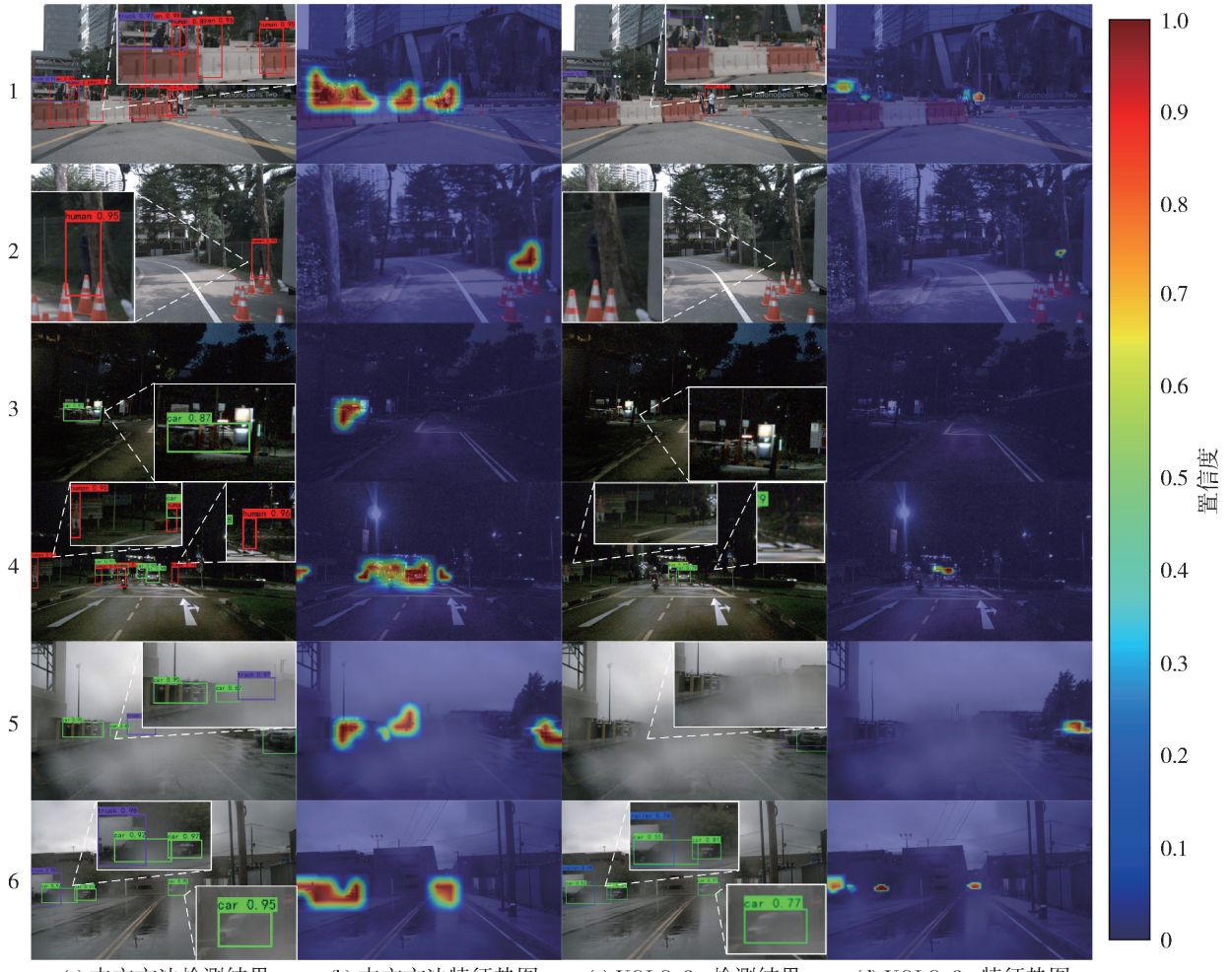
表 4　CBAM 中不同卷积尺度的实验结果

Table 4　Experimental results of different convolutional scales in CBAM

尺度	卷积核大小	mAP/%		
		Test	Rain	Night
单尺度	CBAM(3)	56.88	49.42	50.64
	CBAM(5)	56.42	50.33	48.95
	CBAM(7)	57.29	48.71	46.93
双尺度	双尺度 CBAM(3,5)	57.60	49.37	48.89
	双尺度 CBAM(5,7)	57.26	49.46	**51.21**
	双尺度 CBAM(3,7)	**58.01**	**50.79**	49.20
多尺度	MSCBAM(3,5,7)	56.82	48.62	46.21

2.5 检测结果可视化分析

为了能够更加直观地展示本文方法和纯视觉模型的检测性能,对它们在不同场景下的检测结果以及模型关注区域进行了可视化。图 5 分别展示了白天遮挡、夜间和雨天 3 种挑战环境下,本文方法与 YOLOv8-x 的检测结果和特征热图。在每个图像中,白色框表示局部放大区域,能够更清楚地观察目标检测结果。



(a) 本文方法检测结果 (b) 本文方法特征热图 (c) YOLOv8-x检测结果 (d) YOLOv8-x特征热图

图 5 本文方法与 YOLOv8-x 检测结果可视化对比

Fig. 5 Visual comparison of the detection results between our method and YOLOv8-x

如图 5 所示,相较于纯视觉算法,本文方法在不同挑战性场景下展现了显著的优势。在白天场景下,第 1 组图中路障对行人造成了遮挡,导致纯视觉算法未检测到大部分行人,而本文方法能够准确检测出被遮挡的行人。从相应热图中可以观察到,本文方法能够关注到更多被路障遮挡的行人目标区域。第 2 组图中路边的树遮挡了行人,导致纯视觉算法未检测出行人,而本文方法凭借自身优势检测出了该行人。从其热图可以看到,本文方法关注到了更多被树遮挡的行人目标区域。

在夜间场景下,由于第 3 组图中车辆处于光线较暗的区域,并且受各种障碍物的遮挡,纯视觉算法难以识别出该车辆,而本文方法能够准确识别出该目标的位置和类别。从特征热图中也可以看到,本文方法也可以关注到暗光下的目标。在第 4 组图中,尽管大部分目标都处于光线较暗的区域,但是本文方法仍能够准确检测出大部分目标,优于纯视觉算法。这一点从对应特征热图中也得到了印证。

在雨、雾天气场景下,第 5 组图中的目标几乎无法被纯视觉算法识别,而本文方法借助毫米波雷达的优势准确识别出了完全被遮挡的目标。从特征热图中可以观察到,本文方法也能够关注到完全被雨、雾遮挡的目标区域。在第 6 组图中,尽管纯视觉算法能够检测出被雨、雾遮挡的目标,但本文方法所检测出的目标置信度更高。通过特征热图中颜色的深浅可以明显看出,本文方法对目标的识别具有更高

的置信度。

综上，本文方法通过毫米波雷达数据特征提取和双尺度 CBAM 特征融合方法能够在视觉信息不足甚至人眼也难以识别的条件下关注到目标，实现在挑战性环境下的有效检测。

3　结论

1）针对恶劣天气和低光照等挑战性环境，本文提出了一种基于双尺度 CBAM 的雷达与图像特征融合的目标检测方法。首先，本文采用了基于体素的雷达特征生成网络对雷达数据进行编码和特征提取，以生成雷达特征，用于与图像特征进行多尺度特征融合。其次，在特征融合阶段，使用提出的双尺度 CBAM 对雷达和视觉异构特征分别在通道上以及空间上进行自适应特征优化，改善了融合特征表达，有效提高了网络的检测性能。最后，在检测阶段，网络采用解耦检测头分别对目标进行分类预测和定位，进一步提高了检测性能。

2）在 nuScenes 数据集上，与 SOTA 视觉模型 YOLOv7-x、YOLOv8-x、YOLOX-x 等和代表性 RV 特征融合方法的对比实验和消融实验证明了本文方法的有效性和优越性，尤其是在雨、雾天气条件以及夜间条件下，检测精度得到了较大的提升。

3）由于雷达数据与视觉图像的异构性，雷达数据的预处理也是影响检测性能的重要环节。在未来的工作中，将探索一种通用的雷达数据预处理滤波方法，以提高方法的泛化性。

参考文献：

[1] CHEN Q P, XIE Y F, GUO S F, et al. Sensing system of environmental perception technologies for driverless vehicle: a review of state of the art and challenges[J]. Sensors and Actuators A: Physical, 2021, 319: 112566.

[2] YEONG D J, VELASCO-HERNANDEZ G, BARRY J, et al. Sensor and sensor fusion technology in autonomous vehicles: a review [J]. Sensors, 2021, 21(6): 2140.

[3] FAYYAD J, JARADAT M A, GRUYER D, et al. Deep learning sensor fusion for autonomous vehicle perception and localization: a review [J]. Sensors, 2020, 20(15): 4220.

[4] 张新钰，邹镇洪，李志伟，等. 面向自动驾驶目标检测的深度多模态融合技术 [J]. 智能系统学报，2020，15(4)：758-771.
ZHANG X Y, ZOU Z H, LI Z W, et al. Deep multi-modal fusion in object detection for autonomous driving [J]. CAAI Transactions on Intelligent Systems, 2020, 15(4): 758-771. (in Chinese)

[5] ZHOU Y, LIU L, ZHAO H, et al. Towards deep radar perception for autonomous driving: datasets, methods, and challenges [J]. Sensors, 2022, 22(11): 4208.

[6] 任柯燕，谷美颖，袁正谦，等. 自动驾驶 3D 目标检测研究综述 [J]. 控制与决策，2023，38(4)：865-889.
REN K Y, GU M Y, YUAN Z Q, et al. 3D object detection algorithms in autonomous driving: a review [J]. Control and Decision, 2023, 38(4): 865-889. (in Chinese)

[7] JIAO T Z, GUO C P, FENG X Y, et al. A comprehensive survey on deep learning multi-modal fusion: methods, technologies and applications [J]. Computers, Materials & Continua, 2024, 80(1): 1-35.

[8] YAO S L, GUAN R W, HUANG X Y, et al. Radar-camera fusion for object detection and semantic segmentation in autonomous driving: a comprehensive review [J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(1): 2094-2128.

[9] WEI Z, ZHANG F, CHANG S, et al. MmWave radar and vision fusion for object detection in autonomous driving: a review [J]. Sensors, 2022, 22(7): 2542.

[10] ABDU F J, ZHANG Y, FU M, et al. Application of deep learning on millimeter-wave radar signals: a review [J]. Sensors, 2021, 21(6): 1951.

[11] FENG D, HAASE-SCHÜTZ C, ROSENBAUM L, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: datasets, methods, and challenges [J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(3): 1341-1360.

[12] CHADWICK S, MADDERN W, NEWMAN P. Distant vehicle detection using radar and vision [C] // 2019 International Conference on Robotics and Automation. Piscataway, NJ: IEEE, 2019: 8311-8317.

[13] JOHN V, MITA S. RVNet: deep sensor fusion of monocular camera and radar for image-based obstacle detection in challenging environments [C] // 2019 Pacific-Rim Symposium on Image and Video Technology. Cham: Springer, 2019: 351-364.

[14] NOBIS F, GEISSLINGER M, WEBER M, et al. A deep learning-based radar and camera sensor fusion architecture for object detection[C]//2019 Sensor Data Fusion: Trends, Solutions, Applications. Piscataway, NJ: IEEE, 2019: 1-7.

[15] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition.

Piscataway, NJ: IEEE, 2017: 2117-2125.

[16] CHANG S, ZHANG Y, ZHANG F, et al. Spatial attention fusion for obstacle detection using mmWave radar and vision sensor[J]. Sensors, 2020, 20(4): 956.

[17] LI L Q, XIE Y L. A feature pyramid fusion detection algorithm based on radar and camera sensor[C]//2020 15th IEEE International Conference on Signal Processing. Piscataway, NJ: IEEE, 2020: 366-370.

[18] STACKER L, HEIDENREICH P, RAMBACH J, et al. Fusion point pruning for optimized 2D object detection with radar-camera fusion[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway, NJ: IEEE, 2022: 3087-3094.

[19] ZHOU Y, TUZEL O. VoxelNet: end-to-end learning for point cloud based 3D object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 4490-4499.

[20] SONG G L, LIU Y, WANG X G. Revisiting the sibling head in object detector[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 11563-11572.

[21] CAESAR H, BANKITI V, LANG A H, et al. nuScenes: a multimodal dataset for autonomous driving[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 11621-11631.

[22] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 7464-7475.

[23] GE Z, LIU S T, WANG F, et al. YOLOX: exceeding YOLO series in 2021[EB/OL]. [2023-06-05]. https://arxiv.org/abs/2107.08430.

[24] JOCHER G, CHAURASIA A, QIU J. Ultralytics YOLOv8[EB/OL]. [2023-06-05]. https://github.com/ultralytics/ultralytics.

[25] JOCHER G. Ultralytics YOLOv5[EB/OL]. [2023-06-05]. https://github.com/ultralytics/yolov5.

[26] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//European Conference on Computer Vision. Cham: Springer, 2018: 3-19.

[27] PADILLA R, NETTO S L, DA SILVA E A B. A survey on performance metrics for object-detection algorithms[C]//2020 International Conference on Systems, Signals and Image Processing. Piscataway, NJ: IEEE, 2020: 237-242.

[28] LOSHCHILOV I, HUTTER F. SGDR: stochastic gradient descent with warm restarts[EB/OL]. [2023-06-05]. https://arxiv.org/abs/1608.13983.

[29] ZANG S Z, DING M, SMITH D, et al. The impact of adverse weather conditions on autonomous vehicles: how rain, snow, fog, and hail affect the performance of a self-driving car[J]. IEEE Vehicular Technology Magazine, 2019, 14(2): 103-111.

(责任编辑 梁 洁)