DOI:10.19651/j.cnki.emt.2416756

基于雷达和视觉多级信息融合的目标检测网络*

周志伟 周建江 王佳宾 邓 凯
(南京航空航天大学雷达成像与微波光子技术教育部重点实验室 南京 211100)

摘 要:针对自动驾驶感知任务中由于道路环境复杂、车载雷达和摄像头数据融合不充分导致的一些高危险性动目标检测效果过差的问题,本文在 Centerfusion 的基础之上设计了一种雷达和视觉多层级信息融合的目标检测网络 MLFusionNet。首先在输入层增加了数据级融合,将雷达回波特征以像素值的形式和图像进行拼接后再通过一个二级残差融合模块输入到编解码网络,丰富了网络的输入信息;然后在骨干网络的编码器和解码器之间设计了一种瓶颈结构的上下文模块,通过多分支的卷积结构获取特征图中更广泛的上下文信息,并通过压缩通道的方式降低参数量;最后设计了一种并行注意力融合模块,解决了特征级模态融合不充分的问题。在 nuScenes 数据集上的实验结果表明 MLFusionNet 的 NDS 达到了 46.6%,相比较多模态网络 Centerfusion 汽车、卡车和行人的 mAP 分别提升了 1.4%、3.0%和 1.5%,说明网络更加关注驾驶环境中的高危险性动态目标。

关键词:深度学习;目标检测;特征融合;注意力机制
中图分类号:TP391;TN919.8 **文献标识码**:A **国家标准学科分类代码**:520.60

Target detection network based on multi level information fusion of radar and vision

Zhou Zhiwei Zhou Jianjiang Wang Jiabin Deng Kai
(Key Laboratory of Radar Imaging and Microwave Photonic Technology of Ministry of Education, Nanjing University of Aeronautics and Astronautics, Nanjing 211100, China)

Abstract: In response to the problem of poor detection performance of some high-risk moving targets in autonomous driving perception tasks due to complex road environments and insufficient fusion of onboard radar and camera data, this paper designs an object detection network MLFusionNet that integrates radar and visual multi-level information based on Centerfusion. Firstly, data level fusion is added to the input layer, which concatenates the radar echo features with the image in the form of pixel values, and then inputs them into the encoding and decoding network through a secondary residual fusion module, enriching the input information of the network; then, a bottleneck structured context module was designed between the encoder and decoder of the backbone network, which obtains broader contextual information from the feature map through a multi branch convolutional structure and reduces the number of parameters through compression channels; finally, a parallel attention fusion module was designed to solve the problem of insufficient feature level modal fusion. The experimental results on the nuScenes dataset showed that the NDS of MLFusionNet reached 46.6%, which increased the mAP of cars, trucks, and pedestrians by 1.4、3.0 and 1.5 percentage points respectively compared to the multimodal network Centerfusion. This indicates that the network pays more attention to high-risk dynamic targets in the driving environment.

Keywords: deep learning; target detection; feature fusion; attention mechanism

0 引 言

自动驾驶车辆通常都会进行三维的目标检测,以获取周围目标的状态信息,这其中包括它们的位置、方向、尺寸和速度。感知就像是自动驾驶中的眼睛,是运动预测、路径规划和机动控制等后续功能的基础[1-2]。相机通过记录信号的光强,产生清晰的图像以表达环境信息,但由于相机缺乏物体的位置、尺寸、深度等信息,因此对自动驾驶等应用

收稿日期:2024-08-29
* 基金项目:国家自然科学基金(61502228)项目资助

场景来说有很大局限性[3];而毫米波雷达点云能够提供物体的速度、距离和RCS等信息;融合这两种模态的特征,能够更准确的掌握环境状况,提高自主决策和路径规划的能力。

现有目标检测可以分为基于图像,点云和多传感器融合3种。自动驾驶中基于图像的目标检测方法已经广泛应用于检测车辆、车道线、行人以及交通信号灯等。根据训练步骤的不同,可以分为两阶段和一阶段。两阶段检测算法如SPPNet[4]、Fast R-CNN[5]、Faster R-CNN[6]、FPN[7]等将检测阶段分为生成区域建议和对每个目标的位置进行分类预测两个阶段。两阶段目标检测算法具有较高的准确率和召回率但复杂度相对较高。单阶段检测算法能够同时预测边界框和类别,因此检测速度更快,但精度较低。张建恒等[8]提出了一种融合注意力机制的轻量化道路交通标志检测方法就是一种基于图像的单阶段算法。

基于点云的目标检测方法中文献[9]提出了一种不需要人工提取点云特征的端到端三维目标检测网络VoxelNet。VoxelNet将点云划分为等间距的三维体素,并通过新引入的体素特征编码(voxel feature encoding,VFE)层将每个体素内的点群转换为统一的特征表示,之后与RPN连接,生成检测结果。文献[10]提出了一种融合多注意力机制与PointRCNN的三维点云目标检测,解决了点云格式不规则和密度不均匀的问题。

由于摄像头数据易受环境因素影响,而车载雷达数据又十分稀疏,因此上述两种单模态方案在目标检测任务中往往由于环境因素使得传感器数据不够完整从而导致网络检测精度较差,鲁棒性较低。因此需要考虑利用点云特征来补充和强化图像特征,并通过分析模态的特征优势设计高效的融合模块。

基于融合的方法中,文献[11]提出了一种RRPN算法,通过将雷达探测点映射到图像坐标系,并为每个映射的雷达探测点生成预定义的锚盒,从而生成目标建议以缩小相机图像的检测范围。文献[12]提出了雷达-相机像素深度关联(radar-camera pixel depth association, RC-PDA),这是一种将雷达点云与附近图像像素关联以增强雷达图像的学习方法,他学习从雷达返回到像素的映射。RC-PDA通过输入原始数据,可以利用完整的特征并从中学习联合表示。文献[13]提出了一种基于毫米波雷达和视觉传感器的空间注意融合(spatial attention fusion,SAF)障碍物检测方法,该方法考虑了雷达点的稀疏性,通过3个不同大小卷积核来提取点云特征后与图像逐元素相乘。文献[14]提出了一种两次回归的特征级融合方法,利用摄像头数据预测目标的中心点,并回归得到目标的3D坐标、深度、旋转等信息,然后将雷达检测到的目标数据和上面检测到的目标中心点进行关联完成特征级融合后再次进行回归得到最终结果。上述所有融合网络都是在数据级和特征级一个层级来融合图像和点云。在数据级融合中由于点云极为稀疏且噪声较多,很难与图像做匹配。在特征级融合中,尽管Centerfusion通过两次回归的方式过滤噪声点,但点云特征提取和转换后会丢失大量原始信息。因此,为了利用不同层级的融合优势,本文在Centerfusion的基础之上设计了一种雷达和视觉多层级信息融合的目标检测网络MLFusionNet。

主要工作分为以下三部分:第一,在网络的输入图像数据中增加点云数据,通过将点云投影到图像平面,并在垂直方向上进行拉伸来弥补点云的稀疏性,之后通过一个设计的两级残差结构的注意力融合模块融合点云和图像数据。第二,在特征级融合中,通过一个并行注意力机制模块代替原来直接的通道堆叠融合点云和图像特征,并进一步改进了通道注意力机制强化特征图的边缘和全局特征。第三,由于在Centerfusion中雷达点云和目标的关联依靠第一次回归的结果,所以为了进一步提高初次回归的准确性,通过一个瓶颈结构的上下模块改进图像的骨干网络DLA[15],增强骨干网络对上下文特征的读取能力,获得多感受野下的图像特征。

1 网络结构和算法介绍

1.1 多层级雷达融合网络结构

网络的整体结构如图1所示,主要由4部分组成:

1)解码网络:DLA是一种轻量级的神经网络,它结合了深度分支和跨层特征聚合技术,以实现更好的特征提取和整合,并且DLA相比较ResNet[16]有着更好的性能表现。为了加强DLA编码与解码的关联性,在中间阶段增加了基于RFB[17]改进的上下文模块。

2)和图像的数据级融合:基于文献[18],先将前后多帧的雷达点进行融合弥补毫米波雷达点云的稀疏性。之后通过将点云从二维地平面投影到图像平面中,并在垂直方向上进行拉伸使雷达图像特征在数据上变得更密集。得到的雷达特征图作为图像的额外通道并通过设计的二级残差融合模块输入到DLA网络中。

3)图像的截锥关联:在Centerfusion中提出了一种雷达点云与图像的截锥关联方法。用第一次回归估计的物体深度、尺寸和旋转来创建物体周围的RoI(region of interest),过滤掉与该物体无关的雷达点。

4)和图像的特征级融合:为了更好的融合点云热图和图像特征,设计了一种并行注意力融合模块(parallel attention fusion module,PAFM)代替原来的特征图的直接堆叠,并采用两次池化来改进通道注意力机制以获得更精细的通道特征表达。

1.2 数据级融合

相机与雷达传感器存在特性互补,合理的在网络输入端融合两种传感器数据能够有效增强网络的鲁棒性。因此本文在图像的基础上额外增加雷达点云特征通道并通过一个二级残差融合模块将点云特征融合到图像中。这样能够

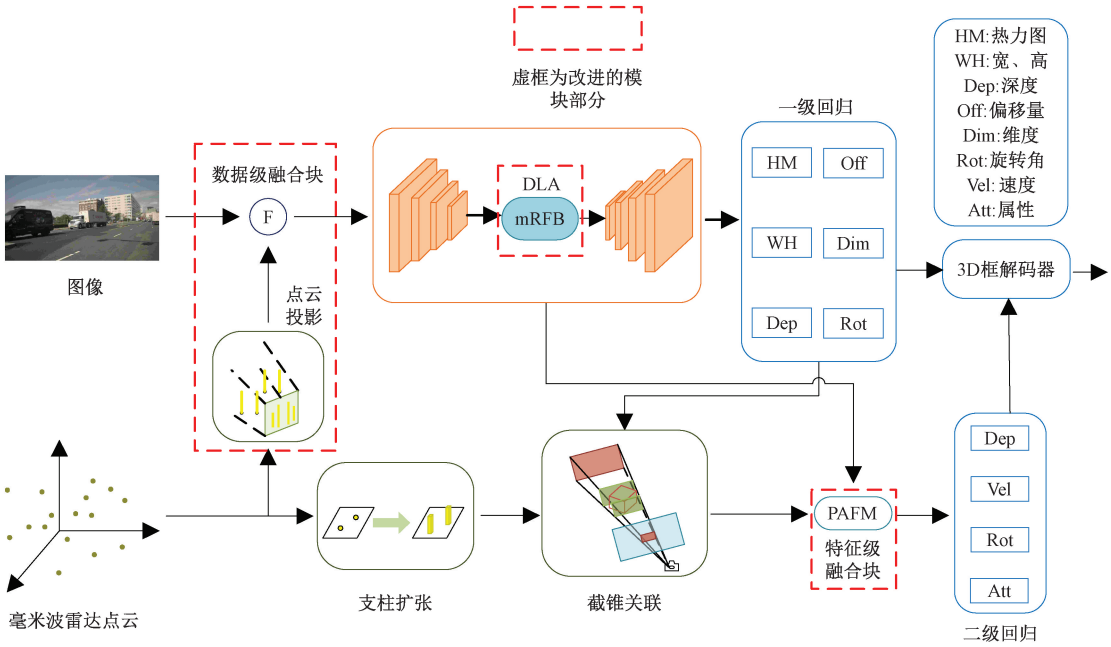
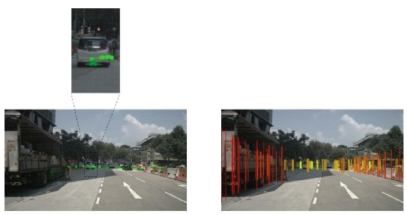


图 1　多层级融合网络 MLFusionNet 的总体网络结构

Fig. 1　Overall network structure of MLFusionNet

使提供初始目标检测结果的主干网络在环境条件恶劣的情况下更加稳健。

SSD 在实际驾驶场景中，目标的速度和深度最能影响网络对目标尺寸和位置的估计，而雷达截面积（radar cross-section，RCS）能够有效区分行人和车辆目标，因此选择点云的深度、横向速度、经向速度和雷达截面作为 4 个特征通道。图 2 为雷达点从二维地平面投影到图像平面后的可视化结果，颜色越浅对应着更大的深度值。为了解决雷达点云的稀疏性问题，将雷达点在垂直方向上进行了拉伸，同时为了强化目标的距离特征，利用较浅的颜色对应更大的深度值，雷达点距离相机原点越远，2D 线的长度则越短。



(a) 点云投影
(a) Projections of point clouds

(b) 垂直方向拉伸图
(b) Vertical stretching diagram

图 2　雷达点云高度拉伸图

Fig. 2　Stretching diagram of radar point cloud height

生成的雷达特征图与图像连接共同输入到网络中。由此得到一个 7 通道的输入特征，增加雷达数据后大大丰富了网络的输入信息，增强了网络的鲁棒性。

为了更好的融合二者特征，额外设计了一种基于二级级残差映射的注意力融合模块，如图 3 所示。模块中所有卷积都采用了深度可分离卷积以此来获取更多的目标信息，减少模块参数。

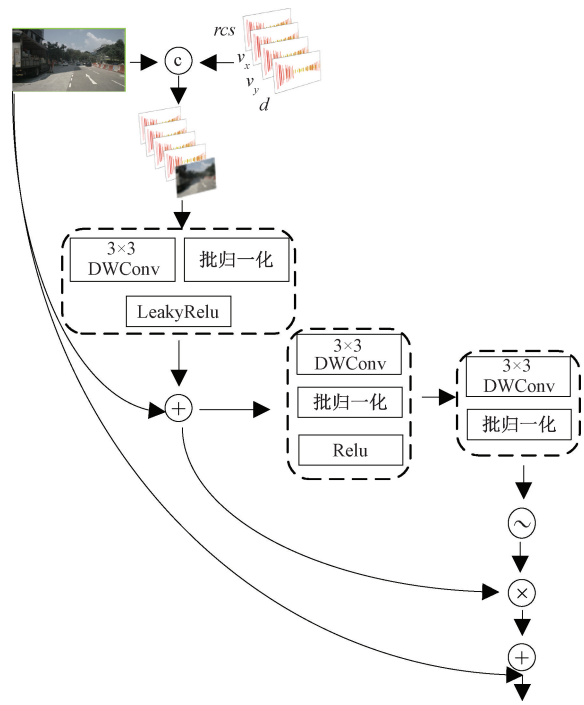


图 3　基于二级残差结构的注意力融合模块

Fig. 3　Attention fusion module based on secondary residual structure

首先将点云数据和图像进行拼接后通过一个深度可分离卷积降低通道数至图像通道数以便后续进行残差连

接,并通过非线性激活函数 LeakRelu 来学习非线性关系。融合后的输出与图像特征进行第一次残差连接,防止图像信息的丢失。

为了更好的保留图像的空间信息,将 SENet 模型中的全连接层替换为深度可分离卷积层。由于点云中可能含有大量的杂波信息,因此为了避免点云噪声过多影响网络性能,将图像和最后输出再次进行残差连接,这样既能强化图像特征防止网络退化又能够缓解梯度消失和加速训练过程中的收敛速度。

1.3 瓶颈结构的上下文模块

感受野指的是神经网络中某一层输出特征图上的一个像素点对输入数据的区域的影响范围。小的感受野用于描述局部信息,而大的感受野则用来描述全局信息。为了捕捉解码器输出特征图中更广泛的上下文信息,本文在主干网络的编解码器之间增加了一个上下文关联模块(bottleneck structure receptive field block, BsRFB),并通过瓶颈结构的设计降低模块的参数量。模块的具体结构如图 4 所示。

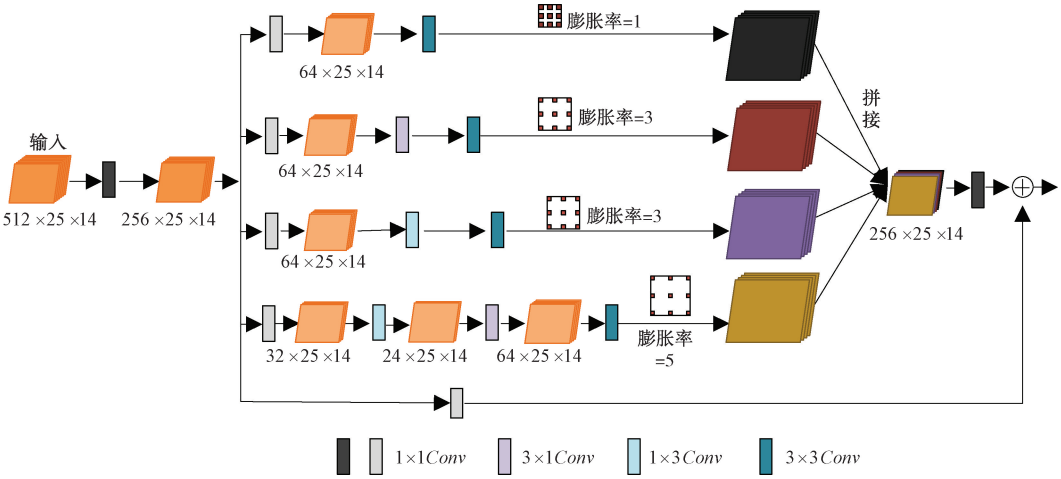


图 4 BsRFB 网络结构

Fig. 4 Structure of BsRFB

BsRFB 由 5 个分支组成,每一个分支都采用瓶颈结构的设计,即首先都会使用一个 1×1 单位卷积来减少通道数量,之后在分别级联一个普通卷积和膨胀率不同的空洞卷积来得到不同感受野大小下的特征图。其中不同卷积核大小的卷积来获得不同的感受野,而不同膨胀率的空洞卷积则能够调整卷积核的采样范围,增大感受野而不增加参数数量,以捕捉更广泛的上下文信息。这使得模型能够更有效地处理图像边缘或角落部分的对象,增强了对于位置变化的鲁棒性。在第 5 条支路中单独通过一个 1×1 来获得特征图的全局特征。最后将不同感受野下的特征图进行堆叠与第 5 路的全局特征进行相加。

多分支的卷积会增加大量的计算资源,因此需要进一步提高网络的计算效率。卷积层的参数量只与网络结构有关,与输入数据无关,考虑到轻量化的设计,在多分支卷积前再使用一个 1×1 的单位卷积来压缩通道数量。输出端使用 1×1 的卷积将通道数复原。卷积计算参数量的公式为:

$$param = (k^2 C_{in} + bias) C_{out} \tag{1}$$

式中:k 是卷积核的尺寸;C_{in} 是输入通道数;C_{out} 是输出通道数;$bias$ 为偏置,默认为 1。

设计的瓶颈结构通过 1×1 的卷积将通道数量降为原来的 1/2,在输出前通过 1×1 的卷积恢复通道数。尽管瓶颈结构压缩了通道数,但通常能够达到与之前相似甚至更好的训练效果,因为瓶颈结构本身是通过额外的卷积来实现,因此也增加了网络深度,从而提升了网络的提取特征能力。表 1 显示了使用瓶颈结构前后模块的参数情况,可以看出,外瓶颈结构的设计能够进一步使参数量降低 63.31%。

1.4 并行注意力特征融合

雷达点云的特征信息通过视锥关联得到,如图 5 所示。通过构建 3D 视锥体,缩小点云和对应目标的匹配范围,进而加快了匹配速度。其中为了解决估计深度值 存在误差问题,添加一个参数 来控制截锥大小。当匹配多个点云时,取深度最小的点来完成锥体内的点云去重。

通过视锥关联得到的雷达特征与主干解码器的输出的特征图共同作为融合模块的输入。

在特征级融合阶段,由于点云在关联后更为稀疏,且通道数量相比较图象特征更少,因此简单的通道堆叠不能够很好利用点云信息来强化图像特征,为了更好的融合两种特征图,本文设计了一种并行注意力融合模块如图 6 所示。

该模块主要有以下 3 个优势:

1)利用深度可分离卷积代替传统卷积来提取两种模态特征堆叠后的特征图和压缩通道。深度可分离卷积有

表 1　参数量对比

Table 1　Comparison of parameter amount

结构分类	具体结构(C_{in},C_{out},k)	参数量
普通结构	Conv(512,128,1)×3 Conv(512,64,1) Conv(512,512,1) Conv(128,128,(3×1))×2 Conv(64,48,(1×3)) Conv(48,128,(3×1)) Conv(128,128,3)×4	881 520
瓶颈结构	Conv(512,256,1) Conv(256,64,1)×3 Conv(256,32,1) Conv(256,256,1) Conv(64,64,(3×1))×2 Conv(32,24,(1×3)) Conv(24,64,(3×1)) Conv(64,64,3)×4	323 448

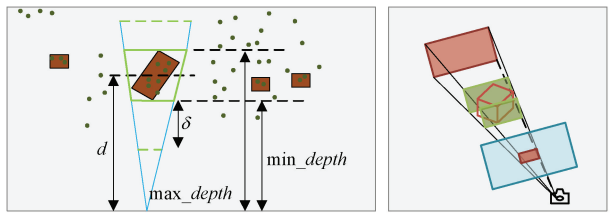


图 5　截锥关联模块

Fig. 5　Visual cone correlation module

着更少的参数和更优的特征提取效果，能够更好的捕捉点云和图像特征堆叠后的空间信息和通道信息。

2)设计了一种并行双重注意力机制，将通道注意力机制和空间注意力机制并行连接，这能够同时考虑通道和空间信息，有助于获取更加综合的特征表达，提升特征提取的效果。同时并行连接可以在两个维度上同时建模，大大强化了特征交互。

3)改进了一种高效的通道注意力机制，全局最大池化层(global max pooling，GMP)和全局平均池化(global average pooling，GAP)分别用来获取纹理信息和背景信息，同时采用两种池化方式能够获得更精细的通道特征信息。

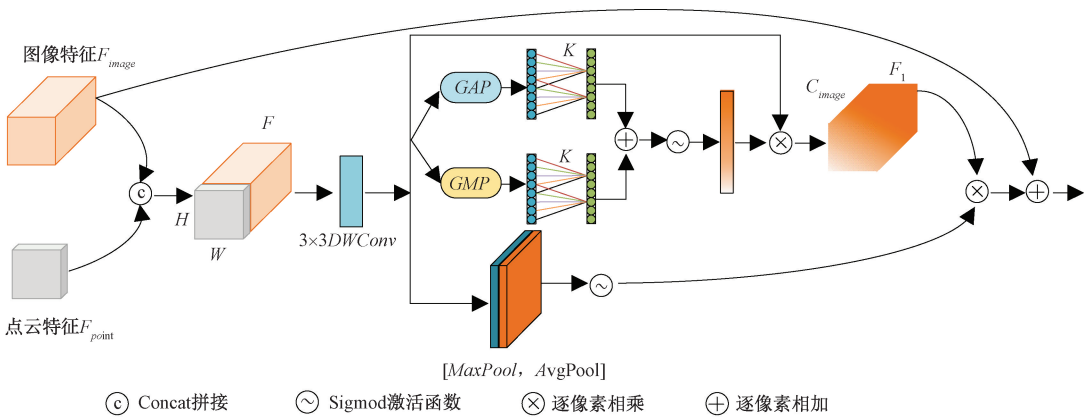


图 6　PAFM 模块结构

Fig. 6　Structure of PAFM

具体的在通道注意力机制中，分别通过 GAP 和 GMP 操作后获得聚合特征，再同时通过一个自适应卷积核大小为 的快速 1D 卷积来生成通道权重。K 是通过通道维度的映射自适应确定的，如式(2)所示，其中 $|t|_{odd}$ 表示 t 最接近的奇数，b 和 γ 分别设置为 1 和 2。经过通道注意力机制后能够获得每个通道的权重，区分特征图的背景和纹理信息，同时强化稀疏点云的通道特征。

$$K=\psi(C)=\left|\frac{\log_2(C)}{\gamma}+\frac{b}{\lambda}\right|_{odd} \tag{2}$$

在空间注意力机制中，沿着通道维度利用全局平均池化和最大池化学习两种空间权重，之后经过 Sigmoid 进行激活，获得两种特征图的权重概率分布后与通道注意力的输出相乘获得通道和空间两个维度的特征。

PAFM 能够更好的融合两种模态特征，过滤背景，强化纹理信息，关注特征图中的可能存在目标的重点区域。

2　实验分析

2.1　数据集和实验参数

实验网络均采用预训练的方式，在 nuScenes 数据集上进行训练、验证和测试。nuScenes 是第一个携带完全自动驾驶汽车传感器套件 3D 目标检测数据集，拥有 6 个摄像头，5 个雷达和一个激光雷达，都具完整的视野。

由于数据集较大，因此只采用前向的雷达和摄像头数据，将整个数据集划分为训练集、验证集和测试集。考虑到点云的稀疏性，通过扫描前后三帧数据来加强点云的特征表达。

使用 Pytorch 框架在两块 Nvidia GeForce RTX 3090 设备上进行分布式训练。模型有 AdamW 优化器训练,损失权重为 0.1,学习率为 2.5×10^{-4}。训练在 60 次迭代时收敛,在第 30 次迭代时学习率衰减至 1/10。批量大小设置为 16。

2.2 评估指标

nuScenes 数据集的检测任务评价指标主要包括:平均精度均值(mean average precision,mAP),nuScenes 检测分数(nuscenes detection scores,NDS)。其中计算 mAP 需要计算精确度 P 和召回率 R,计算公式如下:

$$P=\frac{TP}{TP+FP} \tag{3}$$

$$R=\frac{TP}{TP+FN} \tag{4}$$

其中,TP 为真实样本且预测为真实样本,FP 为错误样本而预测为真实样本,FN 为真实样本而预测为错误样本。平均精度(average precision,AP)计算如下:

$$AP=\int_0^1 P(R)\mathrm{d}R \tag{5}$$

mAP 是所有类别的 AP 之和,如式(7)所示,其中 n 为类别数:

$$mAP=\frac{\sum_{i=1}^{n}AP_i}{n} \tag{6}$$

mAP 作为一项最流行的目标检测指标,并不能够衡量 nuScenes 检测任务的所有方面如速度和属性估计,因此增加一项指标 NDS。NDS 根据 mAP 以及平局度量 mTP 计算得到,其中 mTP 包括 5 个小类,分别是:平均平移误差(average translation error, ATE)、平均尺度误差(average scale error, ASE)、平均角度误差(average orientation erroe, AOE)、平均速度误差(average velocity error, AVE)和平均属性误差(average attribute erroe, AAE)。NDS 的计算如式(7)所示。

$$NDS=\frac{1}{10}\left[5mAP+\sum_{mTP\in TP}(1-\min(1,mTP))\right] \tag{7}$$

2.3 整体网络效果

为了评估模型的目标检测效果,与其他流行的单模态和多模态网络进行了对比。其中 CenterNet 与 FCOS3D 为纯视觉方案,RadarDistill 为基于毫米波雷达点云方案,Centerfusion 为融合方案也是本文的改进的网络结构。具体的目标检测性能如表 2 所示,其中用粗体表示了最优的检测效果。

表 2 NuScenes 数据集下不同网络模型的目标检测性能对比

Table 2 Comparison of performance of different network models of target detection on Nuscenes

网络模型	模态	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE	网络大小/MB
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	
RadarDistill	R	0.437	0.205	**0.461**	0.263	0.525	**0.336**	**0.072**	
CenterFusion	R+C	0.454	0.321	0.655	0.271	0.460	0.520	0.153	245
MLFusionNet	R+C	**0.466**	0.333	0.646	0.260	**0.429**	0.521	0.148	250

从表 2 的实验结果可以看出,本文所设计的网络 NDS 相比较纯视觉网络 CenterNet[19] 和 FCOS3D[20] 提高了 6.6%和 3.8%,虽然 FCOS3D 的 mAP 优于本文网络,但在目标的速度、角度和尺度估计方面性能表现过差不能够适应复杂的交通环境。本文网络 NDS 对比最先进的纯点云方案 RadarDistll[21]提高了 2.9%。对比最近的多模态网络 Centerfusion 也是本文的基线网络,NDS 和 mAP 都提高了 1.2%,并且有着更好的位置、尺度、角度和属性估计,而整体网络的大小仅仅增加了 2%。

表 3 列出了网络在所有类别的精度表现,可以看出,设计的多层级融合网络对实际道路中较为危险的动态目标给予了更高的关注。表现在汽车、卡车、以及行人的检测精度相比较多模态网络 Centerfusion 分别提升了 1.4%、3.0%和 1.5%。这得益于 MLFusionNet 在数据级和特征级两个层级中都融合了毫米波雷达的点云数据,而点云特征中包含了目标的横向和经向速度属性,能够很好的区分出动静态目标。并且 MLFusionNet 在融合模块中都运用了注意力机制,这在自动驾驶环境中能够更好的预测目标的姿态和类别并且更关注较为危险的目标如行人和障碍物,方便决策端做出合理得路径规划。

综上所述,本文所设计的网络综合性能对比最先进的单模态网络表现出明显优势。对比多模态网络 Centerfusion 也是本文的 baseline 也有显著的性能提升,表现在更加关注道路中高危险性的动态目标,这也更符合安全性判断的需求。

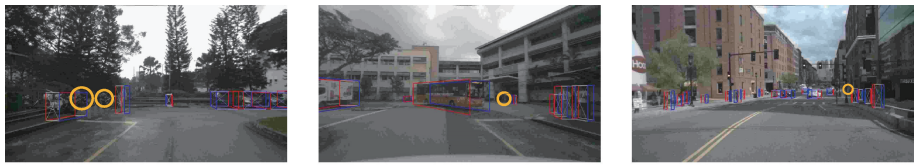
为了更加直观的感受 MLFusionNet 的目标检测效果,图 7 可视化了不同场景下的 CenterFusion 和 MLFusionNet 的目标检测表现。

在第 1 个场景在雨天有遮挡的情况下可以看到在 MLFusionNet 能够准确检测到被遮挡住部分的摩托车掉部分的摩托车。第 2 和第 3 个场景分别是拐弯和道路正常行驶的场景,MLFusionNet 能够跟很好的检测到远距离的

表 3　不同网络模型对不同类别的目标检测精度对比

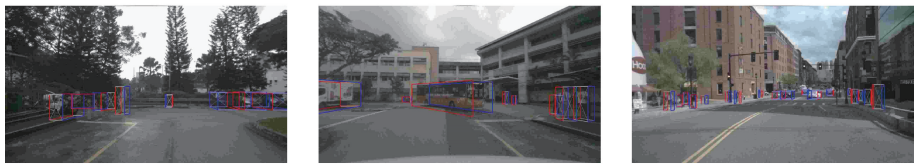
Table 3　Comparison of detection accuracy for different Categories of targets by different network models

网络模型	模态	Car	Truck	Bus	Trailer	Const	Pedest	Motor	Bicycle	Traff	Barrier
CenterNet	C	0.484	0.231	**0.340**	0.131	0.035	0.377	0.249	0.234	0.550	0.456
FCOS3D	C	0.524	0.270	0.277	0.255	0.117	0.397	**0.345**	**0.298**	0.557	0.538
RadarDistill	R	**0.540**	0.153	0.113	**0.295**	0.055	0.092	0.153	0.009	0.217	0.423
CenterFusion	R+C	0.526	0.259	**0.337**	0.132	0.039	0.392	0.283	0.235	0.554	0.453
MLFusionNet	R+C	**0.540**	**0.289**	0.340	0.133	**0.057**	**0.407**	0.300	0.236	**0.560**	**0.466**



(a) Centerfusion检测结果

(a) Centerfusion test results



(b) MLFusionNet检测结果

(b) MLFusionNet test results

图 7　对比检测结果

Fig. 7　Comparision test results

行人。自动驾驶的感知任务中，行人和摩托车作为道路上不确定性较高的两个关键因素，准确地识别和理解它们对于确保驾驶安全至关重要。MLFusionNet 在处理这类具有较高危险性和不确定性的环境因素时展现了出色的性能。其独特的能力使得它能够有效聚焦于潜在的危险情况，提供更加精准和可靠的感知能力，从而大大提升了自动驾驶系统在复杂交通环境中的安全性和稳定性。

2.4　消融实验

为了验证设计的每一种模块的有效性，进行了消融实验。表 4 的消融实验结果分别展示了各模块对 MLFusionNet 网络的贡献。表 4 中：基线表示原始的 Centerfusion 网络；BsRFB 表示使用瓶颈结构；多级融合表示同时利用设计的融合模块在数据级和特征级融合数据。PAFM 表示使用设计的并联的注意力模块进行特征级别融合否则只是进行特征图堆叠操作。

表 4　各模块消融实验结果

Table 4　Ablation experiments on each module

基线	RFB	BsRFB	多级融合	PAFM	mAP	NDS
√					0.321	0.450
√	√				0.322	0.457
√		√			0.323	0.458
√			√		0.327	0.463
√				√	0.327	0.460
√		√	√	√	0.333	0.466

从表 4 第 2、3 组实验可知在 RFB 中使用瓶颈结构的设计即使输入模块时减少通道数也并没有影响网络的整体性能，反而使得 mAP 和 NDS 提升了 0.1%。这进一步说明了，通过瓶颈结构压缩中间过程的通道数量，降低参数量的操作并没有影响网络的性能，反而通过增加网络深度，提高了网络的整体性能。

第 4 组实验可以看到在使用多级融合后 NDS 提升了 0.9%，mAP 提升了 0.6%，即网络不仅能够更好的检测到目标并且能够更好的拟合和估计目标的状态，这得益于雷达点云特征包含目标的距离，速度以及 RCS 信息。

第 5 组实验看到，使用 PAFM 模块代替简单的特征堆叠，mAP 和 NDS 都有了明显的提升。表明并行注意力机制的引入能够更好的完成特征图的融合，使得网络能够提取到关键信息。

3　结　　论

了更好的融合雷达点云和图像数据，提高自动驾驶任务中的目标检测性能，设计了一种雷达和视觉多层级信息融合的目标检测网络，通过数据级和特征级这种多层级融合方式更有效的利用多模态优势，增强了网络整体的鲁棒性。同时在编解码网络之间设计了一种瓶颈结构的上下

文模块,降低模块参数量的同时增强了 DLA 编码器和解码器之间的关联性。实验结果表明以上改进均能够有效提高网络的目标检测性能。由于多模态的融合任务往往伴随着模型的复杂化,因此在未来的工作中考虑对网络整体结构进行轻量化处理。

参考文献

[1] PADEN B, CAP M, YONG S ZH, et al. A survey of motion planning and control techniques for self-driving urban vehicles[J]. IEEE Transactions on Intelligent Vehicles, 2016, 1(1): 33-55.

[2] XIAO Y, CODEVILLA F, GURRAM A, et al. Multimodal end-to-end autonomous driving[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 23(1): 537-547.

[3] 吴文涛,何赟泽,杜旭,等.融合相机与激光雷达的目标检测与尺寸测量[J].电子测量与仪器学报,2023,37(6):169-177.
WU W T, HE Y Z, DU X, et al. Fusing camera and Lidar for object detection and dimensional measurement[J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(6): 169-177.

[4] HE K M, ZHANG X Y, REN SH Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[5] GIRSHICK R. Fast RCNN[C]. IEEE International Conference on Computer Vision, 2015: 1440-1448.

[6] REN SH Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards realtime object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149.

[7] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2117-2125.

[8] 张建恒,杨智宇,夏利红,等.融合注意力机制的轻量化道路交通标志检测方法[J].电子测量技术, 2023, 46(21):85-92.
ZHANG J H, YANG ZH Y, XIA L H, et al. Lightweight road traffic sign detection method with attention mechanism [J]. Electronic Measurement Technology, 2023, 46(21): 85-92.

[9] ZHOU Y, TUZEL O. Voxelnet: End-to-end learning for point cloud based 3d object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4490-4499.

[10] 郑美琳,高建瓴.融合多注意力机制与 PointRCNN 的三维点云目标检测[J].电子测量技术,2022,45(9):127-132.
ZHENG M L, GAO J L. 3D point cloud target detection based on attention mechanism and pointrcnn[J]. Electronic Measurement Technology, 2022, 45(9): 127-132.

[11] NABATI R, QI H R. Rrpn: Radar region proposal network for object detection in autonomous vehicles[C]. 2019 IEEE International Conference on Image Processing(ICIP). IEEE, 2019: 3093-3097.

[12] LONG Y F, MORRIS D, LIU X M, et al. Radar-camera pixel depth association for depth completion[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 12507-12516.

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

[14] NABATI R, QI H R. Centerfusion: Center-based radar and camera fusion for 3d object detection[C]. IEEE/CVF Winter Conference on Applications of Computer Vision, 2021: 1527-1536.

[15] YU F, WANG D Q, SHELHAMER E, et al. Deep layer aggregation[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 2403-2412.

[16] HE K M, ZHANG X Y, REN SH Q, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[17] LIU S T, HUANG D. Receptive field block net for accurate and fast object detection[C]. European Conference on Computer Vision (ECCV), 2018: 385-400.

[18] NOBIS F, GEISSLINGER M, WEBER M, et al. A deep learn-ing-based radar and camera sensor fusion architecture for object detection[C]. 2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF). IEEE, 2019: 1-7.

[19] ZHOU X Y, WANG D Q, KRAHENBUHL P. Objects as points[J]. ArXiv preprint arXiv: 1904.07850, 2019.

[20] WANG T, ZHU X G, PANG J M, et al. FCOS3D: Fully convolutional one-stage monocular 3d object detection[C]. IEEE/CVF International Conference on Computer Vision, 2021: 913-922.

[21] BANG G, CHOI K, KIM J, et al. RadarDistill: Boosting radar-based object detection performance via knowledge distillation from LiDAR features[J]. ArXiv preprint arXiv:2403.05061, 2024.

作者简介

周志伟,硕士研究生,主要研究方向为深度学习、传感器融合、目标检测。

E-mail:937611989@qq.com

周建江(通信作者),博士,教授,主要研究方向为射频隐身技术、雷达目标特性分析。

E-mail:zjjee@nuaa.edu.cn

王佳宾,硕士研究生,主要研究方向为雷达信号处理、目标分类。

E-mail:522035274@qq.com

邓凯,硕士研究生,主要研究方向为雷达信号处理、4D 毫米波雷达目标检测。

E-mail:1741782039@qq.com