基于伪点云特征增强的多模态三维目标检测方法

孔德明 李晓伟 杨庆鑫
(燕山大学电气工程学院 河北 秦皇岛 066004)

摘 要 环境感知是自动驾驶汽车落地的关键技术之一,它对于提高自动驾驶汽车的安全性和可靠性至关重要.三维目标检测是其中的一项核心任务,旨在识别和定位三维空间中的物体,为后续决策提供重要的信息.点云和图像是该任务最常用的输入数据,点云由三维空间中不规则分布的点组成,而图像则是由二维空间上规则分布的像素组成.因此,点云和图像之间难以进行有效的融合.而伪点云作为一种点云表征的图像信息,近几年受到了该领域学者的广泛关注.现阶段基于伪点云的三维目标检测方法还存在伪点云特征提取粗糙和相应感兴趣区域(Region-of-Intersts, RoI)特征表征能力差的问题.本文针对上述问题开展研究,分别提出细粒度注意力卷积和多尺度分组稀疏卷积.细粒度注意力卷积将规则图像处理中常用的深度可分离卷积引入不规则点云的处理流程,并在此基础上嵌入通道和分组注意力机制,进行精细的特征提取,增强伪点云特征;多尺度分组稀疏卷积将格网池化后的RoI特征分组,进行差异化特征学习,获取不同尺度的RoI特征,增强伪点云RoI格网特征的表征能力.基于此,本文在SFD(Sparse Fuse Dense)网络的伪点云特征提取流程中引入细粒度注意力卷积,同时在其伪点云RoI特征学习流程中引入多尺度分组稀疏卷积,构建SFD++多模态三维目标检测网络.在权威KITTI自动驾驶数据集上的实验结果表明,SFD++每秒可以处理8.33帧数据,其精度在简单、中等和困难的三维汽车检测上达到95.74%、88.80%和86.04%,比次优SFD的精度高出0.15%、0.84%和0.58%.除此之外,一系列消融和补充实验结果验证了所提出卷积的有效性和相关参数设置的合理性.

关键词 自动驾驶;三维目标检测;伪点云;注意力机制;深度可分离卷积;组卷积
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2024.00759

Multimodal 3D Object Detection Method Based on Pseudo Point Cloud Feature Enhancement

KONG De-Ming LI Xiao-Wei YANG Qing-Xin
(*School of Electrical Engineering, Yanshan University, Qinhuangdao, Hebei* 066004)

Abstract Environment perception is one of the key technologies to ensure the landing of autonomous vehicle, and it is crucial to improve the safety and reliability of autonomous vehicle. Three dimensional (3D) object detection is a core task in environment perception technology. It aims at identifying and locating different objects in 3D space, and provides important information for the subsequent decision-making and action of autonomous vehicle. Point clouds and images are the most commonly used input data for 3D object detection task. However, point clouds are composed of irregularly distributed scattered points in 3D space, while images are composed of regularly distributed pixels in 2D space. Therefore, it is difficult to map and fuse irregularly distributed point clouds and regularly distributed image pixels in an effective way. In recent years, as a type of image information input in the form of point cloud, image pseudo point clouds have received widespread attention from researchers in this field. But 3D object detection methods

收稿日期:2023-04-19;在线发布日期:2024-01-09. 本课题得到国家自然科学基金面上项目(No. 62173289)、航空科学基金(No. 20200016099002)和国家自然科学基金青年科学基金(No. 61501394)资助. **孔德明**,博士,教授,主要研究领域为计算机视觉、光谱信息分析和激光雷达数据处理. E-mail:demingkong@ysu.edu.cn. **李晓伟(通信作者)**,博士,主要研究领域为计算机视觉、深度学习和三维环境感知. E-mail:qwertycn@hotmail.com. **杨庆鑫**,硕士研究生,主要研究领域为计算机视觉、深度学习.

based on image pseudo point clouds still have some issues. On the one hand, the feature extraction process for image pseudo point clouds is still relatively rough. On the other hand, the representation ability of Region of Interest (RoI) features of image pseudo point clouds is still poor. This paper conducted research on the above two issues, proposed fine-grained attention convolution and multi-scale group sparse convolution, respectively. Fine-grained attention convolution introduces the commonly used depth-wise separable convolution in regular image processing into irregular point cloud processing process. On this basis, channel attention mechanism and group attention mechanism are embedded for fine image pseudo point cloud feature extraction, and to enhance image pseudo point cloud features. This kind of convolution can enhance the fine-grained information of image pseudo point cloud features. Multi-scale group sparse convolution groups the image pseudo point cloud RoI features after grid pooling, performs differential feature learning on the grouped RoI features to obtain RoI features at different scales. This kind of convolution can enhance the representation ability of RoI features grouped from image pseudo point clouds. On this basis, this paper constructed SFD++ multi-modal 3D object detection network, which introduced the proposed fine-grained attention convolution into the feature extraction process of image pseudo point cloud, and introduced multi-scale group sparse convolution into its RoI features learning process of image pseudo point cloud in SFD (Sparse Fuse Dense) 3D object detection network. Experiments were carried out on the authoritative KITTI autonomous driving dataset, and the results show that the constructed SFD++ can process 8.33 frames of data per second. It achieved an average precision of 95.74%, 88.80%, and 86.04% in easy, moderate and hard 3D car detection subsets, respectively. The achieved average precision is 0.15%, 0.84%, and 0.58% higher than the current second-best 3D object detection network SFD. In addition, a series of ablation experiments and supplementary experiments were also conducted on KITTI datasets, the results verified the effectiveness of the proposed fine-grained attention convolution and multi-scale group sparse convolution, as well as the rationality of related parameter settings.

Keywords autonomous driving; 3D object detection; pseudo point cloud; attention mechanism; depth-wise separable convolution; group convolution

1 引 言

自动驾驶系统主要由环境信号获取、环境感知和决策控制三部分组成.环境感知主要为自动驾驶汽车提供“视觉”功能,是该系统的重要组成部分.在环境感知阶段,自动驾驶汽车通过分析摄像机和激光雷达等传感器获取到的数据,对周围环境进行高精度的感知.目标检测是环境感知阶段的核心任务,关于二维目标检测的研究已在计算机视觉领域取得瞩目的成果[1-2].但在自动驾驶场景下,二维信息无法为自动驾驶汽车提供环境目标的三维位置、尺寸和行车方向等重要信息,从而无法保障行车安全.由此可见,对自动驾驶场景下的三维目标检测任务开展研究具有重要的现实意义.

图像和点云是三维目标检测网络中常用的两种输入数据.基于激光雷达点云中蕴含的丰富三维信息,研究者可以获取高质量的感兴趣区域(Region-of-Intersts,RoI),从而获得较好的检测结果,因此点云三维目标检测方法在过去几年保持着较高的检测精度.由于图像缺失深度信息,图像三维目标检测方法通常先通过图像进行深度估计,生成伪点云;再使用点云三维目标检测网络进行后续处理,完成三维目标检测.然而深度估计过程中会不可避免地引入误差,现有研究表明[3],单独以图像伪点云作为输入难以进行准确的三维目标检测,因此图像三维目标检测方法的检测精度通常较低.不过,在点云三维目标检测网络能输出准确 RoI 的情况下,伪点云特征可在 RoI 中与点云特征进行融合,提供必要的图像信息,从而提升三维目标检测精度.基于该思路,

Wu 等人[4]构建 SFD,取得了目前最优的三维目标检测精度,但该网络在伪点云处理过程中还存在以下问题:

(1)PointNet[5]是目前最常用的点云特征提取结构,SFD 基于 PointNet 构建伪点云特征提取网络,但 PointNet 直接采用多层感知机(Multi-Layer Perception,MLP)进行表征学习,提取到的特征通常较粗糙.因此,该结构无法提供细粒度的图像伪点云特征,影响伪点云 RoI 特征的表征能力,进而影响最终的检测精度;

(2)在三维目标检测领域,格网池化是目前效果较好的一类 RoI 池化方式.该池化方式通常预设固定数量的格网点,将骨干网络获得的高级特征聚合至格网点,再进行后续 RoI 级别的特征学习.但由于实际场景中的物体尺寸不同、距离远近不同,其伪点云分布也不同,固定尺寸的格网难以满足不同伪点云分布 RoI 的特征学习,对多尺度信息的表征较差,这将影响三维目标检测精度.

针对问题(1),本文拟引入注意力机制实现精细特征提取.近几年注意力机制已被证明是一种有效的学习技巧,许多研究者提出了多种多样的注意力机制应用方式,提升了各类视觉和语言任务的精度.但现有方法如 SENet[6]和 CBAM[7]的注意力权重通常是直接从输入特征中学习得到的,虽然 SFD 中生成的点云特征权重较粗糙,但对于点云数据来说,其采用的分组距离残差相比点云特征更适合进行精细的注意力权重学习,进而指导点云特征的精细调整.此外,深度可分离卷积是一种经典的卷积解耦方式[8],它将常规卷积分解为深度卷积和点卷积,初衷是降低常规卷积的计算量.但这种解耦方式将空间相互作用和通道相互作用区分开,为伪点云特征的精细化提取提供了条件,在该过程中引入注意力机制可以方便地实现对伪点云特征的针对性调整.

针对问题(2),本文拟基于组卷积实现伪点云 RoI 格网特征的差异化和多尺度学习.对 RoI 进行多次不同尺寸的格网池化是获取多尺度 RoI 格网特征最直接的方法,但这种做法的计算量较大.组卷积通过将特征进行分组,然后对各个分组特征进行卷积,可以避免多次 RoI 格网池化引入的计算消耗.在每个分组特征的卷积过程中设计不同的卷积步长,可以方便地实现 RoI 格网特征的差异化学习,获取多尺度的 RoI 格网特征.

基于以上分析,本文提出细粒度注意力卷积(Fine-grained Attention Convolution,FAC)和多尺度分组稀疏卷积(Multi-scale Group Sparse Convolution,MGSC)增强伪点云特征,从特征提取和 RoI 特征学习两方面改进 SFD.FAC 将图像视觉任务中常用的深度可分离卷积引入点云视觉任务中,将点云特征提取中常用的常规卷积分解为深度卷积和点卷积,并在此过程中嵌入两种注意力机制.相比于平等看待点云中各点的常规卷积,FAC 根据各点的注意力权重,分别在伪点云通道维度和分组维度上进行特征调整,可以精细地提取伪点云特征.MGSC 先是借鉴组卷积的思想对 RoI 格网特征进行分组,再使用不同的卷积步长对分组特征进行卷积计算,并基于此自适应地调整分组后的 RoI 特征.相比于常规的稀疏卷积,MGSC 通过不同尺度的卷积计算,可以获取多尺度的 RoI 格网特征.

本文的贡献可以总结如下:

(1)本文提出一种针对点云数据表征的卷积算子 FAC,结合注意力机制和深度可分离卷积进行精细的特征提取,增强伪点云特征的表征能力.

(2)本文提出一种针对 RoI 格网特征学习的卷积算子 MGSC,基于组卷积进行多尺度特征学习,增强伪点云 RoI 格网特征的表征能力.

(3)基于 FAC 和 MGSC,本文构建 SFD++三维目标检测网络,在权威 KITTI 自动驾驶数据集上的实验表明,SFD++取得了具有竞争力的三维目标检测精度,在保持检测效率的基础上提升了 SFD 的三维目标检测精度,且优于 3D-CVF[9]和 SRIF-RCNN[10]等多模态三维目标检测方法.

2　相关工作

随着人工智能理论的发展,深度学习技术(Deep Learning,DL)和神经网络(Neural Network,NN)已在目标检测领域淘汰传统的人工建模方法,成为相关研究的首选工具.因此,本节主要介绍基于深度学习技术的三维目标检测方法.从输入数据模态的角度来看,现阶段基于深度学习技术的三维目标检测方法可以分为图像、点云和多模态三维目标检测方法.

2.1　图像三维目标检测方法

由于图像缺少三维空间中的深度信息,一些研究者首先尝试从图像中恢复出深度信息,然后再进行三维目标检测,如 Xu 等人[11]和 Weng 等人[12]使用单目图像进行深度估计,而 Pseudo-LiDAR 系列方法[3,13-14]则使用双目图像进行深度估计.Chen 等人[15]提出 DSGN,将二维图像转化至一种可微分的三维几何结构(3D Geometric Volume,3DGV),该

结构可以有效地将深度分布编码到统一的三维目标检测网络中,并进行联合深度估计和三维目标检测;Reading 等人[16]提出 CaDDN,对深度区间进行离散化操作,将连续值回归问题转化为离散值回归问题,减小了深度估计的难度,并提出一种视锥特征网络来预测带有深度信息的特征图,以此辅助三维目标检测.

还有一些研究者借助二维和三维空间之间的映射关系进行三维目标检测,通常做法为先从图像中获取二维边界框,再根据映射关系将其映射至三维空间得到三维边界框. Li 等人[17]提出 Gs3d,利用"二维边界框上边的中点应与三维边界框上表面的中心相对应"这一约束关系从二维图像边界框中恢复出三维边界框;Mousavian 等人[18]提出 Deep3DBox,该网络也采用了类似的思想,利用"二维边界框在三维空间的透视投影至少应该紧贴二维边界框的一条边"这一约束关系进行位姿和三维边界框尺寸估计,实现三维目标检测;Li 等人[19]提出 Stereo R-CNN,先利用双目图像之间的语义和几何约束关系预测目标底部的四个语义关键点,再通过二维和三维空间之间的几何投影关系估计出三维边界框.

总的来说,这类方法恢复出的信息相较真实值存在一定误差,因此图像三维目标检测方法的精度通常较低.

2.2 点云三维目标检测方法

激光雷达输出的点云数据可以提供环境中物体的高精度三维信息,因此近年来众多研究者均选择使用点云进行三维目标检测.

2017 年,Qi 等人[20]提出 PointNet 和 PointNet++直接处理原始点云数据,学习点云特征,进行点云分类和分割任务. 一些研究者受该网络启发,直接使用原始点云作为输入数据进行三维目标检测. Qi 等人[21]先在 PointNet++的基础上引入霍夫投票机制生成 RoI,再进行回归细化完成检测. Shi 等人[22]提出的 Point-RCNN 先使用 PointNet++作为骨干网络对原始点云进行语义分割,在点云前景点上生成 RoI;再进行后续处理完成检测. Yang 等人[23]提出的 STD 也使用类似的思想,提出使用球形锚框进行 RoI 生成. Li 等人[24]提出的 3DIoUNet 也使用了 PointNet++结构进行点云特征提取,该网络通过设计一项交并比预测任务和对应的交并比特征提取模块提升了三维目标检测精度. 此外,Shi 等人[25]提出的 Point-GNN 使用图神经网络(Graph Neural Networks,GNN)构建原始点云特征提取骨干网络,并基于此实现三维目标检测.

2018 年美国苹果公司提出 VoxelNet[26],该方法将点云数据转化为体素,使用三维卷积处理体素,提取点云特征,解决了不规则分布点云的高效特征提取问题,拉开了点云体素三维目标检测研究的序幕. Yan 等人在 VoxelNet 的基础上引入稀疏卷积(Sparse Convolution, SC)和子流形稀疏卷积(Submanifold Sparse Convolution,SSC)[27-28],提出的 SECOND[29]构建了一种高效的体素特征提取骨干网络,降低 VoxelNet 计算消耗的同时提升了其检测精度. 基于此,该领域后续出现了一系列优秀的三维目标检测方法. 如 Lang 等人[30]在体素化过程中生成高度为 1 的柱状体素,并基于此提出 PointPillars,使用二维卷积神经网络(Convolutional Neural Network, CNN)对柱状体素进行特征提取. 和 SECOND 相比,该网络的检测速率较快,但精度较差. He 等人[31]在 SECOND 的基础上引入中心回归和前景背景分割两项辅助任务,在不增加推理计算成本的情况下提升了其三维目标检测精度. Shi 等人[32]提出的 PartA^{2}Net 利用三维数据标签中分离的点云真值信息设计了一项目标内部点云空间分布预测任务,辅助三维目标检测. Shi 等人[33]提出的 PV-RCNN 借助一组关键点,有效利用了 PointNet++和 SECOND 所设计骨干网络的优点,大幅提升了三维目标检测精度. Deng 等人[34]提出的 Voxel-RCNN 设计了一种体素 RoI 池化方法改进了 PV-RCNN,进一步提升了其检测精度. Yin 等人[35]以 VoxelNet 和 PointPillars 为基础,将二维目标检测网络 CenterNet 扩展至三维目标检测任务中,取得了较好的效果. 总的来说,SECOND 设计的骨干网络展现出强大的生命力,基于该骨干网络的三维目标检测方法近几年快速提升该领域的最优检测精度.

2.3 多模态三维目标检测方法

虽然激光雷达点云数据蕴含着三维目标检测任务所需的丰富三维信息,但图像数据蕴含的颜色信息也是一种重要的目标识别信息. 因此,一些研究者同时使用图像和点云进行三维目标检测. F-PointNet[36]和 F-ConvNet[37]采用串联融合方式处理多模态数据,首先从图像中获取目标二维边界框,然后投影至三维点云空间中进行边界框细化,得到最终的检测结果.

PointPainting[38]采用前融合方式,先对点云进行语义分割,再将语义信息串联至输入点云上,丰富输入特征. 而 Chen 等人[39]提出 MV3D,在特征提取阶段采用深度融合方式融合不同模态数据多个视角

的特征.Ku 等人和 Liang 等人依次提出 AVOD[40]、ContFuse[41]和 MMF[42],分别通过全分辨率特征融合和多尺度特征融合进一步改进了 MV3D,取得了更优的检测精度.Yoo 等[9]提出的 3D-CVF 也采用深度融合方式,设计了一种交叉视图特征映射模块,通过融合体素和图像的特征获得联合相机—激光雷达特征,并基于此构建三维目标检测网络.然而由于图像和点云的数据表征完全不同,这两种模态特征难以直接进行有效的融合,因此该类方法的检测水平均低于同期的点云三维目标检测方法.

为了解决多模态特征难以有效融合的问题,CLOC[43]采用后融合方式,先借助成熟的二维和三维目标检测网络获取相应的检测结果,再对两种检测网络得到的检测结果进行融合.然而这种方法的检测水平极其依赖其采用的目标检测网络,且由于需要使用两个目标检测网络,因此其计算成本较高.Wu 等人[4]在 2022 年提出 SFD,通过深度补全任务生成图像伪点云(Pseudo Point Cloud),以此作为颜色信息输入,在 RoI 中进行多模态特征融合,避免了多模态特征的直接融合,取得了目前最优的三维目标检测精度.但如引言所述,该网络还存在特征提取粗糙和 RoI 表征能力差的问题.因此本文以 SFD 为基础开展研究,提出的 FAC 和 MGSC 可以从两方面增强伪点云特征,改进 SFD 伪点云处理支路的缺陷,从而提升其三维目标检测精度.

3　方　法

3.1　SFD 三维目标检测网络

图 1 为 SFD 的整体结构.如图 1 所示,SFD 主要包括点云特征提取、区域提议网络、伪点云特征提取、RoI 特征学习网络和检测头网络五部分,主要流程如下.

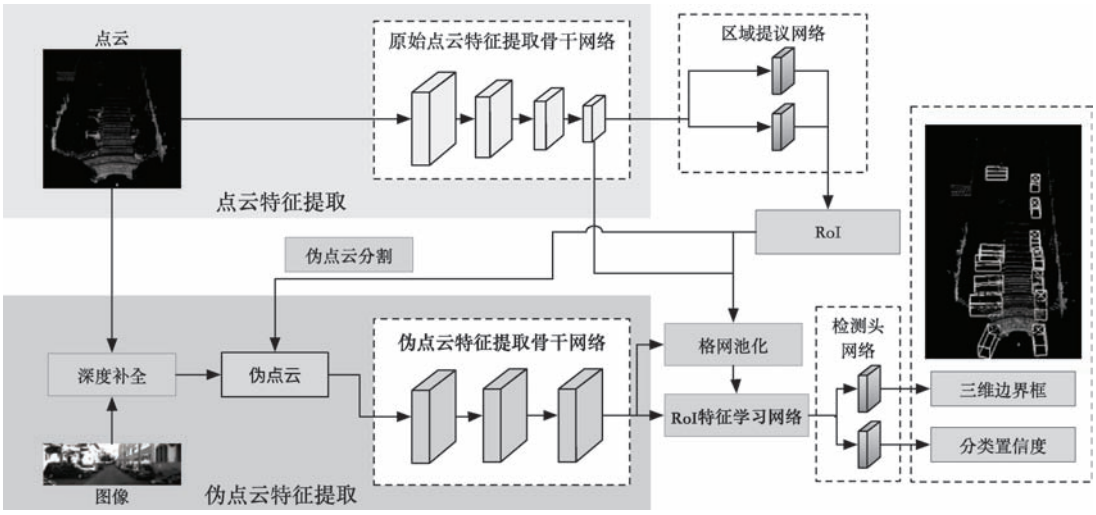

图 1　SFD 网络结构示意

首先,对点云进行特征提取,并获取 RoI;然后,对图像进行深度补全,根据 RoI 分割伪点云,得到 RoI 内的伪点云;其次,对分割后的伪点云进行特征提取,分别进行体素点云特征提取和伪点云特征提取;再次,基于 RoI 对得到的特征进行格网池化和特征学习;最后,将 RoI 格网特征送入检测头网络,得到三维目标检测结果.

如第 1 节和第 2 节所述,SFD 三维目标检测网络存在特征提取粗糙和 RoI 表征能力差的问题.针对上述问题,本文提出 FAC,并基于此构建伪点云特征提取骨干网络,对伪点云进行精细的特征提取,增强伪点云特征;提出 MGSC,并基于此构建 RoI 特征学习网络,增强 RoI 特征.本文基于上述两种网络改进 SFD 的伪点云特征提取和 RoI 特征学习流程,进而提升其三维目标检测精度.

3.2　细粒度注意力卷积

本节主要介绍 FAC 的具体处理流程,首先对输入数据进行定义.

伪点云分组特征:记 $\boldsymbol{P}\in \mathrm{R}^{N\mathrm{p}\times C_{\mathrm{in}}}$ 为一组点,则 $\boldsymbol{P}$ 中点可被表示为 $\{(f_{1i},f_{2i},\cdots,f_{C_{\mathrm{in}}i}),i=1,2,3,\cdots,N_{\mathrm{p}}\}$,其中 R 表示实数域,$f$ 为点的特征值,C_{in} 为特征维度(初始输入中一般包括点的三维坐标 x、y、z,对应的图像坐标 u、v 和对应的 RGB 颜色值 r、g、b 共 8 种特征),N_{p} 为点的数量.对 $\boldsymbol{P}$ 进行邻域搜索生成点云分组,进而得到点云分组特征 $\boldsymbol{F}_g\in \mathrm{R}^{N\mathrm{p}\times C_{\mathrm{in}}\times L}$,其中 L 为每组点包含的点数量,单个点 p_i 的分组特征记为 $\boldsymbol{a}_i\in \mathrm{R}^{C_{\mathrm{in}}\times L}$.

点云分组距离残差:根据点云的坐标构建点云分组距离残差,即各点与其邻域内其他点的距离差值.以点 p_i 的第 l 个邻域点 p_i^l 为例,它们之间的距离残差通常表示为 $\boldsymbol{b}_i^l=(x_i-x_i^l,\ y_i-y_i^l,\ z_i-z_i^l,\ \|p_i-p_i^l\|,\ u_i-u_i^l,\ v_i-v_i^l)$.堆叠 p_i 所在邻域其他 L 个点的距离残差,可以得到 p_i 的分组距离残差 $\boldsymbol{b}_i\in \mathrm{R}^{Cr\times L}$,其中 C_r 为距离残差的特征维度,进一步堆叠所有点的分组距离残差得到 $\boldsymbol{B}\in \mathrm{R}^{Np\times Cr\times L}$.

图 2 为 FAC 的整体结构,为了清晰简明地进行展示,图中仅画出某点 p_i 的特征提取过程.如图 2 所示,FAC 可以分为注意力权重学习和精细特征学习两部分.

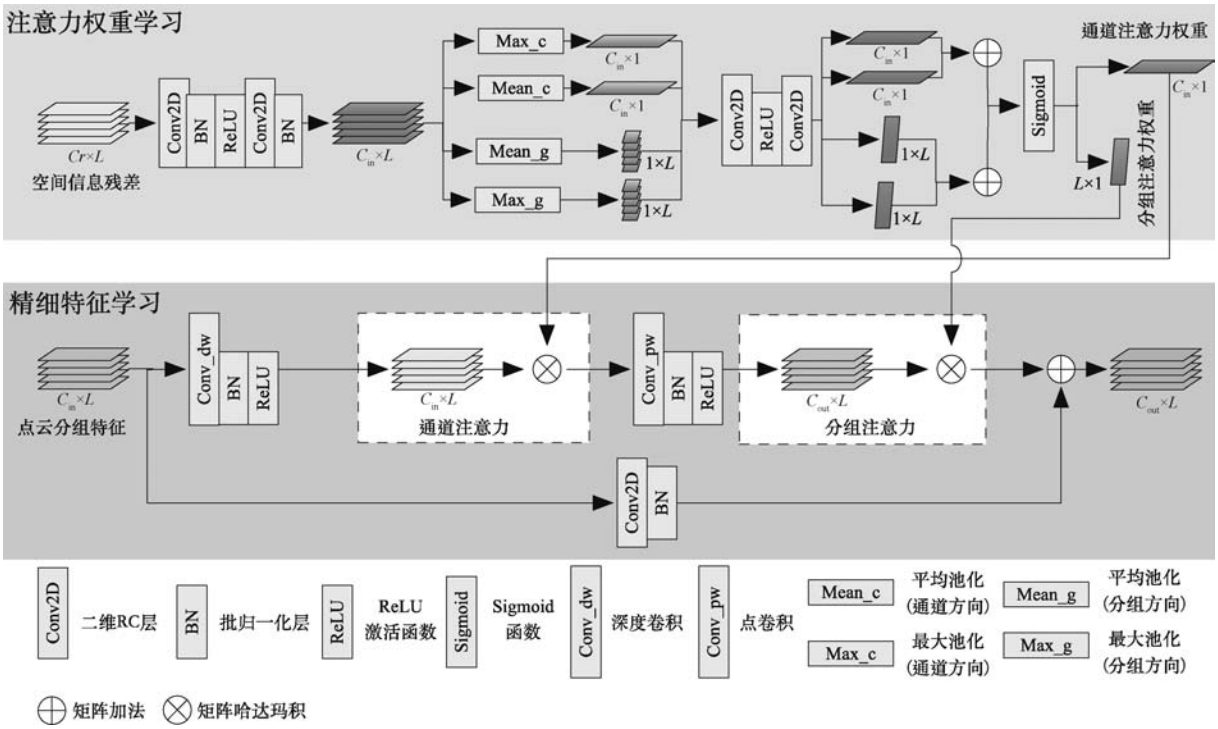

图 2 FAC 整体结构示意

3.2.1 注意力权重学习

以点云中的单个点为例,注意力权重学习具体流程如下.

首先,将分组距离残差 $\boldsymbol{b}$ 送入一个 MLP 结构,得到特征 $\boldsymbol{f}_b\in \mathrm{R}^{C_{in}\times L}$,过程如公式(1)所示.

$$\boldsymbol{f}_b=\mathrm{BN}(\mathrm{Conv}(\mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv}(\boldsymbol{b}))))) \tag{1}$$

其次,沿特征通道方向和分组方向分别进行最大池化和平均池化,得到通道最大值特征向量 $\boldsymbol{v}_{\mathrm{Max\_c}}$、通道均值特征向量 $\boldsymbol{v}_{\mathrm{Mean\_c}}$、分组最大值特征向量 $\boldsymbol{v}_{\mathrm{Max\_g}}$ 和分组均值特征向量 $\boldsymbol{v}_{\mathrm{Mean\_g}}$.

再次,将上述特征向量送入另一个 MLP 结构进行学习得到 $\boldsymbol{v}_{\mathrm{maxcl}}\in \mathrm{R}^{C_{in}\times 1}$、$\boldsymbol{v}_{\mathrm{meancl}}\in \mathrm{R}^{C_{in}\times 1}$、$\boldsymbol{v}_{\mathrm{maxgl}}\in \mathrm{R}^{1\times L}$ 和 $\boldsymbol{v}_{\mathrm{meangl}}\in \mathrm{R}^{1\times L}$.

最后,将 $\boldsymbol{v}_{\mathrm{maxcl}}$、$\boldsymbol{v}_{\mathrm{meancl}}$、$\boldsymbol{v}_{\mathrm{maxgl}}$ 和 $\boldsymbol{v}_{\mathrm{meangl}}$ 两两相加,并使用 sigmoid 函数进行处理,得到通道注意力权重 $\boldsymbol{w}_c\in \mathrm{R}^{C_{in}\times 1}$ 和分组注意力权重 $\boldsymbol{w}_g\in \mathrm{R}^{1\times L}$.过程如公式(2)所示.

$$\begin{cases}\boldsymbol{w}_c=\mathrm{sigmoid}(\boldsymbol{v}_{\mathrm{maxcl}}+\boldsymbol{v}_{\mathrm{meancl}})\\ \boldsymbol{w}_g=\mathrm{sigmoid}(\boldsymbol{v}_{\mathrm{maxgl}}+\boldsymbol{v}_{\mathrm{meangl}})\end{cases} \tag{2}$$

3.2.2 精细特征学习

以点云中的单个点为例,精细特征学习的具体流程如下.

首先,对分组特征进行深度卷积和批归一化,并使用 ReLU 函数激活输出特征,得到分组特征 $\boldsymbol{a}_c\in \mathrm{R}^{C_{in}\times L}$,过程如公式(3)所示.

$$\boldsymbol{a}_c=\mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv\_dw}(\boldsymbol{a}))) \tag{3}$$

其次,进行通道注意力计算,使用 $\boldsymbol{w}_c$ 对 $\boldsymbol{a}_c$ 的不同特征通道进行重加权,得到分组特征 $\boldsymbol{a}_{c1}\in \mathrm{R}^{C_{in}\times L}$,过程如公式(4)所示.

$$\boldsymbol{a}_{c1}=\boldsymbol{a}_c\otimes\begin{bmatrix}\boldsymbol{w}_c\\ \boldsymbol{w}_c\\ \cdots\\ \boldsymbol{w}_c\end{bmatrix} \tag{4}$$

再次,对 $\boldsymbol{a}_{c1}$ 进行点卷积和批归一化,利用点卷积将特征维度由 C_{in} 扩展为 C_{out},将低维浅层特征

抽象为高维深层特征,并使用 ReLU 函数激活输出特征,得到分组特征 $\boldsymbol{a}_g \in \mathrm{R}^{C_{\mathrm{out}} \times L}$,过程如公式(5)所示.

$$\boldsymbol{a}_g = \mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv\_pw}(\boldsymbol{a}_{c1}))) \tag{5}$$

从次,进行分组注意力计算,使用 $\boldsymbol{w}_g$ 对 $\boldsymbol{a}_g$ 的不同点进行重加权,得到特征 $\boldsymbol{a}_{g1} \in \mathrm{R}^{C_{\mathrm{out}} \times L}$,过程如公式(6)所示.

$$\boldsymbol{a}_{g1} = \boldsymbol{a}_g \otimes \begin{bmatrix} \boldsymbol{w}_g & \boldsymbol{w}_g & \cdots & \boldsymbol{w}_g \end{bmatrix} \tag{6}$$

最后,通过一个跳跃连接结构将浅层特征传至深层特征上,获得该阶段的输出 $\boldsymbol{a}_{\mathrm{fac}} \in \mathrm{R}^{C_{\mathrm{out}} \times L}$,过程如公式(7)所示.

$$\boldsymbol{a}_{\mathrm{fac}} = \boldsymbol{a}_{g1} + \mathrm{BN}(\mathrm{Conv}(\boldsymbol{a})) \tag{7}$$

3.2.3　差异性分析

3.2.1 和 3.2.2 节基于深度可分离卷积和两种注意力机制提出 FAC,描述了其具体实现流程.本节针对 FAC 和深度可分离卷积以及常用注意力机制的差异进行分析和讨论.

(1)和深度可分离卷积的使用目的差异

从使用目的上来看,深度可分离卷积最初被用来降低卷积计算量,而 FAC 中深度可分离卷积的目的在于引入两种注意力机制以进行精细的特征调整.深度可分离卷积和常规卷积的计算量之比如公式(8)所示,其中 C_{kn} 和 C_k 分别为输出特征通道数和卷积核尺寸.

$$\frac{\mathrm{T}_{\mathrm{dsc}}}{\mathrm{T}_{\mathrm{rc}}} = \frac{1}{C_{\mathrm{kn}}} + \frac{1}{C_k^2} \tag{8}$$

如公式(8)所示,当 $C_k > 1$ 时,深度可分离卷积的计算量小于常规卷积.但由于点云特征提取时通常使用 C_k 为 1 的卷积,在此情况下,深度可分离卷积解耦过程相当于在原卷积基础上增加了一组深度卷积,增加了计算量.因此,FAC 并不能达到降低计算量的效果.但由于在本文设计的 FAC 中 C_{kn} 通常较大,因此这种做法并不会在常规卷积基础上增加太多计算量.

(2)注意力权重的来源差异

现有注意力机制如通道注意力和空间注意力的注意力权重通常由输入特征(如图像特征或点云特征)直接学习得到.而如引言所述,点云分组距离残差更适合用于点云注意力权重学习,得到点云分组邻域内不同点之间的关系.因此,本文提出的 FAC 选择从点云分组距离残差中学习点云特征的注意力权重,从而在应用深度可分离卷积时引导点云特征调整,得到更有判别力的点云特征.

(3)注意力机制的差异

通道注意力和空间注意力是目前两种常用的注意力机制[7],但由于点云数据分布不规则的特点,目前常用的空间注意力机制难以直接应用于点云数据.因此 FAC 在点云分组内引入注意力机制,通过学习到的注意力权重重新衡量每个点云分组内点的重要性,以达到与空间注意力类似的效果.

3.3　多尺度分组稀疏卷积

3.3.1　多尺度分组稀疏卷积流程

本节介绍 MGSC 的处理流程.图 3 为 MGSC 的整体结构.

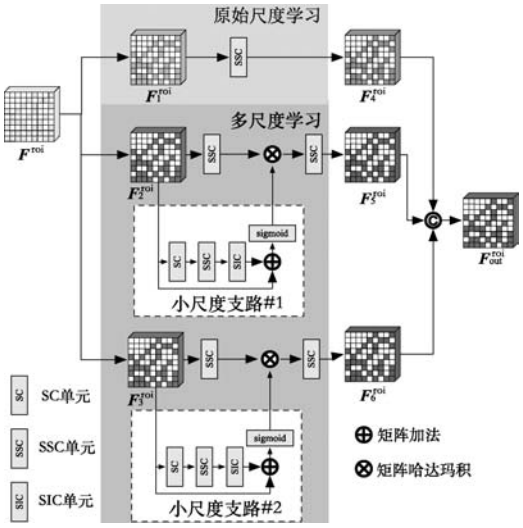

图 3　MGSC 整体结构示意

如图 3 所示,MGSC 先是借鉴分组卷积的思想,对伪点云 RoI 格网特征进行分组,得到 $\boldsymbol{F}_1^{\mathrm{roi}}$、$\boldsymbol{F}_2^{\mathrm{roi}}$ 和 $\boldsymbol{F}_3^{\mathrm{roi}}$;再分别进行原始尺度和多尺度学习.MGSC 可以从两方面改善 RoI 格网特征的学习过程:一是通过小尺度支路的特征学习,指导原始尺度特征空间中的特征学习过程,自适应地调整原始尺度空间特征;二是在小尺度支路中进行特征学习时引入了不同尺度的特征信息,使输出特征同时具有原始尺度空间特征信息和小尺度特征空间信息,增强了 RoI 格网特征的多尺度表征能力.

对于原始尺度信息,MGSC 使用 SSC 单元对 $\boldsymbol{F}_1^{\mathrm{roi}}$ 进行处理得到 $\boldsymbol{F}_4^{\mathrm{roi}}$,从而保留原始尺度信息,本文中一个卷积单元包括 1 个相应卷积层、1 个 BN 层和 1 个 ReLU 激活函数层.对于多尺度信息,MGSC 基于自校准卷积设计了两个小尺度学习支路获取不同尺度空间下的 RoI 格网特征,多尺度学习的具体流程如下.

首先,MGSC 将 $\boldsymbol{F}_2^{\mathrm{roi}}$ 送入小尺度支路 #1,先使用一个 SC 单元将 $\boldsymbol{F}_2^{\mathrm{roi}}$ 降采样至小尺度空间中,再

使用一个 SSC 单元在下采样的小尺度空间中对特征进行学习.

其次,使用一个稀疏反卷积(Sparse Inverse Convolution,SIC)单元将小尺度空间下的特征映射至原始尺度空间,得到伪点云的自适应 RoI 格网特征 $\boldsymbol{F}_{\text{adp}}^{\text{roi}}$. 上述过程如公式(9)所示.

$$\boldsymbol{F}_{\text{adp}}^{\text{roi}}=\boldsymbol{F}_{2}^{\text{roi}}\oplus(\text{SIC}(\text{SSC}(\text{SC}(\boldsymbol{F}_{2}^{\text{roi}})))) \tag{9}$$

再次,使用 $\boldsymbol{F}_{\text{adp}}^{\text{roi}}$ 自适应地调整原始尺度空间中的特征,得到 $\boldsymbol{F}_{5}^{\text{roi}}$. 由于小尺度空间下的特征具有较大的感受野,因此该特征映射回原始尺度空间后可以有效指导原始尺度的特征变化,过程如公式(10)所示.

$$\boldsymbol{F}_{5}^{\text{roi}}=\text{SSC}(\text{SSC}(\boldsymbol{F}_{2}^{\text{roi}})\otimes\text{sigmoid}(\boldsymbol{F}_{\text{adp}}^{\text{roi}})) \tag{10}$$

从次,将 $\boldsymbol{F}_{3}^{\text{roi}}$ 送入小尺度支路 #2 进行处理得到 $\boldsymbol{F}_{6}^{\text{roi}}$.

最后,串联 $\boldsymbol{F}_{4}^{\text{roi}}$、$\boldsymbol{F}_{5}^{\text{roi}}$ 和 $\boldsymbol{F}_{6}^{\text{roi}}$,得到 MGSC 的输出特征 $\boldsymbol{F}_{\text{out}}^{\text{roi}}$.

3.3.2 差异性分析

3.3.1 节基于分组卷积提出 MGSC,描述了其具体实现流程. 本节针对 MGSC 和常用多尺度学习方法进行差异化分析和讨论.

现有多尺度特征提取方法通常设置多个不同尺寸的卷积核分别对输入特征进行处理,最后将得到的各尺度特征进行合并得到多尺度特征. 如文献[44]中所设计的多尺度特征提取模块,采用了 1、2、3、4、5 共 5 种卷积核尺寸的卷积单元处理输入特征. 这种做法相当于对输入特征分别进行多次特征提取,计算量较大. 而本文所提出 MGSC 则先将输入特征分为三组,降低输入特征维度以降低计算量;再借鉴自校准卷积,以小尺度空间下的特征信息自适应地调整原始尺度空间特征;最后合并三种特征得到多尺度特征.

3.4 基于伪点云特征增强的三维目标检测网络

3.4.1 网络细节

(1)体素点云特征提取

对于体素点云特征,SFD++采用 SECOND 设计的骨干网络作为原始点云特征提取骨干网络,本文使用的网络参数如表 1 所示,其中一个完整的括号代表一个卷积单元,括号中的数字从左至右依次代表输入特征维度、输出特征维度、卷积核尺寸和卷积步长.

(2)区域提议网络

区域提议网络主要用于生成检测任务所需的 RoI. 本文采用目前通用的区域提议网络结构[32,34],不再赘述其具体结构和流程.

表 1 三维稀疏卷积网络参数

	稀疏卷积	子流形稀疏卷积
卷积块 #1	—	(4,16,3,1),(16,16,3,1)
卷积块 #2	(16,32,3,2)	(32,32,3,1),(32,32,3,1)
卷积块 #3	(32,64,3,2)	(64,64,3,1),(64,64,3,1)
卷积块 #4	(64,64,3,2)	(64,64,3,1),(64,64,3,1)

(3)伪点云特征提取

SFD++先采用 TWISE[45] 进行深度补全,获得稠密的深度图,从而获得对应的伪点云,再将伪点云送入伪点云特征提取网络进行细粒度的伪点云特征提取,获得具判别力的伪点云特征. 基于 FAC,本节构建伪点云特征提取网络 FACNet 如图 4 所示.

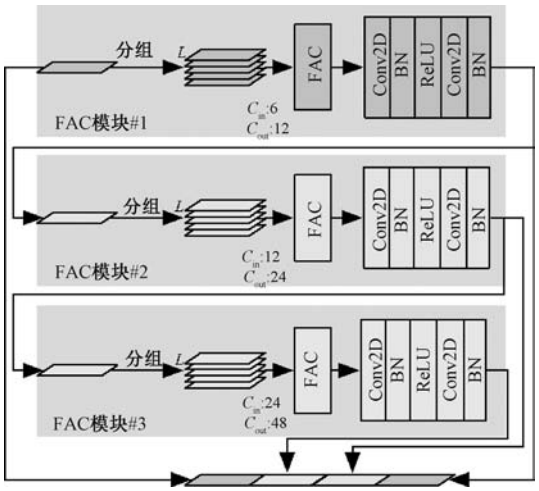

图 4 FACNet 结构示意

如图 4 所示,FACNet 主要由三个 FAC 模块组成,每个 FAC 模块中包含一个 FAC 和一个 MLP 结构. 该 MLP 结构包含两个卷积单元,输入特征维度和输出特征维度均等于图中 FAC 模块的 C_{out}. 在将分组特征聚合至一点时,FAC 模块并未使用目前常用的最大池化[34]或展平再卷积方法[4],而是先使用一个卷积核尺寸为 L 的卷积单元将分组特征聚合到一点,再使用一个卷积核尺寸为 1 的卷积单元进一步学习点云特征. 通过此方法聚合分组特征,将伪点云分组中每个点的特征值都引入该过程,保留了信息的全面性.

以点云中的单个点为例,FACNet 的具体流程如下.

首先,将点送入 FAC 模块 #1,以输入点为中心进行邻域搜索,得到分组特征 $\boldsymbol{a}$ 和分组距离残差 $\boldsymbol{b}$;

其次,使用 FAC 处理 $\boldsymbol{a}$ 和 $\boldsymbol{b}$,并送入图 4 所示 MLP 结构得到该点的特征 $\boldsymbol{a}_{\text{fac1}}\in\text{R}^{12\times1}$.

再次,将输出依次送入 FAC 模块 #2 和 FAC 模块 #3 进行处理得到 $\boldsymbol{a}_{\text{fac2}} \in \mathrm{R}^{24\times1}$ 和 $\boldsymbol{a}_{\text{fac3}} \in \mathrm{R}^{48\times1}$,这里不再赘述.

最后,将初始特征和三个 FAC 模块的输出特征串联起来,得到多层级的伪点云特征 $\boldsymbol{a}_{\text{facnet}} \in \mathrm{R}^{90\times1}$.

(4)RoI 特征学习网络

基于 MGSC,本节构建 RoI 特征学习网络结构如图 5 所示.

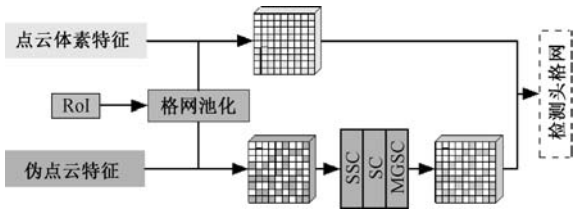

图 5　RoI 特征学习网络结构示意

对于点云体素特征,SFD++采用 Voxel-RCNN 中使用的体素 RoI 池化(Voxel RoI Pooling)方法进行格网池化;对于伪点云特征,SFD++先使用 PartA^{2}Net 中的 RoI 感知池化(RoI-aware Pooling)方法进行格网池化,再通过 SSC、SC 和 MGSC 三个卷积单元进行 RoI 特征学习,各卷积单元的相关参数如表 2 所示. MGSC 中小尺度支路的相关参数如表 3 所示.

表 2　RoI 特征学习网络的参数

卷积单元类型	卷积核尺寸	步长	输入特征维度	输出特征维度
SSC	3	1	90	128
SC	3	2	128	128
MGSC	3	1	128	128

表 3　MGSC 中小尺度支路的参数

	小尺度支路 #1			小尺度支路 #2		
	SC	SSC	SIC	SC	SSC	SIC
输入特征维度	32	32	32	32	32	32
输出特征维度	32	32	32	32	32	32
卷积核尺寸	3	3	3	3	3	3
步长	2	1	—	(2,1,1)	1	—

在进行 RoI 池化时,SFD++并未使用 Voxel-RCNN 中使用的 6×6×6 格网和 PartA^{2}Net 中使用的 12×12×12 格网,而是根据检测目标的三维尺寸划分格网池化尺寸,以期在不影响检测精度的前提下控制计算成本.

(5)检测头网络

SFD++采用与 SFD 相同的检测头网络,这里不再赘述其流程和细节.

3.4.2　损失函数

SFD++的损失函数主要由区域提议网络损失 L_{rpn}、检测头网络的两个辅助任务损失 L_{aux1}、L_{aux2} 和检测头网络的预测损失 L_{rcnn} 三部分组成,其总损失函数可以表示为公式(11).

$$L_{\text{total}} = L_{\text{rpn}} + 0.5L_{\text{aux1}} + 0.5L_{\text{aux2}} + L_{\text{rcnn}} \tag{11}$$

L_{rpn} 由平滑 L1 损失和焦点损失组成,L_{aux1}、L_{aux2} 和 L_{rcnn} 由平滑 L1 损失和二元交叉熵损失组成. 平滑 L1 损失函数如公式(12)所示.

$$\text{Smooth}-\text{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x|<1 \\ |x|-0.5 & \text{otherwise} \end{cases} \tag{12}$$

式中 $x = \Delta\text{res} - \Delta\text{res}^*$,$\Delta\text{res}$ 为预测边界框的尺寸偏移,Δres^* 为预测边界框尺寸偏移的优化目标.

焦点损失函数如公式(13)所示,其中 $\alpha=0.25$,$\gamma=2$,p_{seg} 为前景分数,$|b|$ 代表目标的前景背景标签:1 代表前景,0 代表背景.

$$\text{Focal}(p_{\text{seg}}) = \begin{cases} -\alpha(1-p_{\text{seg}})^{\gamma}\log(p_{\text{seg}}) & \text{if } |b|=1 \\ -(1-\alpha)p_{\text{seg}}{}^{\gamma}\log(1-p_{\text{seg}}) & \text{if } |b|=0 \end{cases} \tag{13}$$

二元交叉熵损失函数如公式(14)所示,式中 p_u 为预测得到的置信度;y_u 为置信度优化目标.

$$\text{BCE}(p_u, y_u) = -y_u\log(p_u) + (1-y_u)\log(1-p_u) \tag{14}$$

4　实验及分析

4.1　实验设置

4.1.1　数据集

为了和其他方法进行公平的对比,本文采用自动驾驶领域权威的 KITTI 数据集[46]进行实验. 对于点云体素化,本文在三维空间的深度,宽度和高度维度上采用的体素化步长为(0.05 m, 0.05 m, 0.1 m),以体素内点云的特征均值作为体素的初始特征. KITTI 数据集包含 7481 个训练样本,本文遵从该领域研究人员的通用做法将其分为 3712 和 3769 个样本两部分,分别作为训练集和验证集.

4.1.2　评估标准

KITTI 数据集将场景中的待测目标分为简单、中等和困难三种难度,因此本文实验遵循该难度分类进行定量评估. 由于汽车类目标是自动驾驶场景中种类和数量最多的目标,且检测结果较稳定,能较好地体现网络性能. 因此,本文主要以汽车作为实验评估目标.

本文遵从 KITTI 数据集官方的要求,主要使用

40 个召回位置的平均精度(Average Precision at 40 Recall Positions,AP@R40)对检测结果进行评估.同时由于该领域大部分研究工作中都给出了 11 个召回位置的平均精度(Average Precision at 11 Recall Positions,AP@R11),因此本文也以 AP@R11 为评估指标对网络进行了评估和对比.

4.1.3 训练参数

在训练过程中,SFD++使用自适应矩估计(Adaptive Moment Estimation,Adam)优化器[47]和余弦退火策略调整学习率,批大小、学习率和训练周期分别为 1、0.00125 和 50.本文使用一台配备 RTX 2080ti GPU 和志强银牌 4210 CPU 的计算机完成实验.

4.2 定量结果和分析

4.2.1 和现阶段先进方法的对比分析

为了和其他方法进行公平的对比,本文首先在同一实验环境下复现了近几年较先进的三维目标检测方法.表 4 为 SFD++和其他先进方法在 KITTI 验证集上的汽车三维目标检测 AP@R40,其中 L 和 I+L 分别代表点云三维目标检测方法和多模态三维目标检测方法.

如表 4 所示,由于引入了额外的图像数据,多模态三维目标检测方法总体上取得了比点云三维目标检测方法更高的精度.但本文提出的 SFD++在点云和多模态三维目标检测方法中均取得了最优的三维检测精度,在简单、中等和困难难度上分别比次优的 SFD 高出 0.15%、0.84%和 0.58%,展现了优秀的汽车三维目标检测性能.

表 4 SFD++和其他先进方法的汽车三维目标检测 AP@R40

方法	提出时间	模态	汽车(%)			
			简单	中等	困难	平均
SECOND[29]	2018	L	90.55	81.61	78.61	83.59
PointPillars[30]	2019	L	87.75	78.39	75.18	80.44
PointRCNN[22]	2019	L	91.82	80.57	78.08	83.49
PartA^{2}Net[32]	2020	L	91.56	84.27	82.05	85.96
PV-RCNN++[48]	2021	L	91.79	84.62	82.49	86.30
VoTr[49]	2021	L	91.95	84.63	82.50	86.36
E^2-PV-RCNN[50]	2022	L	92.12	84.83	82.63	86.53
VoxSeT[51]	2022	L	89.55	80.64	78.14	82.78
SRIF-RCNN[10]	2022	I+L	92.23	85.43	83.17	86.94
SFD[4]	2022	I+L	95.59	87.96	85.46	89.67
SFD++	—	I+L	**95.74**	**88.80**	**86.04**	**90.19**
提升	—	—	+0.15	+0.84	+0.58	+0.52

表 5 为 SFD++和其他先进方法在 KITTI 验证集上的行人和自行车手三维目标检测 AP@R40,其中 L 和 I+L 分别代表点云三维目标检测方法和多模态三维目标检测方法.由于行人和自行车手都属于小目标,需要综合看待这类目标的检测精度.因此表 5 中最后一列的平均精度计算的是行人和自行车手检测精度的平均精度.如表 5 所示,本文提出的 SFD++虽然在自行车手各项检测上并未取得最高的 AP@R40,但其平均 AP@R40 达到 70.67%,优于表 5 中其他方法,比次优的 SFD 高出 1.83%,展现了良好的小目标检测性能.

表 5 SFD++和其他先进方法的行人和自行车手三维目标检测 AP@R40

方法	模态	行人(%)			自行车手(%)			—
		简单	中等	困难	简单	中等	困难	平均
SECOND[29]	L	55.94	51.14	46.16	82.96	66.74	62.78	60.95
PointPillars[30]	L	57.30	51.41	46.87	81.58	62.94	58.98	59.85
PointRCNN[22]	L	61.69	54.97	48.48	91.64	**73.26**	68.78	66.47
PartA^{2}Net[32]	L	66.89	59.68	54.62	90.34	70.14	66.93	68.10
PV-RCNN++[48]	L	65.05	58.23	53.31	91.11	70.00	65.50	67.20
VoTr[49]	L	54.65	49.44	45.86	88.50	70.09	65.61	62.36
E^2-PV-RCNN[50]	L	66.47	58.58	53.30	**92.17**	72.71	**68.90**	68.69
VoxSeT[51]	L	58.56	53.48	48.58	87.77	69.48	65.20	63.85
SRIF-RCNN[10]	I+L	65.30	57.82	52.25	90.20	69.29	64.51	66.56
SFD[4]	I+L	72.37	64.91	57.99	90.00	66.12	61.64	68.84
SFD++	I+L	**74.99**	**65.48**	**58.53**	89.90	69.84	65.29	**70.67**
提升	—	+2.62	+0.57	+0.54	−0.10	+3.72	+3.65	+1.83

除此之外,本文也遵循该研究领域学者的传统做法,以 AP@R11 为指标对近几年提出的方法进行定量评估,结果如表 6 所示,表中所有数据均引自相应参考文献.如表 6 所示,SA-SSD 的 AP@R11 在简单难度上取得了最高的精度,其 AP@R11 比 SFD++高出 0.20%,这主要是由于简单目标可见程度高,包含的点较多,SA-SSD 提出的两种能够感知目标几何结构的辅助任务在这类目标中能够发挥较大优势.但 SFD++在重要的中等和难以检测的困难难度上分别比 SA-SSD 高出 7.38%和 7.26%,远远优于简单难度检测上存在的缺陷,平均高出 4.81%.综合表 4 和表 6 的实验结果可以看出,无论使用 AP@R40 还是 AP@R11 进行评估,本文提出的 SFD++均取得了较优的三维目标检测精度.

表 7 为 SFD++和其他先进方法在 KITTI 测试集上的三维检测 AP@R40,表中所有数据均引自相应参考文献.如表 7 所示,SFD++的 AP@R40 明显高于除 SFD 以外的三维目标检测网络.与 SFD 相比,SFD++在三个难度下的检测精度略低.

表 6　SFD++和其他先进方法的汽车三维目标检测 AP@R11

方法	提出时间	模态	汽车(%)			
			简单	中等	困难	平均
SECOND[29]	2018	L	87.43	76.48	69.10	77.67
PointRCNN[22]	2019	L	88.88	78.63	77.38	81.63
STD[23]	2019	L	89.70	79.80	79.30	82.93
MVLSN[52]	2020	L	86.31	77.32	73.21	78.95
3DSSD[53]	2020	L	89.71	79.45	78.67	82.61
SA-SSD[31]	2020	L	**90.15**	79.91	78.78	82.95
PartA^{2}Net[32]	2020	L	89.48	79.47	78.54	82.50
PV-RCNN[33]	2020	L	—	83.90	—	—
Point-GNN[25]	2020	L	87.89	78.34	77.38	81.20
DVFENet[54]	2021	L	89.81	79.52	78.32	82.55
CIA-SSD[55]	2021	L	90.04	79.81	78.80	82.88
VoTr[49]	2021	L	89.04	84.04	78.68	83.92
VoxSeT[51]	2022	L	88.45	78.48	77.07	81.33
MV3D[39]	2017	I+L	71.29	62.68	56.56	63.51
AVOD[40]	2018	I+L	84.41	74.44	68.65	75.83
ContFuse[56]	2018	I+L	86.32	73.25	67.81	75.79
F-PointNet[36]	2018	I+L	83.76	70.92	63.65	72.78
F-ConvNet[37]	2019	I+L	89.02	78.80	77.09	81.64
PI-RCNN[57]	2020	I+L	88.27	78.53	77.75	81.52
3D-CVF[9]	2020	I+L	89.67	79.88	78.47	82.67
Azimuth-aware Fusion[58]	2020	I+L	86.77	76.84	75.92	79.84
Multi-Layer Fusion[59]	2021	I+L	83.77	73.84	67.37	74.99
SFD[4]	2022	I+L	89.74	87.12	85.20	87.35
SFD++	—	I+L	89.95	**87.29**	**86.04**	**87.76**
提升	—	—	−0.20	+0.17	+0.84	+0.41

表 7　SFD++和其他先进方法在 KITTI 测试集上的三维检测 AP@R40

方法	提出时间	模态	汽车(%)			
			简单	中等	困难	平均
SECOND[29]	2018	L	83.34	72.55	65.82	73.90
PointRCNN[22]	2019	L	86.96	75.64	70.70	77.77
STD[23]	2019	L	87.95	79.71	75.09	80.92
3DSSD[53]	2020	L	88.36	79.57	74.55	80.83
SA-SSD[31]	2020	L	88.75	79.79	74.16	80.9
PartA^{2}Net[32]	2020	L	87.81	78.49	73.51	79.94
PV-RCNN[33]	2020	L	90.25	81.43	76.82	82.83
Point-GNN[25]	2020	L	88.33	79.47	72.29	80.03
DVFENet[54]	2021	L	86.20	79.18	74.58	79.99
CIA-SSD[55]	2021	L	89.59	80.28	72.87	80.91
VoTr[49]	2021	L	90.07	82.09	**79.14**	83.77
VoxSeT[51]	2022	L	88.53	82.06	77.46	82.68
MV3D[39]	2017	I+L	74.97	63.63	54.00	64.20
AVOD[40]	2018	I+L	83.07	71.76	65.73	73.52
ContFuse[56]	2018	I+L	83.68	68.78	61.67	71.38
F-PointNet[36]	2018	I+L	82.19	69.79	60.59	70.86
F-ConvNet[37]	2019	I+L	87.36	76.39	66.69	76.81
PI-RCNN[57]	2020	I+L	84.37	74.82	70.03	76.41
3D-CVF[9]	2020	I+L	89.20	80.05	73.11	80.79
SFD[4]	2022	I+L	**91.73**	**84.76**	77.92	**84.80**
SFD++	—	I+L	90.93	82.46	77.27	83.55

综合表 4、表 5、表 6 和表 7 的实验结果可以看出:SFD++在同一实验环境下的验证集实验上取得了先进的检测精度,而在不同实验环境下的测试集实验上精度欠佳.本文认为这是由于计算资源有限,本文实验仅可使用 1 为批大小训练 SFD++,因此相比使用 8 为批大小进行训练得到的 SFD 测试集模型,本文得到的 SFD++测试集模型泛化性能欠佳.

表 8 为 SFD 和 SFD++各部分的运行时长.其中 A、B、C、D、E、F 和 G 分别代表深度估计、点云特征提取、RoI 生成、伪点云分割、伪点云特征提取、RoI 特征学习和检测头网络.本文所提出的 SFD++分别提出 FAC 和 MGSC 改进了 SFD 网络中的 E 和 F 过程,因此表中 E、F 所需时间不同.此外,由于 SFD++使用较小的格网尺寸,因此最后的检测头网络所需时间也不同.

表 8　SFD++各部分的运行时长　(单位:毫秒)

	A	B	C	D	E	F	G
SFD++	21.0	23.2	6.3	5.0	10.3	51.1	3.0
SFD[4]					5.4	52.9	3.6

如表 8 所示,和常规卷积相比,FAC 计算量较大,因此在进行伪点云特征提取时,SFD++所需时间比 SFD 长.虽然本文提出的 MGSC 结构比常规的稀疏卷积复杂,但由于使用了较小的格网池化尺寸,SFD++在 RoI 特征学习过程中所需时间比 SFD 在该过程中所需时间少.同时,在最终的检测头网络中所需时间也较 SFD 少.

表 9 为 SFD++和一些典型三维目标检测方法的运行速度对比,以每秒可处理帧数(Frame Per Second,FPS)评估.如表 9 所示,从总体上看,由于仅需要处理一种数据,点云三维目标检测方法的运行速度要快于多模态三维目标检测方法,其中 VoxSeT 的 FPS 最高,但综合表 4、表 5 和表 6 的数据来看,VoxSeT 无法为自动驾驶汽车提供高质量的三维感知结果.由表 9 还可以看出,SFD++和 SFD 的运行速度并无太大差别.原因有二:一是 FAC 在卷积解耦过程中仅引入较小的计算量;二是 MGSC 通过特征分组进行差异化学习,也控制了相应计算量.这使得 SFD++在几乎不影响运行速度的基础上,提升了 SFD 的三维目标检测性能.

表 9　SFD++和其他典型方法的运行速度(单位:帧/秒)

PV-RCNN[33]	VoxSeT[51]	F-ConvNet[37]	SFD[4]	SFD++
L	L	I+L	I+L	I+L
11.00	29.41	2.30	8.42	8.33

4.2.2 消融实验

本节进行一系列消融实验验证本文提出 FAC 和 MGSC 的有效性. 表 10 为 FAC 和 MGSC 的消融实验结果,以 AP@R40 进行评估. 第一行不使用 FAC 和 MGSC,代表原始 SFD 网络;第二行在第一行的基础上加入 FAC,即使用 FACNet 进行伪点云特征提取;第三行在第一行的基础上加入 MGSC,即使用 3.4.1 节中所述 RoI 特征学习网络进行伪点云 RoI 特征学习;第四行同时使用 FAC 和 MGSC,代表完整的 SFD++网络.

表 10 FAC 和 MGSC 的消融实验结果

FAC	MGSC	汽车(%)			
		简单	中等	困难	平均
N/A	N/A	95.59	87.96	85.46	89.67
√	N/A	95.67	88.56	85.98	90.07
N/A	√	95.50	88.46	85.62	89.86
√	√	**95.74**	**88.80**	**86.04**	**90.19**

FAC 的作用:如表 10 中第一行和第二行所示,在使用基于 FAC 的 FACNet 后,SFD 的简单、中等和困难三维汽车检测 AP@R40 分别提升了 0.08%、0.60% 和 0.52%,验证了本章所设计 FAC 和 FACNet 的有效性. 本文认为这是由于网络在使用 FACNet 后可以获得精细的伪点云特征,增强了伪点云特征的表征能力.

MGSC 的作用:由表 10 中第一行和第三行可以看出,在 RoI 特征学习网络中加入 MGSC 后,SFD 在简单三维汽车检测上的 AP@R40 下降了 0.09%,在中等和困难难度的检测上分别提高了 0.50%和 0.16%,平均提高了 0.19%,验证了本章所设计 MGSC 的有效性. 本文认为这是由于在使用 MGSC 后,伪点云 RoI 格网特征能表征多尺度的格网信息,为 RoI 细化提供了有判别力的特征,从而提升了检测出的三维边界框质量.

由表 10 中第一行和第四行可以看出,同时使用 FAC 和 MGSC 后,完整的 SFD++取得了最高的检测精度,验证了本文所提出 FAC 和 MGSC 的有效性.

4.2.3 补充实验

本节针对检测网络中所使用的伪点云 RoI 格网划分尺寸进行了补充实验.

表 11 为 RoI 特征学习网络使用不同伪点云 RoI 格网划分尺寸时的三维汽车检测 AP@R40,以 AP@R40 进行评估.

表 11 RoI 特征学习网络使用不同格网划分尺寸时的三维汽车检测 AP@R40

格网划分尺寸(长度×宽度×高度)	汽车(%)		
	简单	中等	困难
12×12×12	95.70	88.77	**86.07**
12×6×6	**95.74**	**88.80**	86.04
12×4×4	95.67	86.82	83.97

如第一行和第二行所示,当宽度和高度方向上的格网尺寸降低为 6 时,SFD++的检测性能和格网尺寸降低前差距不大,在三种难度上互有优劣. 本文认为这是由于对绝大多数汽车来说,其长、宽和高大致成 2∶1∶1 的比例,RoI 宽度和高度方向上并不需要太多的格网即可为最终的检测提供足够有效的特征信息. 另一方面,当格网尺寸由 12×12×12 降低为 12×6×6 后,每个 RoI 的格网点由 1728 个降低为 432 个,仅占原始版本的四分之一,这可以有效降低检测网络的计算成本. 但如第二行和第三行所示,当格网尺寸由 12×6×6 降低为 12×4×4 之后,SFD++的检测精度出现了不同程度的下降,这说明在宽度和高度上的格网已不能为最终检测提供有效的特征信息,同时这也符合之前研究工作中得到的结论:稠密的格网能提供更加精细的 RoI 特征[34,48]. 由此可见,12×6×6 是相对合理的伪点云 RoI 池化格网尺寸.

4.3 定性结果和分析

图 6 为 SFD++、SFD、SRIF-RCNN 以及 VoxSeT 在 KITTI 数据集 4 个场景中的三维汽车检测结果,粗方框表示错误检测结果. 如图 6 所示,场景 1 中检测目标较少,4 种方法均能给出准确的检测结果. 在较为复杂的场景 2 和场景 3 中,SFD++仍然能够给出准确的检测结果,但其他 3 种方法均出现了不同类型的检测错误. 在场景 4 中,SFD 错将路右侧一处挡板检测为汽车,而本文提出的 SFD++避免了该错检. 虽然和 SRIF-RCNN 相比,SFD++在该场景中漏检了较远处的一辆汽车,但在这 4 个场景中 SFD++仅出现一处错误,而包括 SRIF-RCNN 在内的其他方法均出现了至少两处检测错误. 由此可见,在上述场景中 SFD++优于其他 3 种三维目标检测方法.

图 7 为 SFD++在 KITTI 数据集 12 个场景中的三维汽车检测结果. 如图 7 所示,在 FAC 和 MGSC 的帮助下,本文构建的 SFD++在拥挤和宽松的现实场景中均能给出良好的三维检测结果,一定程度上佐证了本文提出 FAC、MGSC 的有效性.

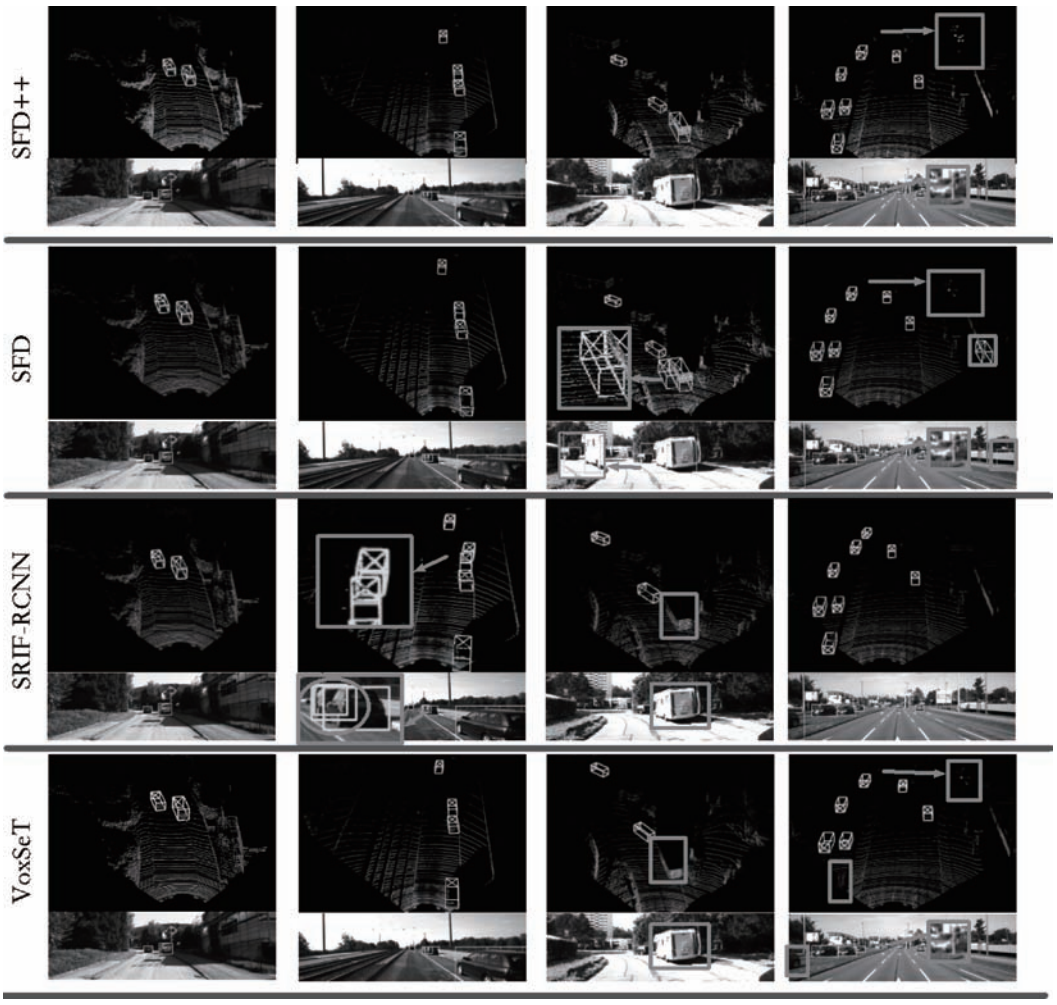

图 6　SFD++、SFD、SRIF-RCNN 和 VoxSeT 在 KITTI 数据集 4 个场景中的检测结果

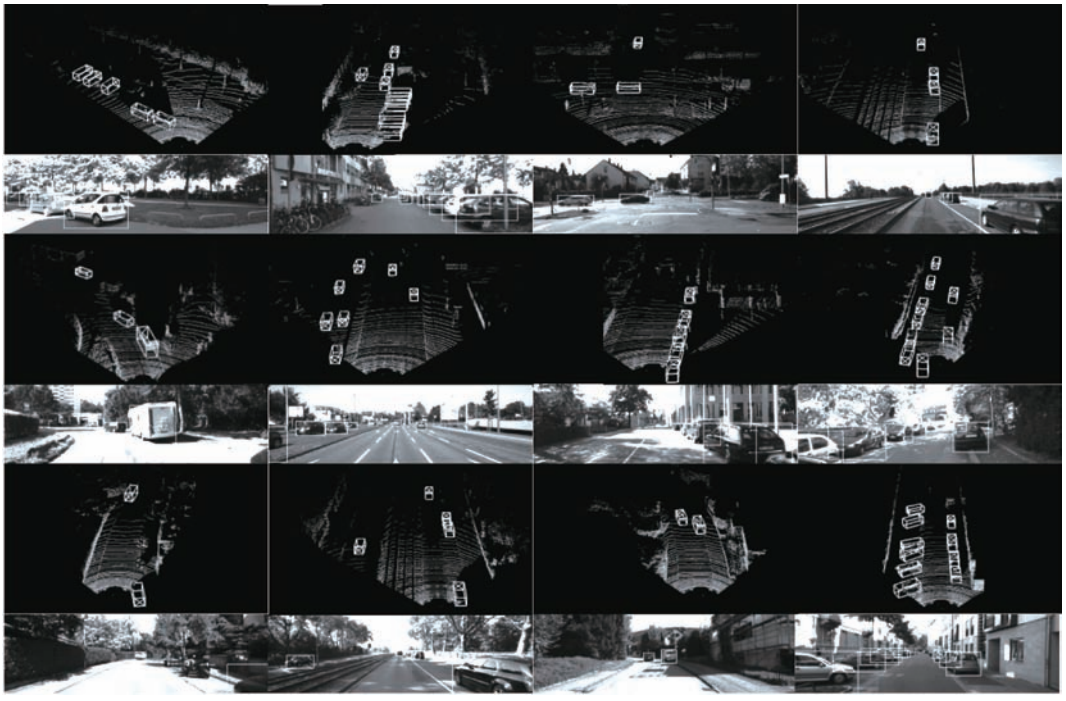

图 7　SFD++在 KITTI 数据集上的检测结果

5 总结与展望

针对自动驾驶场景下的三维目标检测任务，本文面向图像伪点云数据提出了细粒度注意力卷积FAC，该卷积依照深度可分离卷积解耦常规卷积，并在此过程中引入注意力机制，借助点之间的几何关系引导注意力权重的学习，进而实现了精细化的点云特征学习；还提出了多尺度分组稀疏卷积MGSC，将RoI格网特征分组并进行差异化学习，利用小尺度空间扩大的感受野获取原始尺度空间的权重，自适应地调整原始尺度空间下的RoI格网特征，增强其对于多尺度信息的表征能力. 基于此，本文构建了SFD++三维目标检测网络.

在权威的KITTI自动驾驶数据集上的实验结果表明，SFD++以8.33的FPS在简单、中等和困难难度汽车检测上达到95.74%、88.80%和86.04%的AP@R40，比SFD的AP@R40高出0.15%、0.84%和0.58%. 此外，本文进行的一系列消融实验验证了FAC和SGSC的有效性；进行的一系列补充实验验证了RoI特征学习网络中相关结构和参数设置的合理性.

虽然相比其他多模态三维目标检测方法，SFD++的运行速度尚可接受，但和点云三维目标检测方法相比还有一些差距. 因此，在今后的研究工作中，作者将针对网络轻量化方法开展研究，以期缩近与点云三维目标检测网络的速度差距. 此外，在后续研究工作中也将对如何在计算资源有限的情况下提升模型的泛化能力进行探索和研究.

参 考 文 献

[1] Li Peng-Fang, Liu Fang, Li Ling-Ling, et al. Meta-feature relearning with embedded label semantics and reweighting for few-shot object detection. Chinese Journal of Computers, 2022, 45(12): 2561-2575 (in Chinese)
(李鹏芳，刘芳，李玲玲，等. 嵌入标签语义的元特征再学习和重加权小样本目标检测. 计算机学报，2022，45(12)：2561-2575)

[2] Xie Xing-Xing, Cheng Gong, Yao Yan-Qing, et al. Dynamic feature fusion for object detection in remote sensing images. Chinese Journal of Computers, 2022, 45(4): 735-747 (in Chinese)
(谢星星，程塨，姚艳清，等. 动态特征融合的遥感图像目标检测. 计算机学报，2022，45(4)：735-747)

[3] Qian R, Garg D, Wang Y, et al. End-to-end pseudo-LiDAR for image-based 3d object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 5881-5890

[4] Wu X, Peng L, Yang H , et al. Sparse fuse dense: Towards high quality 3d detection with depth completion//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 5408-5417

[5] Qi C R, Su H, Mo K C, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 652-660

[6] Jie H, Li S, Gang S, et al. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(08): 2011-2023

[7] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 3-19

[8] Sifre L, Mallat S. Rigid-motion scattering for texture classification. arXiv preprint arXiv:1403.1687, 2014

[9] Yoo J H, Kim Y, Kim J, et al. 3d-cvf: Generating joint camera and LiDAR features using cross-view spatial feature fusion for 3d object detection//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 720-736

[10] Li X W, Kong D M. SRIF-RCNN: Sparsely represented inputs fusion of different sensors for 3D object detection. Applied Intelligence, 2022,53(5):5532-5553

[11] Xu B, Chen Z Z. Multi-level fusion based 3d object detection from monocular images//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 2345-2353

[12] Weng X, Kitani K. Monocular 3d object detection with pseudo-LiDAR point cloud//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Seoul, Korea, 2019

[13] Wang Y, Chao W L, Garg D, et al. Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 8445-8453

[14] You Y, Wang Y, Chao W L, et al. Pseudo-LiDAR++: Accurate depth for 3D object detection in autonomous driving. arXiv preprint arXiv:1906.06310, 2019

[15] Chen Y L, Liu S, Shen X Y, et al. Dsgn: Deep stereo geometry network for 3d object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 12536-12545

[16] Reading C, Harakeh A, Chae J, et al. Categorical depth distribution network for monocular 3d object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online, 2021: 8555-8564

[17] Li B, Ouyang W, Sheng L, et al. Gs3d: An efficient 3d object detection framework for autonomous driving//Pro-

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1019-1028

[18] Mousavian A, Anguelov D, Flynn J, et al. 3D bounding box estimation using deep learning and geometry//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 7074-7082

[19] Li P, Chen X, Shen S. Stereo R-cnn based 3D object detection for autonomous driving//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 7644-7652

[20] Qi C R, Yi L, Su H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in Neural Information Processing Systems, 2017, 30: 5099-5108

[21] Qi C R, Litany O, He K M, et al. Deep hough voting for 3D object detection in point clouds//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 9277-9286

[22] Shi S S, Wang X G, Li H S. Pointrcnn: 3D object proposal generation and detection from point cloud//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 770-779

[23] Yang Z, Sun Y, Liu S, et al. Std: Sparse-to-dense 3D object detector for point cloud//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 1951-1960

[24] Li J, Luo S J, Zhu Z Q, et al. 3d iou-net: Iou guided 3D object detector for point clouds. arXiv preprint arXiv: 2004.04962v1, 2020

[25] Shi W J, Rajkumar R. Point-gnn: Graph neural network for 3D object detection in a point cloud//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 1711-1719

[26] Zhou Y, Tuzel O. Voxelnet: End-to-end learning for point cloud based 3D object detection//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4490-4499

[27] Graham B, van der Maaten L. Submanifold sparse convolutional networks. arXiv preprint arXiv:1706.01307, 2017

[28] Graham B, Engelcke M, Van Der Maaten L. 3D semantic segmentation with submanifold sparse vonvolutional networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 9224-9232

[29] Yan Y, Mao Y X, Li B. Second: Sparsely embedded convolutional detection. Sensors, 2018, 18(10): 3337

[30] Lang A H, Vora S, Caesar H, et al. Pointpillars: Fast encoders for object detection from point clouds//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 12697-12705

[31] He C H, Zeng H, Huang J Q, et al. Structure aware single-stage 3D object detection from point cloud//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 11873-11882

[32] Shi S S, Wang Z, Shi J P, et al. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2020, 43(8): 2647-2664

[33] Shi S S, Guo C X, Jiang L, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 10529-10538

[34] Deng Jia-Jun. Object detection for video sequence and point clouds[Ph. D. Thesis]. Hefei: University of Science and Technology of China, 2021 (in Chinese)
(邓家俊. 面向视频和点云数据的目标检测方法研究[博士论文]. 合肥: 中国科学技术大学, 2021)

[35] Yin T W, Zhou X Y, Krahenbuhl P. Center-based 3D object detection and tracking//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online, 2021: 11784-11793

[36] Qi C R, Liu W, Wu C, et al. Frustum pointnets for 3D object detection from RGB-D data//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 918-927

[37] Wang Z X, Jia K. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Macau, China, 2019: 1742-1749

[38] Vora S, Lang A H, Helou B, et al. Pointpainting: Sequential fusion for 3D object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 4604-461

[39] Chen X, Ma H M, Wan J, et al. Multi-view 3D object detection network for autonomous driving//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1907-1915

[40] Ku J, Mozifian M, Lee J, et al. Joint 3D proposal generation and object detection from view aggregation//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Macau, China, 2018: 1-8

[41] Liang M, Yang B, Wang S L, et al. Deep continuous fusion for multi-sensor 3D object detection//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 641-656

[42] Liang M, Yang B, Chen Y, et al. Multi-task multi-sensor fusion for 3D object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 7345-7353

[43] Pang S, Morris D, Radha H. Clocs: Camera-LiDAR object candidates fusion for 3D object detection//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and

Systems. Las Vegas, USA, 2020: 10386-10393

[44] Su Y, Jiang L and Cao J. Point cloud semantic segmentation using multi scale sparse convolution neural network. arXiv preprint arXiv:2205.01550, 2022

[45] Imran S, Liu X, Morris D. Depth completion with twin surface extrapolation at occlusion boundaries//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 2583-2592

[46] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The kitti vision benchmark suite//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA, 2012: 3354-3361

[47] Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014

[48] Shi S S, Jiang L, Deng J J, et al. PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection. arXiv preprint arXiv: 2102.00463v1, 2021

[49] Mao J G, Xue Y J, Niu M Z, et al. Voxel transformer for 3D object detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 3164-3173

[50] Li X W, Zhang Y C, Kong D M. E-2-PV-RCNN: Improving 3D object detection via enhancing keypoint features. Multimedia Tools and Applications, 2022, 81(25): 35843-35874

[51] He C H, Li R H, Li S, et al. Voxel set transformer: A set-to-set approach to 3D object detection from point clouds//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 8417-8427

[52] Yang Y G, Chen F, Wu F, et al. Multi-view semantic learning network for point cloud based 3D object detection. Neurocomputing, 2020, 397: 477-485

[53] Yang Z T, Sun Y N, Liu S, et al. 3DSSD: Point-based 3D single stage object detector//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 11040-11048

[54] He Y Q, Xia G H, Luo Y K, et al. Dvfenet: Dual-branch voxel feature extraction network for 3d object detection. Neurocomputing, 2021, 459: 201-211

[55] Zheng W, Tang W L, Chen S J, et al. CIA-SSD: Confident iou-aware single-stage object detector from point cloud//Proceedings of the AAAI Conference on Artificial Intelligence. Online, 2021, 35(04): 3555-3562

[56] Liang M, Yang B, Wang S, et al. Deep continuous fusion for multi-sensor 3D object detection//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 641-656

[57] Xie L, Xiang C, Yu Z X, et al. PI-RCNN: An efficient multi-sensor 3d object detector with point-based attentive cont-conv fusion module//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(07): 12460-12467

[58] Tian Y L, Wang K F, Wang Y, et al. Adaptive and azimuth-aware fusion network of multimodal local features for 3D object detection. Neurocomputing, 2020, 411: 32-44

[59] Wu Y, Jiang X Y, Fang Z J, et al. Multi-modal 3D object detection by 2d-guided precision anchor proposal and multi-layer fusion. Applied Soft Computing, 2021,108: 107405

KONG De-Ming, Ph. D., professor. His research interests include computer vision, spectral information analysis and LiDAR data processing.

LI Xiao-Wei, Ph. D., His research interests include computer vision, deep learning and 3D scene perception.

YANG Qing-Xin, M. S. candidate. His research interests include computer vision, deep learning.

Background

With the development of the automotive industry and the improvement of computing power, autonomous driving technology has been widely studied in recent years. Environment perception is a core issue of autonomous vehicles. It needs to perceive the surrounding environment with high accuracy by analyzing the data obtained by camera, LiDAR, and other sensors. Real driving scenarios have extremely strict requirements for safety, and any seemingly insignificant mistake can lead to traffic accidents, causing casualties and economic losses. Based on comprehensive 3D information, the autonomous driving system can make accurate driving decisions and control operations in complex real scenes to ensure the safety and reliability of autopilot. Therefore, this article researches 3D object detection in autonomous driving scenarios.

In 2022, Wu et al. proposed SFD, which converts LiDAR point clouds and image data into voxels and image pseudo point clouds, then effectively integrates point cloud voxels and image pseudo point cloud features based on Region of Interests, achieving advanced 3D object detection

precision. It achieved average precision at 40 recall positions (AP@R40) of 95.59%, 87.96%, and 85.46% for easy, moderate, and hard car detection in KITTI autonomous driving dataset. However, this network has problems with rough feature extraction of image pseudo point clouds and poor feature representation ability of image pseudo point cloud RoI grid features. To this end, this paper proposes Fine-grained Attention Convolution (FAC) for point cloud representation data, which combines attention mechanism and depth-wise separable convolution for fine feature extraction to enhance the representation ability of image pseudo point cloud features. In addition, Multi-scale Group Sparse Convolution (MGSC) is also proposed for RoI grid feature learning, which helps RoI to obtain multi-scale features, and enhances the representation ability of image pseudo point cloud RoI grid features. Based on FAC and MGSC, this paper constructs SFD++ 3D object detection network. Experiments on KITTI datasets show that SFD++ leads SFD by 0.15%, 0.84%, and 0.58% in 3D car detection.

This paper is supported by the National Natural Science Foundation of China (No. 62173289), the Aviation Science Foundation (No. 20200016099002) and the Youth Science Fund of National Natural Science Foundation of China (No. 61501394).