DOI: 10.19650/j.cnki.cjsi.J2413542

基于体素的全稀疏三维目标检测器*

周伟典[1,2],洪 濡[1,2],盖绍彦[1,2],达飞鹏[1,2]
(1. 东南大学自动化学院 南京 210096; 2. 东南大学复杂工程系统测量与控制教育部重点实验室 南京 210096)

摘 要:针对目前基于体素的三维目标检测方法由于过于依赖密集二维骨干网络而导致在大范围点云感知上实时性不佳问题,提出了一种基于体素的全稀疏三维目标检测器 VoxelFSD,有效提升在大范围点云上检测的实时性表现。该模型由 3 个关键部分组成:首先,并行卷积分支模块(PCB),扩大模型的感受野,充分提取物体特征,并且有效处理物体中心特征丢失对结果的影响;其次,稀疏候选框生成(SRPN)检测头,以稀疏的方式预测物体定位框,在点云模态下,相比密集预测方式能够减少冗余计算,从而提升模型在大范围点云预测中的计算效率;最后,注意力融合模块候选区域检测头(AFM-ROI),在二阶段检测中,利用交叉注意力机制有效融合提取的三维骨干特征和压缩后的鸟瞰图特征,进一步精炼物体特征,得到更好的检测效果。在现有基于体素检测框架上舍弃密集 2D 骨干,并引入 PCB 模块和 SRPN 检测头,提出了全稀疏结构的单阶段轻量级检测器 VoxelFSD-S。VoxelFSD-S 在速度和精度上相比现有基于体素的轻量化模型达到了更好的平衡,并且能够在大范围点云场景中满足实时性要求。在 VoxelFSD-S 基础上,进一步引入 AFM-ROI 提出了两阶段检测器 VoxelFSD-T。VoxelFSD-T 牺牲部分推理速度但能够显著提升模型精度。VoxelFSD-S 和 VoxelFSD-T 在 KITTI 数据集测试集上精度分别达到 77.67% 和 81.50%。

关键词: 三维目标检测;体素化;轻量化;交叉注意力;全稀疏检测器
中图分类号: TP391.4 TH865 **文献标识码**: A **国家标准学科分类代码**: 520.20

VoxelFSD: voxel-based fully sparse detector with sparse convolution for 3D object detection

Zhou Weidian[1,2], Hong Ru[1,2], Gai Shaoyan[1,2], Da Feipeng[1,2]
(*1. School of Automation, Southeast University, Nanjing 210096, China; 2. Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University, Nanjing 210096, China*)

Abstract: Voxel-based 3D object detection methods often suffer from poor real-time performance when processing large-scale LiDAR point clouds due to their heavy dependence on dense 2D backbone networks. In this paper, we propose VoxelFSD, a voxel-based fully sparse 3D object detector that significantly enhances the real-time capability of long-range detection. The model features three core components: Firstly, parallel convolutional branches (PCB), which expand the receptive field and comprehensively extract object features while mitigating the impact of missing object center features; Then, a sparse region proposal network (SRPN) head that predicts objects sparsely, reducing redundant computations compared to dense prediction and thus improving efficiency for large-scale point clouds; Finally, an ROI head with an attention fusion module (AFM-ROI) that employs cross-attention to effectively fuse 3D backbone features with compressed bird's eye view (BEV) features in the second stage, refining object representation for improved detection accuracy. By removing the dense 2D backbone from traditional voxel-based detectors and integrating PCB and SRPN, we first present VoxelFSD-S, a fully sparse, single-stage, lightweight detector that achieves a superior balance between speed and accuracy relative to existing lightweight voxel-based models. Building upon VoxelFSD-S, we introduce VoxelFSD-T, a two-stage detector enhanced with AFM-ROI, which boosts accuracy with minimal additional computational cost. On the KITTI test set, VoxelFSD-S and VoxelFSD-T achieve

收稿日期:2024-11-26 Received Date: 2024-11-26
*基金项目:江苏省前沿引领技术基础研究专项(BK20192004C)、江苏省重大科技专项(BG2024003)项目资助

accuracies of 77.67% and 81.50% , respectively.

Keywords:3D object detection; voxel-based; lightweight; cross attention; fully sparse detector

0 引　　言

随着自动驾驶、机器人等领域研究的兴起,基于激光雷达点云的三维目标检测任务被广泛研究,深度学习方法也被应用于三维目标检测任务中[1-2]。激光雷达点云检测任务主要有 2 种方法:基于点的方法通常首先采用聚类、下采样等措施[3]或者借助图像中的二维检测框和几何投影关系[4]获取包含物体的目标点云,然后将重构清洗后的点云送入点云模型[5]预测前景点并回归物体包围框;然而,这些模型通常表现出有限的学习能力和效率,限制了其适用性。基于体素的方法首先对激光雷达点进行体素化,然后引入稀疏卷积,从而提高了检测性能和效率。

SECOND[6]是体素化处理点云的开创性工作,奠定了基于体素方法的半稀疏的三维检测框架的基础,整个检测框架包括前处理、三维稀疏骨干、二维密集骨干和检测头模块。点云经过体素化等前处理后,形成分辨率规整的稀疏三维特征图。三维稀疏骨干引入稀疏卷积操作,对稀疏三维特征图进行高效的特征提取。之后,稀疏特征图沿高度堆叠,并转换为密集的二维特征图。二维密集骨干通过卷积神经网络(convolutional neural network, CNN)进一步提取特征,同时向中心增强物体边缘特征,从而提高检测性能。目前,大多数现有工作都是在此框架基础上进行改进。Yin 等[7]提出了一种新颖的无需预设锚框的检测头,省去了人工选取锚框参数步骤,使模型更具鲁棒性。有的方法[8]尝试在三维骨干中引入了 Transformer[9]结构,相比于 CNN,Transformer 结构可以为特征图上的每个特征设置全局注意力窗口,扩大感受野。Voxel-R-CNN[10]在此流程中引入了两阶段网络,提出了感兴趣区域(region of interest,ROI)检测头,针对一阶段获取的 ROI,提出了基于体素的 ROI 池化操作,以纳入三维骨干包含的高度信息。ROI 头重新引入了高度信息,相比一阶段模型显著提高了检测效果,但是 ROI 头只利用了浅层三维骨干特征,这可能会导致最终结果的不稳定性。文献[11]在二阶段框架中引入对称形状生成。通过在生成候选框时预测前景点的镜像对称点,恢复目标整体形状。二阶段利用自注意力池化层从原始点和对称点聚合特征,用于修正候选框并完成三维框预测。Lang 等[12]则采取柱状编码网络将点云编码成伪二维图像特征,紧接着二维密集网路进行检测。文献[13]在 Lang 等[12]基础上设计了更为有效的柱状特征编码网络,引入注意力编码网络融合全局上下文信息提高特征图表征能力。

以上工作对现有体素框架做出了改进并取得了不错的效果,但未考虑大范围点云场景,半稀疏结构的模型设计在大范围点云感知中无法满足实时性要求。激光雷达扫描得到的物体点云往往比较稀疏,大多只能捕捉到物体的边缘点,投射到鸟瞰图(bird′s-eye-view,BEV)空间物体中心特征往往是缺失的。点云的稀疏性不利于物体的精确定位,而目前受于计算开销的限制,现有基于体素的检测器使用的三维骨干结构相对简单,特征提取能力相对较弱,因此需要密集的二维骨干的帮助,将边缘特征向中心扩散。但是,随着模型输入分辨率的增加,二维密集骨干的引入会导致计算成本的显著增加,这不利于远距离检测。现实自动驾驶场景中,激光雷达扫描范围往往会达到上百米距离,为保证检测效果,体素化后得到的特征图分辨率一般比较大,但由于激光雷达点的稀疏性,扫描的空间体积通常只被一小部分点占据,导致体素化特征图中出现大量冗余零值,在传统的二维 CNN 中,这些零值会造成不必要的计算,从而导致效率低下。目前仍缺乏能有效执行远距离感知的检测器。

针对上述挑战,一种名为 VoxelFSD 的全稀疏三维目标检测器被提出。首先,轻量级单阶段检测器 VoxelFSD-S 被提出。VoxelFSD-S 在三维骨干中引入了并行卷积分支(parallel convolutional branches, PCB)结构来增强三维稀疏骨干的特征表达能力。同时,VoxelFSD-S 放弃了使用密集的二维网络,提出了一种稀疏候选区域生成(sparse region proposal network, SRPN)检测头,提升检测效率,并去除非极大值抑制(non-maximum suppression, NMS)后处理操作,实现端到端检测。VoxelFSD-S 性能优于其他基于体素的轻量化检测器并且全稀疏设计能够实时处理长距离激光雷达点云感知任务。此外,为进一步提升模型性能,在 VoxelFSD-S 基础上引入注意力融合模块候选区域(attention fusion module-ROI, AFM-ROI)检测头,推出了两阶段检测器 VoxelFSD-T,进一步提高了模型的性能以满足更高精度要求的检测场景。AFM-ROI 检测头分别提取三维骨干的 ROI 特征和压缩高度后 BEV 的 ROI 特征,并利用交叉注意力将二者融合,从而增强了物体 ROI 特征的表示,并进一步精炼了候选框。VoxelFSD-S 和 VoxelFSD-T 在大规模激光雷达点感知任务中表现出了卓越的实时性能,这是目前现有基于体素方法检测器所不具备的。作为全稀疏检测器,VoxelFSD-S 可以实时满足现实场景中的大范围点云感知任务,适用于一些工业机器人等边缘设备,运用有限的算

力满足大部分场景的检测。VoxelFSD-T 可以以小部分额外计算开销为代价,提供更高精度的检测,适用于自动驾驶等高算力、高精度的场景。

1 全稀疏三维目标检测算法

全稀疏检测器 VoxelFSD-S 和 VoxelFSD-T 被提出,其中 VoxelFSD-S 是单阶段检测器,舍去密集二维骨干并引入 PCB 模块和 SRPN 检测头,实现全稀疏的结构设计,在大范围点云感知任务场景中能够实时检测。VoxelFSD-T 是双阶段检测器,在 VoxelFSD-S 基础上引入 AFM-ROI 检测头,性能更佳。两者在执行大规模激光雷达点感知任务时都表现出了高效性。VoxelFSD 流程如图 1 所示。

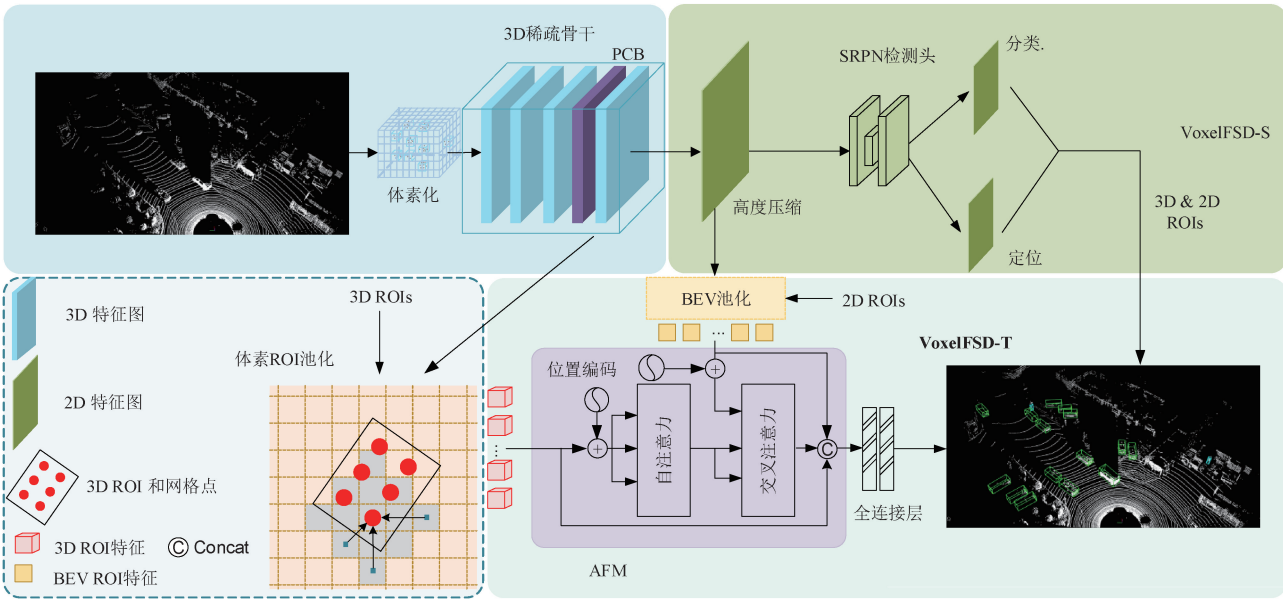


图 1 VoxelFSD 整体流程

Fig. 1 Overall pipeline of VoxelFSD

激光雷达点云在经过体素化操作之后被送入含有 PCB 模块的三维稀疏骨干,它可以代替原有的密集二维骨干,有效提高模型的特征提取能力,并针对物体中心点缺失问题(如图 2 所示),有效将物体边缘特征向中心扩散。经过三维稀疏骨干后的特征图沿着高度维度压缩,转换为二维 BEV 特征图,紧接着被送入 SRPN 检测头进行稀疏预测。SRPN 检测头只针对非空特征计算,这种稀疏设计可以提高模型的远距离感知能力,同时采用一对一匹配策略的训练方式省去了后处理 NMS 步骤,实现了端到端的检测。VoxelFSD-T 则进一步挑选出 SRPN 检测头中的候选框送入 AFM-ROI 头进行二阶段检测。AFM-ROI 头使用 AFM 模块,利用交叉注意力来融合从三维骨干中采集的 ROI 特征和从 BEV 特征图中汇集的 ROI 特征,从而使 SRPN 头提出的 ROI 得到更精确的细化得到物体最终的定位框,进一步提高了模型的性能。

1.1 PCB 特征提取模块

二维密集骨干在特征提取中起着重要作用,追求全稀疏的结构设计而放弃密集的二维骨干网络会导致模型感知能力的下降。在三维稀疏骨干中,通常选择大小为 3 的

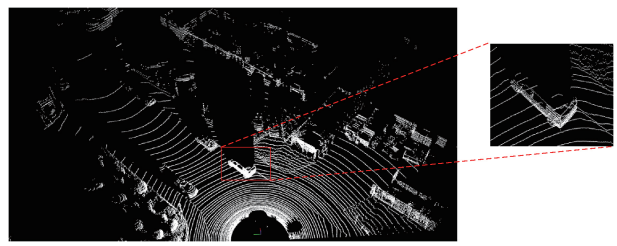


图 2 激光雷达点云

Fig. 2 LiDAR points

小卷积核进行稀疏卷积。然而,对于稀疏的激光雷达点云,这种小卷积核的特征提取能力是不够的。一些模型试图在三维骨干中引入 Transformer 结构来扩大模型的感受野[8,14-15]。此外,文献[16]还提出了一种适合三维稀疏骨干的大核结构。

上述方法在一定程度上提高了模型的性能,但模型结构相对复杂,额外引入较大的计算开销。相比于使用复杂的网络设计,一种简单的解决方案被提出:PCB 模块。记三维稀疏骨干 4 个阶段的特征图分别为 $\boldsymbol{F}_1$、$\boldsymbol{F}_2$、$\boldsymbol{F}_3$ 和 $\boldsymbol{F}_4$,具体来说,PCB 模块在三维骨干的最后一个阶段添加两个不同尺寸的卷积核进行并行处理,最后将每

个分支的结果进行融合。稀疏特征图的获取过程如式(1)所示。

$$\boldsymbol{F}_4 = net_1(\boldsymbol{F}_3) \oplus net_2(\boldsymbol{F}_3) \oplus net_3(\boldsymbol{F}_3) \tag{1}$$

其中,net 表示三维骨干中最后一个阶段的不同分支网络,⊕表示元素相加,具体网络结构见图 3。

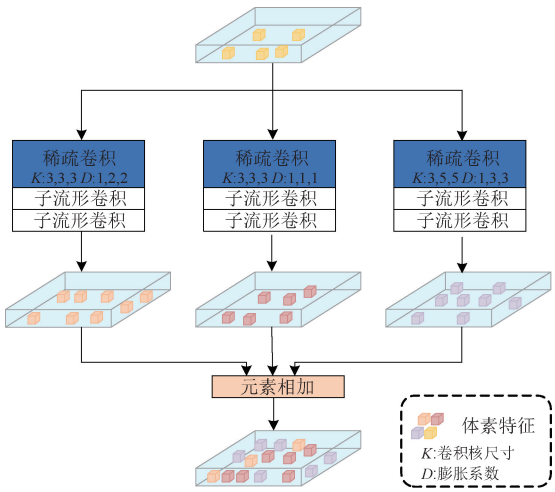


图 3　PCB 模块结构

Fig. 3　Structure of the PCB module

引入 PCB 模块有 3 个优势。首先,对于非均匀分布的激光雷达点云,采用不同尺寸卷积核的多分支结构可以扩大模型感受野,更好地提取特征。其次,在处理稀疏的激光雷达点时,经常会遇到物体中心特征缺失的问题,通过融合上述多分支的结果,有助于通过物体边缘特征学习得到中心特征,利于后续检测头检测,从而提高模型性能。最后,只在骨干网络的最后阶段添加并行卷积分支,这只会略微增加时间成本,这样的操作简单而有效。

1.2　SRPN 检测头

密集检测头通常用于二维检测和当前大多数基于体素的三维检测方法中,这些方法根据最终特征图的尺寸与原始空间尺寸的缩放比将特征图各个特征位置映射到原始空间中的位置来设置人工锚框,最后对部分锚框计算偏移得到目标预测框。这种方法对于图像是可行的,但是,激光雷达点本身就很稀疏,大部分没有点云的位置是不需要预设锚框参与后续计算的,在这些位置预设锚框,导致计算负担增加,尤其是在远距离感知任务中。相比之下,根据点云中的非空区域稀疏地设置锚框更为可行和高效。

在本节中,一种新颖的 SRPN 检测头被提出,直接处理三维主干的输出,而不是将特征图转换为密集形式。SRPN 检测头忽略特征图中的零值,并根据非空体素索引到原始空间的映射设置了锚框。与密集形式相比,这种方法大大减少了冗余锚框的数量,提高了效率。图 4 展示了密集预测方式(图 4(a))与稀疏预测形式(图 4(b))的对比,可以看到稀疏预测方式减少了很多没有点云区域的锚框放置,大大提高计算效率。

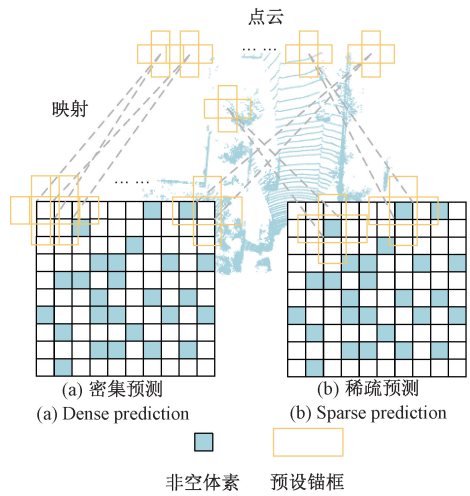


图 4　密集预测与稀疏预测对比

Fig. 4　Comparison of dense and sparse prediction

此外,SRPN 检测头引入 YOLOv10[17] 中的样本分配方式省去了 NMS 后处理的步骤。在训练过程中,设置了一个目标分配多个候选框的检测头和一个目标分配一个候选框的检测头,对二者进行联合训练,以提供额外的监督信号并增强训练收敛性。但是在推理过程中,只使用一对一匹配方式的检测头进行预测,从而省去了 NMS 后处理步骤,实现了端到端的检测。

1.3　AFM-ROI 检测头

设计二阶段模型时,在第 2 阶段引入了 BEV 特征图的 ROI 特征,与仅使用三维骨干 ROI 特征相比,能带来更稳定的结果。具体来说,将 SRPN 头生成的候选框 $\{x,y,z,l,w,h,\theta\}$ 投射到 BEV 特征图中得到 BEV 框 $\{x,y,l,w,\theta\}$。与普通的二维矩形框不同,得到的 BEV 框是旋转的,不能直接应用二维 ROI 池化,因此一种快速 BEV 池化方法被引入。首先,与普通 ROI 池化类似,将 BEV 框分割成足够多的区域,选择每个区域中心的网格点,得到它们的坐标 $\boldsymbol{P}_g \in \boldsymbol{R}^{B\times N\times g^2\times 2}$,如式(2)所示。

$$\boldsymbol{P}_g = \{(x_p, y_p) \mid p \in \boldsymbol{P}_g\} \tag{2}$$

其中,B 表示批次大小,N 是挑选的候选框数量,g 是候选框在每个维度上的分割网格数。然后 BEV 特征图 $\boldsymbol{F}_{bev}$ 上对应 $\boldsymbol{P}_g$ 的坐标位置,通过双线性插值得到网格点特征 $\boldsymbol{F}_g \in \boldsymbol{R}^{B\times N\times g^2\times C}$,其中 C 表示通道数。双线性插值具体过程如式(3)所示。

$$\begin{cases} \boldsymbol{f}_g(x,y_1) = \dfrac{x_2-x}{x_2-x_1}\cdot\boldsymbol{f}_{bev}(x_1,y_1) + \dfrac{x-x_1}{x_2-x_1}\cdot\boldsymbol{f}_{bev}(x_2,y_1) \\ \boldsymbol{f}_g(x,y_2) = \dfrac{x_2-x}{x_2-x_1}\cdot\boldsymbol{f}_{bev}(x_1,y_2) + \dfrac{x-x_1}{x_2-x_1}\cdot\boldsymbol{f}_{bev}(x_2,y_2) \\ \boldsymbol{f}_g(x,y) = \dfrac{y_2-x}{y_2-y_1}\cdot\boldsymbol{f}_g(x,y_1) + \dfrac{y-y_1}{y_2-y_1}\cdot\boldsymbol{f}_g(x,y_2) \\ \boldsymbol{F}_g = \{\boldsymbol{f}_g(x,y)\,|\,(x,y)\in\boldsymbol{P}_g\} \end{cases} \tag{3}$$

其中，(x,y) 是 $\boldsymbol{P}_g$ 中的坐标，(x_1,y_1)、(x_1,y_2)、(x_2,y_1)、(x_2,y_2) 是 BEV 特征图 $\boldsymbol{F}_{bev}$ 中与 (x,y) 相邻的坐标，$\boldsymbol{f}_g(x,y)$ 是 $\boldsymbol{F}_{bev}$ 在位置 (x,y) 的特征。最后，BEV ROI 特征 $\boldsymbol{F}_{2d}\in\boldsymbol{R}^{B\times N\times C}$ 经过全连接层调整得到，如式（4）所示。

$$\boldsymbol{F}_{2d} = FC(rs(\boldsymbol{F}_g)) \tag{4}$$

其中，rs 表示维度变换操作，FC 为全连接层。

然后使用 Voxel-R-CNN[10] 中的体素 ROI 池化操作得到三维骨干的 ROI 特征 $\boldsymbol{F}_{3d}$，并提出 AFM 将 $\boldsymbol{F}_{3d}$ 和 $\boldsymbol{F}_{2d}$ 进行融合，AFM 整体结构如图 5 所示。

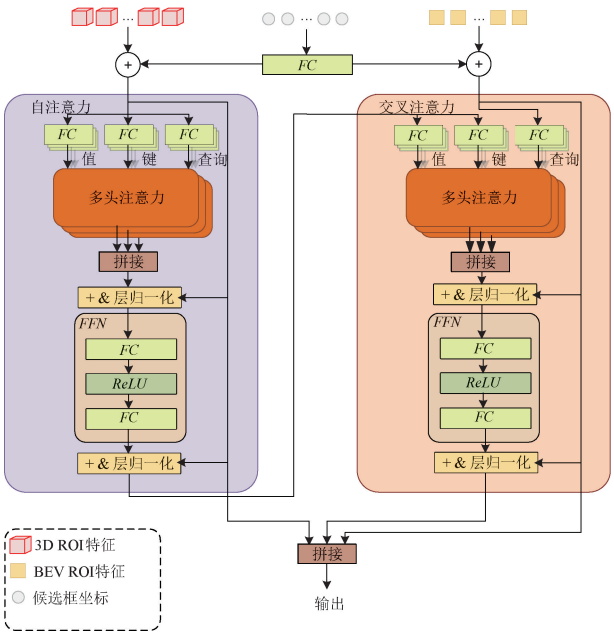


图 5 AFM 模型结构

Fig. 5 Structure of the AFM module

首先使用候选框坐标信息 $\boldsymbol{P}_c\in\boldsymbol{R}^{B\times N\times 7}$ 进行位置编码。形式上，候选框坐标通过全连接层得到位置编码特征 $\boldsymbol{P}_e\in\boldsymbol{R}^{B\times N\times C}$ 并分别与 $\boldsymbol{F}_{3d}$ 和 $\boldsymbol{F}_{2d}$ 相加，得到编码特征如式（5）所示。

$$\begin{cases} \boldsymbol{P}_e = FC(\boldsymbol{P}_c) \\ \hat{\boldsymbol{F}}_{3d} = \boldsymbol{F}_{3d} + \boldsymbol{P}_e \\ \hat{\boldsymbol{F}}_{2d} = \boldsymbol{F}_{2d} + \boldsymbol{P}_e \end{cases} \tag{5}$$

紧接着，使用多头注意力机制提升 $\hat{\boldsymbol{F}}_{3d}$ 的特征表达能力，计算公式如式（6）所示。

$$\begin{cases} \hat{\boldsymbol{F}}_{3d}^{i'} = \sigma\left(\dfrac{\boldsymbol{W}_q^i\cdot\hat{\boldsymbol{F}}_{3d}^i(\boldsymbol{W}_k^i\cdot\hat{\boldsymbol{F}}_{3d}^i)^{\mathrm{T}}}{\sqrt{C}}\right)\boldsymbol{W}_v^i\cdot\hat{\boldsymbol{F}}_{3d}^i \\ \hat{\boldsymbol{F}}_{3d}' = FFN(\delta(\,|\hat{\boldsymbol{F}}_{3d}^{1'},\hat{\boldsymbol{F}}_{3d}^{2'},\cdots,\hat{\boldsymbol{F}}_{3d}^{H'}|+\hat{\boldsymbol{F}}_{3d})) \end{cases} \tag{6}$$

其中，σ 表示 *softmax* 操作，H 是注意力头数，δ 表示层归一化操作，$\boldsymbol{W}$ 是用于特征映射的全连接层权重，FFN 是由全连接层和 *ReLU* 激活函数组成的前向神经网络。$|\cdots|$ 表示维度拼接操作。紧接着融合 $\hat{\boldsymbol{F}}'_{3d}$ 和 $\hat{\boldsymbol{F}}_{2d}$ 获得总的 ROI 特征 $\boldsymbol{F}_{roi}$，计算公式如式（7）所示。

$$\begin{cases} \boldsymbol{F}_{roi}^i = \sigma\left(\dfrac{\boldsymbol{W}_q^i\cdot\hat{\boldsymbol{F}}_{2d}^i(\boldsymbol{W}_k^i\cdot\hat{\boldsymbol{F}}_{3d}^{\prime i})^{\mathrm{T}}}{\sqrt{C}}\right)\boldsymbol{W}_v^i\cdot\hat{\boldsymbol{F}}_{3d}^{\prime i} \\ \boldsymbol{F}_{roi} = FFN(\delta(\,|\boldsymbol{F}_{roi}^1,\boldsymbol{F}_{roi}^2,\cdots,\boldsymbol{F}_{roi}^H|+\hat{\boldsymbol{F}}_{2d})) \end{cases} \tag{7}$$

最后，$\boldsymbol{F}_{3d}$、$\boldsymbol{F}_{roi}$、$\boldsymbol{F}_{2d}$ 被送入由全连接层组成的检测头进行检测。

2 实验结果及分析

2.1 数据集与评价指标

实验在 KITTI 数据集上进行以验证 VoxelFSD 的有效性。KITTI 数据集使用激光雷达和摄像头等传感器收集城市道路数据，其中包括图像、激光雷达点等模态数据。该数据集被认为是评估三维物体检测算法的主流数据集。KITTI 数据集分为 3 712 个训练帧、3 769 个验证帧和 7 518 个测试帧。KITTI 数据集主要包括 3 个注释类别：汽车、行人和骑自行车者；根据距离和遮挡等因素，每个对象类别又被分为 3 个不同的等级：容易、中等和困难。性能评估通常使用 11 个召回点（recall points 11，R11）或 40 个召回点（recall points 40，R40）下的平均精度（mean average precision，mAP）来评估算法在不同级别中的有效性。

2.2 实验细节

KITTI 侧重于对汽车的感知能力，VoxelFSD 中采取预设的汽车锚框的尺寸如下：长 3.9 m，宽 1.6 m，高 1.56 m，沿 Z 轴的高度坐标为 −0.85 m，放置在 0° 和 90° 两个方向上。模型的构建和实验基于 OpenPCDet 框架，单个 GPU 的批量大小设置为 4，初始学习率为 0.003，并使用了 Adam 优化器。

2.3 实验结果

实验在 KITTI 数据集验证集和测试集上进行，采用 mAP 指标来衡量模型的表现。其中，验证集上采用 R11，测试集上采用 R40。

1)验证集结果

表 1 显示了 VoxelFSD 与之前基于体素的方法在 KITTI 验证集上的比较结果,其中粗体结果表示最优。在单阶段轻量化模型中,在汽车类别的所有级别上,VoxelFSD-S 都优于已有的半密集检测器[6-7]。与另一全稀疏单阶段检测器 VoxelNext[18] 相比,VoxelFSD-S 在简单、中等和困难级别上分别提高了 2.51%、1.60% 和 1.53%,表明所提的全稀疏方法 VoxelFSD-S 相比 VoxelNext 效果更好。另外,VoxelFSD-S 的结果也明显优于具有 transformer 结构的检测器[8,14-15],从而证明了所提出的 PCB 模块在三维骨干网中的有效性。此外在双阶段模型中,VoxelFSD-T 在验证集上的表现也优于 PVRCNN[20] 和 Voxel R-CNN[10] 等模型,充分证明了 AFM-ROI 检测头的有效性。

表 1　不同方法在 KITTI 验证集上结果比较

Table 1　Comparison of results of different methods on the KITTI validation set

方法	*mAP* 简单	*mAP* 中等	*mAP* 困难
SECOND[6]	88.61	78.62	77.22
Pointpillars[12]	88.46	77.28	74.65
Centerpoint[7]	88.03	78.39	77.18
VoTR[8]	87.86	78.27	76.93
VoxSet[14]	88.45	78.48	77.07
OcTR[15]	88.43	78.57	77.16
VoxelNext[18]	86.95	77.47	76.51
PointRCNN[19]	88.88	78.63	77.38
PVRCNN[20]	89.35	83.69	78.70
Voxel R-CNN[10]	89.41	84.52	78.93
VoxelFSD-S	89.24	79.12	78.04
VoxelFSD-T	**89.46**	**84.68**	**79.01**

2)测试集结果

将 VoxelFSD 的结果提交给了 KITTI 基准测试。表 2 显示了 VoxelFSD-S 与其他基于体素的轻量化模型在 KITTI 测试集的结果对比,与验证集的结果表现一致,VoxelFSD-S 各项结果表现均取得了最优,相比于 Centerpoint[7] 模型,VoxelFSD-S 在简单、中等、困难 3 个级别上分别提高了 2.82%、1.55%、1.01%。

表 3 显示了 VoxelFSD-T 与其他两阶段方法在 KITTI 测试集上的检测结果。

与最新方法 Xview[26] 相比,VoxelFSD-T 在汽车类别的简单和中等级别上分别提高了 0.67% 和 0.15%,表明了其优越性能。在表中所列方法中,VoxelFSD-T 在中等级别中表现最佳,mAP 为 81.50%,超过了 PVRCNN[20] 等两阶段检测器,并在平均 mAP 中名列榜首,优于众多已有优秀的检测器。

表 2　轻量化模型在 KITTI 测试集上结果比较

Table 2　Comparison of results of lightweight models on the KITTI test set

方法	*mAP* 简单	*mAP* 中等	*mAP* 困难	平均 *mAP*
SECOND[6]	83.13	73.66	66.20	74.33
Pointpillars[12]	82.58	74.31	68.99	75.29
Centerpoint[7]	83.47	76.12	71.17	76.92
VoxelNext[18]	83.88	75.58	70.77	76.74
VoxelFSD-S	**86.29**	**77.67**	**72.18**	**78.71**

表 3　两阶段方法在 KITTI 测试集上结果比较

Table 3　Comparison of the results of the two-stage approach on the KITTI test set

方法	*mAP* 简单	*mAP* 中等	*mAP* 困难	平均 *mAP*
PointRCNN[19]	86.96	75.64	70.70	77.77
PI-RCNN[21]	84.32	74.82	70.03	79.39
MMF[22]	88.40	77.43	70.22	78.68
STD[23]	87.95	79.71	75.09	80.92
Fast-CLOCs[24]	89.10	80.35	76.99	82.15
CAT-Det[25]	89.87	81.32	76.68	82.62
Xview[26]	89.21	81.35	76.87	82.48
PVRCNN[20]	**90.25**	81.43	76.82	82.83
VoxelFSD-T	**89.89**	**81.50**	**76.82**	**82.74**

3)实时性及可视化分析

图 6 显示了 VoxelFSD-S 与其他轻量级方法在 70 m 范围时在 KITTI 测试集上实时性和准确性方面的比较。VoxelFSD-S 在速度方面仅次于 Pointpillars[12],但在准确性方面表现最佳,整体性能最好,也充分表明了 VoxelFSD-S 相比其他轻量化模型在速度和精度上达到了更好的平衡。

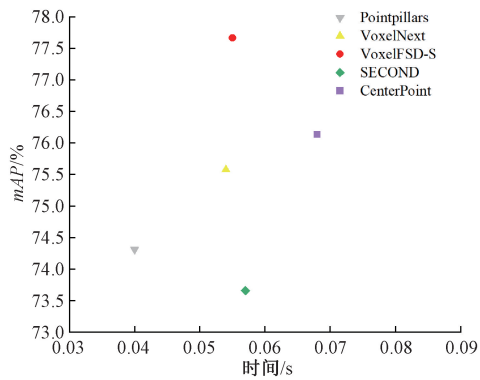


图 6　轻量化模型精确度于实时性对比

Fig. 6　Comparison of precision and real-time performance of lightweight models

表 4 定量展示了在处理不同范围的激光雷达点云时，VoxelFSD-S 与其他轻量化模型的成本对比。在推理开销上，随着点云检测范围的增加，VoxelFSD-S 模型速度和内存使用量上升跨度最小，从 70~200 m 范围，推理时间只从 0.055 s 增加到 0.059 s，显存消耗从 2.11 GB 增加到 2.58 GB。由于引入了 AFM-ROI 头，VoxelFSD-T 的时间和内存成本相比 VoxelFSD-S 略有增加，但仍在可接受的范围内。相比之下，不论是半稀疏结构的 SECOND[6]、Centerpoint[7] 还是 2D 密集网络 Pointpillars[12]，由于密集 2D CNN 的存在，推理时间都增加了 0.1 s 以上，显存占用增加超 2 GB。

表 4 不同感知距离下的模型推理成本比较

Table 4 Comparison of model inference cost at different perceptual distances

方法	成本	70 m	120 m	150 m	180 m	200 m
VoxelFSD-S	时间/s	0.055	0.056	0.058	0.059	0.059
	内存/GB	2.11	2.38	2.45	2.51	2.58
	mAP/%	79.12	79.02	78.85	78.56	78.49
VoxelFSD-T	时间/s	0.067	0.076	0.087	0.091	0.092
	内存/GB	2.38	2.67	2.81	2.96	3.25
	mAP/%	84.68	84.48	84.32	83.96	83.67
SECOND[6]	时间/s	0.057	0.078	0.093	0.135	0.159
	内存/GB	2.32	2.87	3.50	3.78	4.45
	mAP/%	78.62	78.11	77.80	77.35	77.33
Pointpillars[12]	时间/s	0.04	0.073	0.11	0.154	0.187
	内存/GB	2.36	2.82	3.31	4.05	4.70
	mAP/%	77.28	76.26	75.7	75.08	74.94
Centerpoint[7]	时间/s	0.068	0.087	0.107	0.160	0.194
	内存/GB	2.39	2.87	3.31	3.65	4.45
	mAP/%	78.39	77.99	77.73	77.42	77.09

表 4 中的结果充分证明了全稀疏结构设计的 VoxelFSD 在大规模激光雷达点云感知中的优势。图 7 展示了 VoxelFSD-T 在 KITTI 数据集上的汽车类别检测结果。其中方框为 VoxelFSD-T 检测框。通过结果可视化可知，VoxelFSD-T 能够对数据集中出现的汽车实例完成精确且鲁棒的检测。

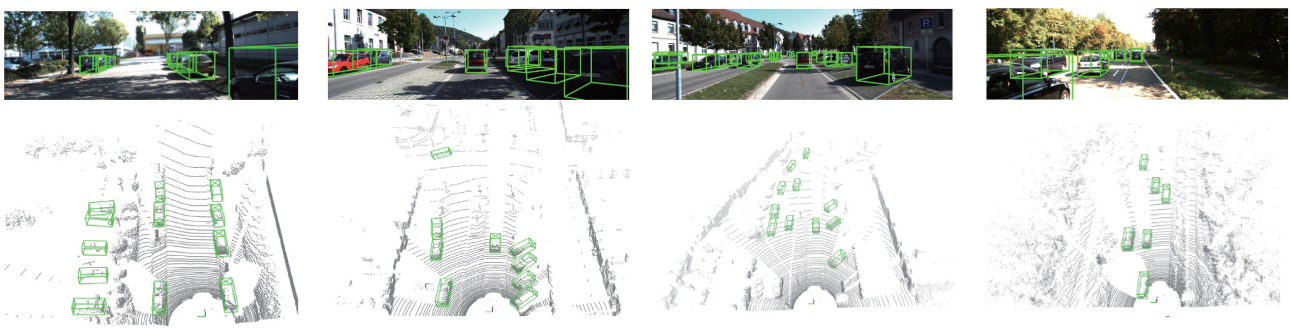


图 7 KITTI 测试集检测结果可视化

Fig. 7 Visualisation of KITTI test set detection result

2.4 消融实验

消融实验在 KITTI 验证集上进行，以验证各个模块的有效性，结果见表 5。

其中 baseline 为 SECOND[6] 去掉二维密集骨干，模块 1 表示 PCB 模块，模块 2 为 SRPN 检测头，模块 3 为 AFM-ROI 检测头。Baseline 在汽车类别上简单、中等和困难级别分别降低了 2.24%、7.17% 和 8.61%，说明在原有框架中，二维密集骨干起到了很重要的特征提取作用，

表 5　在 KITTI 验证集上不同模块的消融结果

Table 5　Ablation study results for different modules on the KITTI validation set

方法	模块 1	模块 2	模块 3	*mAP* 简单	*mAP* 中等	*mAP* 困难
baseline				86.37	71.45	68.61
baseline	√			88.82	78.65	77.70
baseline	√	√		89.24	79.12	78.04
baseline	√	√	√	**89.46**	**84.68**	**79.01**

但当在 baseline 中引入 PCB 模块时,汽车类各级别的结果分别提高了 2.45%、7.2% 和 8.91%,这表明 PCB 模块非常有效,可完全替代原来的二维密集骨干。通过进一步引入模块 2,各个级别的 mAP 分别再提高了 0.42%、0.47% 和 0.34%,最后引入模块 3 后,结果再一步得到了大幅度提高,最后完整模型在汽车各个级别上分别达到了 89.46%、84.68% 和 79.01% 的 mAP,表明了 AFM-ROI 头的有效性。

3　结　　论

首先分析了现有基于体素三维目标检测方法存在的不足之处,分析了激光雷达点云的稀疏特性,并提出了新颖的全稀疏检测器 VoxelFSD,用于精确的三维目标检测。首先提出了单阶段轻量级检测器 VoxelFSD-S,VoxelFSD-S 舍弃了以往基于体素的方法中使用的密集二维骨干,引入了 PCB 模块和 SRPN 头,进一步提高了模型性能,并实现了端到端的目标检测。在大规模点云感知方面,与之前的半密集检测器相比,VoxelFSD-S 表现出更优越的实时性,并且在 KITTI 测试集上的 mAP 达到了 77.64%,准确性和速度方面都优于之前的半密集轻量级检测器,其实时性和准确性足以满足大部分边缘设备上的大范围点云检测场景。同时,VoxelFSD-T 在 VoxelFSD-S 的基础上引入了 AFM-ROI 头,牺牲部分模型推理速度进一步提升检测精度,满足更高精度要求的检测场景需求。在 KITTI 测试集上,它的性能达到了 81.50%,也超过了之前众多双阶段检测器。

参考文献

[1] 陈慧娴,吴一全,张耀. 基于深度学习的三维点云分析方法研究进展[J]. 仪器仪表学报,2023,44(11):130-158.
CHEN H X, WU Y Q, ZHANG Y. Research progress of 3D point cloud analysis methods based on deep learning[J]. Chinese Journal of Scientific Instrument, 2023, 44(11):130-158.

[2] 葛俊彦,史金龙,周志强, 等. 基于三维检测网络的机器人抓取方法[J]. 仪器仪表学报,2021,41(8):146-153.
GE J Y, SHI J L, ZHOU ZH Q, et al. A robotic grasping method based on three-dimensional detection network[J]. Chinese Journal of Scientific Instrument, 2021,41(8): 146-153.

[3] 毕雪婷,刘小军,邵文远. 基于聚类方法的自动驾驶场景下的三维目标检测[J]. 电子测量技术,2021,44(6):103-107.
BI X T, LIU X J, SHAO W Y. 3D object detection in automatic driving scene clustering [J]. Electronic Measurement Technology, 2021, 44(6):103-107.

[4] 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.

[5] QI C R, SU H, MO K, et al. Pointnet: Deep learning on point sets for 3D classification and segmentation[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 77-85.

[6] YAN Y , MAO Y X, LI B. Second: Sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.

[7] YIN T W, ZHOU X Y, KRAHENBUHL P. Center-based 3D object detection and tracking[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 11779-11788.

[8] MAO J G, XUE Y J, NIU M ZH, et al. Voxel transformer for 3D object detection[C]. 2021 IEEE/CVF International Conference on Computer Vision, 2021: 3144-3153.

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. Advances in Neural Information Processing Systems, 2017, 30.

[10] DENG J J, SHI SH SH, LI P W, et al. Voxel R-CNN: Towards high performance voxel-based 3D object detection [C]. 35th AAAI Conference on Artificial Intelligence, 2021, 35(2): 1201-1209.

[11] 涂新奎,郑少武,于善虎, 等. 基于对称形状生成的三维目标检测网络[J]. 仪器仪表学报,2023,44(6):252-263.
TU X K, ZHENG SH W, YU SH H, et al. 3D object detection network based on symmetric shape generation[J]. Chinese Journal of Scientific Instrument, 2023,

44(6): 252-263.

[12] LANG A H, VORA S, CAESAR H, et al. Pointpillars: Fast encoders for object detection from point clouds[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 12697-12705.

[13] 汤新华,代道文,陈熙源, 等. 基于 PointPillars 的改进三维目标检测算法[J]. 仪器仪表学报,2024,45(9): 260-269.
TANG X H, DAI D W, CHEN X Y, et al. Improved three-dimensional object detection algorithm based on PointPillars[J]. Chinese Journal of Scientific Instrument, 2024, 45(9): 260-269.

[14] HE CH H, LI R H, LI SH, et al. Voxel set transformer: A set-to-set approach to 3D object detection from point clouds[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 8407-8417.

[15] ZHOU CH, ZHANG Y N, CHEN J X, et al. Octr: Octree-based transformer for 3D object detection[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 5166-5175.

[16] CHEN Y K, LIU J H, ZHANG X Y, et al. Largekernel 3D: Scaling up kernels in 3D sparse cnns[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 13488-13498.

[17] WANG AO, CHEN H, LIU L H, et al. YOLOv10: Real-time end-to-end object detection[J]. ArXiv preprint arXiv:2405.14458, 2024.

[18] CHEN Y K, LIU J H, ZHANG X Y, et al. Voxelnext: Fully sparse voxelnet for 3D object detection and tracking[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 21674-21683.

[19] SHI SH SH, WANG X G, LI H SH. Pointrcnn: 3D object proposal generation and detection from point cloud[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 770-779.

[20] SHI SH SH, GUO CH X, JIANG L, et al. PV-RCNN: Point-voxel feature set abstraction for 3D object detection[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 10526-10535.

[21] XIE L, XIANG CH, YU ZH X, et al. PI-RCNN: An efficient multi-sensor 3D object detector with point-based attentive cont-conv fusion module [C]. 34th AAAI Conference on Artificial Intelligence, 2020, 34(7): 12460-12467.

[22] LIANG M, YANG B, CHEN Y, et al. Multi-task multi-sensor fusion for 3D object detection[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 7345-7353.

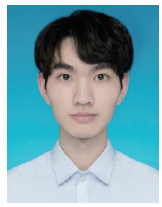
[23] YANG Z T, SUN Y N, LIU SH, et al. STD: Sparse-to-dense 3D object detector for point cloud [C]. 2019 IEEE/CVF International Conference on Computer Vision, 2019: 1951-1960.

[24] PANG S, MORRIS D, RADHA H. Fast-CLOCs: Fast camera-LiDAR object candidates fusion for 3D object detection [C]. 2022 IEEE Winter Conference on Applications of Computer Vision, 2022: 3747-3756.

[25] ZHANG Y N, CHEN J X, HUANG D. Cat-det: Contrastively augmented transformer for multi-modal 3D object detection [C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 898-907.

[26] XIE L, XU G D, CAI D, et al. X-View: Non-egocentric multi-view 3D object detector[J]. IEEE Transactions on Image Processing, 2023, 32: 1488-1497.

作者简介



周伟典,2022 年于江南大学获得学士学位,现为东南大学硕士研究生,主要研究方向为 3D 目标检测。

E-mail: 220222000@seu.edu.cn

Zhou Weidian received his B.Sc. degree from Jiangnan University in 2022. He is currently a M.Sc. student at Southeast University, with a major research interest in 3D target detection.

盖绍彦(通信作者),2008 年于东南大学获得博士学位,现为东南大学副教授、博士生导师,主要研究方向为计算机视觉、三维测量以及图像处理。

E-mail: qxxymm@163.com

Gai Shaoyan (Corresponding author) received his Ph.D. degree from Southeast University in 2008. He is currently an associate professor and Ph.D. supervisor at Southeast University, with main research interests in computer vision, 3D measurement, and image processing.