

基于CCF-PointPillars的3D点云目标检测

李子菲, 黄影平

(上海理工大学 光电信息与计算机工程学院, 上海 200093)

摘要: 3D目标检测是自动驾驶环境感知中的一项重要任务。为了解决自动驾驶场景中点云目标检测在稀疏小目标物体检测上精度低的问题,提出一种基于CCF-PointPillars的3D点云目标检测方法。首先,针对二维伪图像缺乏有效特征信息的问题,引入CBAM卷积注意力机制,动态调整对不同区域特征的关注,提取更具代表性的有效特征,抑制无关特征;其次,针对骨干网络仅使用简单卷积且特征提取能力不足的问题,将ConvNeXt V2集成到下采样子网络中,从而增强骨干网络的特征提取能力;最后,针对在特征融合阶段直接进行通道拼接,未能充分利用低层特征图空间信息的问题,借鉴FPN多尺度特征融合思路,重新设计了特征融合模块,通过融合浅层与深层的特征图,充分利用不同层次的特征信息。在KITTI数据集上的实验与评估结果表明,与现有的多种目标检测算法和原始算法相比,该方法在检测精度上有较好提升,尤其针对小目标物体,由此证明了该方法对点云稀疏小目标检测的有效性。

关键词: 3D目标检测;注意力机制;多尺度特征融合;自动驾驶

DOI: 10.11907/rjdk.251037

中图分类号: TP391.4

文献标识码: A

文章编号: 1672-7800(2026)002-0164-08



扫描二维码阅读全文

3D Point Cloud Object Detection Based on CCF-PointPillars

LI Zifei, HUANG Yingping

(School of Opto-Electronic and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: 3D object detection is critical in autonomous driving environment perception. To address the issue of low accuracy in detecting sparse and small objects in point cloud-based object detection for autonomous driving scenarios, this paper proposes a 3D point cloud object detection method based on CCF-PointPillars. Firstly, to tackle the lack of effective feature information in 2D pseudo-images, the CBAM convolutional attention mechanism is introduced to dynamically adjust the attention to features in different regions, extracting more representative and effective features while suppressing irrelevant ones. Secondly, to enhance the feature extraction capability of the backbone network, which relies solely on simple convolutions, ConvNeXt V2 is integrated into the downsampling sub-network. Finally, to address the issue of insufficient utilization of spatial information from low-level feature maps in the feature fusion stage, where channel concatenation is directly applied, this method redesigns the feature fusion module by incorporating the multi-scale feature fusion concept from FPN. The proposed approach fully leverages features at different levels by fusing shallow and deep feature maps. Experimental and evaluation results on the KITTI dataset demonstrate that, compared to existing detection algorithms and the original algorithm, the proposed method achieves significant improvements in detection accuracy, particularly for small objects, thereby validating its effectiveness for sparse small-object detection in point clouds.

Key Words: 3D object detection; attention mechanism; multi-scale feature fusion; autonomous driving

0 引言

目标检测是自动驾驶环境感知的重要任务,也是计算机视觉领域中兼具挑战性和应用价值的问题。其主要目的是在场景中定位出目标的位置并判断其类别,使车辆能

够实时了解周围环境变化,从而为自动驾驶系统的决策提供可靠依据。传统的目标检测方法主要依赖于二维图像信息,通过从图像中提取特征以识别和定位物体,但由于图像仅提供了平面上的信息,缺乏深度维度,无法准确捕捉物体的三维空间结构和距离信息。这导致在复杂场景中,尤其是多个物体叠加、遮挡或旋转的情况下,检测准确

收稿日期:2025-02-19

基金项目:国家自然科学基金面上项目(61374197);上海市自然科学基金项目(20ZR1437900)

作者简介:李子菲(1999-),女,上海理工大学光电信息与计算机工程学院硕士研究生,研究方向为计算机视觉;黄影平(1966-),男,上海理工大学光电信息与计算机工程学院教授、博士生导师,研究方向为汽车电子、计算机视觉。本文通讯作者:黄影平。

度和稳定性都会受到限制。

与此相对,激光雷达(LiDAR)点云数据可以捕捉场景的深度信息,直接反映物体的三维结构,如高度、宽度和纵深,极大地提高了目标检测准确性^[1]。具体方法包括:将LiDAR点云数据投影到二维平面并结合图像进行多模态检测(如SCANet^[2]、AVOD^[3]等);或是直接利用三维点云数据进行检测,构建3D卷积神经网络(如STD^[4]、Vote3Deep^[5]等),从而充分利用三维信息,捕捉物体的空间分布和几何特征。

然而,3D目标检测引入了深度信息和空间特征,使得检测网络面临更高的复杂度。一方面,点云数据稀疏且无序,需要设计适合点云结构的特征提取方法。比如,PointNet^[6]和PointNet++^[7]利用逐点特征学习的方式处理无序的点云;而VoxelRCNN^[8]、SECOND^[9]、PVT-SSD^[10]等方法则通过体素化处理,将点云转换为规则的三维栅格结构,以便于3D卷积操作。另一方面,3D目标检测需要预测物体的空间坐标、旋转角度和尺寸,因而比二维检测多了位置和角度的回归任务,增加了网络计算负担。

PointPillars(Fast Encoders for Object Detection from Point Clouds)是基于纯激光雷达点云的目标检测算法^[11]。该算法将三维点云划分成二维的柱子(pillars),每个柱子中的点被聚合为特征向量,也即将3D点云转换为二维伪图像,减少了处理复杂三维体素(voxel)的计算成本,同时也巧妙地将三维卷积转换为二维伪图像的卷积,提升了处理速度和效率,能够在保持较高精度的同时又有较快的推理速度。虽然该方法在处理大规模、高稠密度点云数据时表现出色,但是点云在空间中的分布具有稀疏性、不规则的特点,导致小目标(如行人、骑行者)因为体积较小、易被遮挡,从而使得目标检测效果不理想。

针对上述问题,本文提出了基于点云的3D目标检测模型CCF-PointPillars(Compounded Convolutional Attention Feature Fusion-Enhanced PointPillars)。主要贡献如下:①基于ConvNeXt V2重新设计了特征提取网络,将ConvNeXt V2集成到下采样子网络中,串联在原有的卷积层后面,增加网络深度,提高骨干网络的特征提取能力;②加入卷积块注意力模块(Convolutional Block Attention Module, CBAM)这一注意力机制以调整对不同区域特征的关注,帮助模型在处理稀疏或噪声较大的点云数据时,提取更具代表性的有效特征,抑制无关特征,解决伪图像特征表达能力不足、缺乏有效特征信息的问题;③借鉴特征金字塔网络(Feature Pyramid Networks, FPN)的多尺度特征融合思路,设计了特征融合模块,将浅层和深层特征图进行融合,以充分利用不同层次的特征信息,在保持较小计算量的前提下,解决物体检测中多尺度变化问题;④在KITTI数据集上进行了实验与评估,与多种现有目标检测算法进行比较并作可视化分析。结果表明,本文方法在检测精度上有较好提升,尤其针对行人和骑行者的小目标

物体。

1 本文方法

如图1所示,本文方法的网络结构主要由4部分组成:体柱特征网络(Pillar Feature Net)、卷积注意力模块(Convolutional Block Attention)、骨干网络(Backbone)和SSD检测头(Detection Head)。

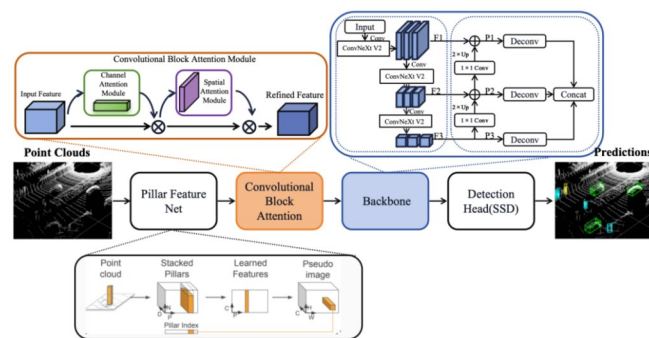


Fig. 1 The structure of the CCF-PointPillars network

图1 CCF-PointPillars网络结构

本文借鉴PointPillars的思想,将三维点云划分成二维的柱子(pillars),每个柱子中的点被聚合为特征向量。首先,在体柱特征网络后面引入卷积注意力模块,以增强对不同区域特征的关注,帮助模型在处理稀疏或噪声较大的点云数据时,提取更加有效的特征,解决伪图像特征表达不足的问题。其次,PointPillars模型采用简单的卷积网络进行特征提取,未能充分捕捉全局语义信息,影响检测精度。为此,本文基于ConvNeXt V2重新设计了特征下采样模块,构建了一种新的骨干网络,以提升点云的特征提取能力。最后,针对原算法上采样后直接进行通道拼接,未有效结合高分辨率与低分辨率特征的问题,本文借鉴FPN的多尺度特征融合思想,在骨干网络中设计了新的特征融合模块,通过浅层与深层特征图的融合,充分利用各层次特征,在保持较低计算成本的情况下,解决多尺度变化问题。

1.1 体柱特征网络

将点云数据划分为 $H \times W$ 的栅格,每个栅格在 Z 轴方向上对应一个柱状结构(pillar)。栅格内的点云数据被编码为九维向量 $D=(x, y, z, r, X_c, Y_c, Z_c, X_p, Y_p)$,其中 x, y, z, r 表示点的空间坐标和反射强度, X_c, Y_c, Z_c 为柱内点的几何中心, X_p, Y_p 为点相对于栅格中心的偏移量。

由于点云稀疏性,许多柱为空。为减少计算复杂度,限制柱的数量为 $P=12\ 000$,每个柱最多包含 $N=100$ 个点。如果点数超过 N 则随机采样,不足 N 则用0填充。这样,点云数据被编码为 (D, P, N) 的张量。随后,通过使用共享的多层感知机进行特征提取,将维度从 $D=9$ 压缩到 C ,生成 (C, P, N) 的张量。经过最大池化操作,得到 (C, P) 的张量,并最终映射为 (C, H, W) 的伪图像。

1.2 卷积注意力模块

在体柱特征提取网络中,点云特征的提取局限于预先划分的网格内,这可能导致目标被划分到不同的网格中。并且,在Pillar编码过程中,采用随机采样和最大池化操作,忽略了局部的点云信息,导致生成的伪图像特征表达能力不足,缺乏有效的特征信息。因此,本文在体柱特征网络后加入了CBAM卷积注意力机制。CBAM是一种深度学习领域的轻量级注意力机制,旨在增强卷积神经网络对图像特征的建模和表示能力^[12]。其主要目标是在卷积中引入通道注意力(Channel Attention Module, CAM)和空间注意力(Spatial Attention Module, SAM)以提高模型感知能力,以便网络能够自适应地关注和强调重要的特征,并抑制不重要的特征。

如图2所示,CBAM结构中的通道注意力能够增强各通道的特征表达,而空间注意力有助于捕捉不同空间位置的关键信息。CBAM注意力模块首先通过通道注意力,再通过空间注意力,从而形成一个串联的特征提取和学习过程。假设输入特征图为 $F \in \mathbb{R}^{C \times H \times W}$,将 F 分别应用于一维通道注意力 $M_c \in \mathbb{R}^{C \times 1 \times 1}$ 和二维空间注意力 $M_s \in \mathbb{R}^{1 \times H \times W}$ 。整个注意力机制的处理过程如式(1)和式(2)所示。

$$F' = M_c(F) \otimes F \quad (1)$$

$$F'' = M_s(F') \otimes F' \quad (2)$$

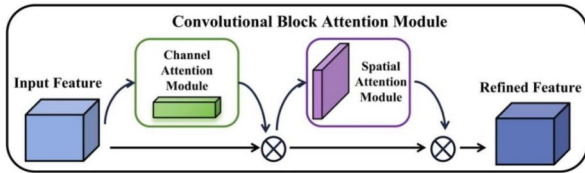


Fig. 2 The structure of CBAM module

图2 CBAM模块结构

(1)通道注意力机制。通道注意力机制旨在通过动态调整各通道的权重,使模型更高效地利用输入数据中的关键信息^[12]。该过程如图3所示,首先对输入特征图 F 进行全局最大池化和全局平均池化,以生成两种不同的特征描述。其次,这两个特征描述共享一个多层感知器网络(MLP):一个全连接层降低特征维度,另一个全连接层恢复原始维度。将两个特征图在通道维度上进行拼接,使用Sigmoid激活函数对拼接通道的权重进行归一化处理。最后,将归一化的权重逐元素与输入特征图相乘,生成强化后的特征图。其计算过程如式(3)所示。

$$M_c(F) = \sigma \left(\text{MLP} \left(\text{AvgPool}(F) \right) + \text{MLP} \left(\text{MaxPool}(F) \right) \right) \quad (3)$$

$$= \sigma \left(W_1 \left(W_0 \left(F_{avg}^c \right) \right) + W_1 \left(W_0 \left(F_{max}^c \right) \right) \right)$$

其中, W_0, W_1 是多层感知器的可学习参数。

(2)空间注意力机制。空间注意力机制是一种用于处理空间信息的注意力机制,主要用于在深度学习模型中动态地调整输入数据不同空间位置的重要性,从而突出关键信息并抑制无关信息^[12]。如图4所示,首先,输入的特征图沿通道维度进行最大池化和平均池化,以获得不同的特

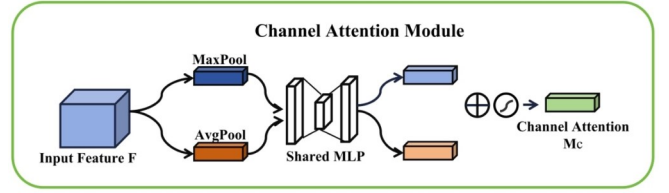


Fig. 3 The structure of channel attention module

图3 通道注意力模块结构

征表示;其次,沿通道维度合并池化特征,创建包含不同尺度上下文信息的特征图;再次通过卷积层处理该特征图,生成空间注意力权重,并通过Sigmoid激活函数将其归一化为0~1的范围;最后,这种空间注意力加权被应用于特征图,对每个空间位置的特征进行加权,以加强重要区域的信息,削弱背景区域的信息。其计算过程如式(4)所示。

$$M_s(F) = \sigma \left(f^{7 \times 7} \left(\left[\text{AvgPool}(F); \text{MaxPool}(F) \right] \right) \right) \quad (4)$$

$$= \sigma \left(f^{7 \times 7} \left(\left[F_{avg}^s; F_{max}^s \right] \right) \right)$$

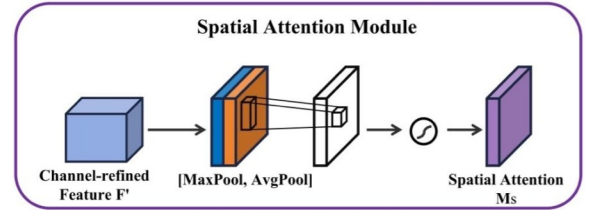


Fig. 4 The structure of spatial attention module

图4 空间注意力模块结构

本文通过引入CBAM注意力机制,调整对不同区域特征的关,从而帮助模型在处理稀疏或噪声较大的点云数据时,提取更具代表性的特征,从而提升目标检测精度。为了减少改进后网络的计算量并提升整体性能,将CBAM模块嵌入骨干网络中。通过强化和抑制相关特征信息,保留关键位置信息,从而提高网络检测精度。

1.3 基于ConvNeXt V2的特征提取骨干网络

骨干网络采用的是基于图像的卷积神经网络,对伪图像依次进行3次下采样操作,将特征图的尺寸不断进行压缩,并同时增加通道数。但是在原始算法中,骨干网络采用的简单卷积层特征提取能力不足,未能有效捕捉一些关键的局部特征,从而影响检测准确性。因此,本文重新设计了特征提取网络,将ConvNeXt V2集成到下采样子网络中,串联在原有的卷积层后面,在避免较大计算量的同时,增加网络深度,从而提升整体特征提取性能。图5为改进后的骨干网络结构。

ConvNeXt V2^[13]是一种卷积神经网络架构,它是在ConvNeXt V1^[13]的基础上改进而来,进一步增强了卷积网络的性能。ConvNeXt V2的架构设计融合了现代卷积网络和Transformer网络的设计精华。

(1)更大的卷积核。ConvNeXt V2使用了更大的卷积核尺寸(例如7×7或更大),这使得网络能够在减少参数数量的同时获取更大的感受野,提升网络的全局感知能力。

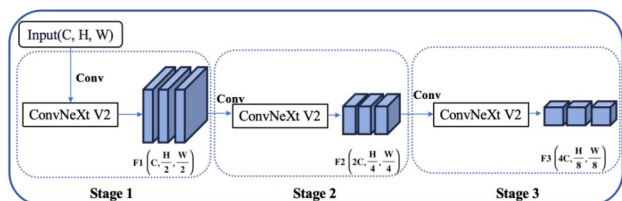


Fig. 5 The structure of feature extraction network based on ConvNeXt V2

图5 基于ConvNeXt V2的特征提取网络结构

(2)深度可分离卷积。该架构还引入了深度可分离卷积(Depthwise Separable Convolution, DSC),将传统卷积操作分解成两个更简单的操作:深度卷积和点卷积。深度卷积对于每个输入通道使用一个独立的卷积核,卷积核大小为 $k \times k$ 。因此,每个输入通道仅与一个 $k \times k$ 卷积核进行卷积操作,参数量为 $k \times k$ 。点卷积使用 1×1 卷积核将每个通道的输出与其他通道的信息相结合,生成最终的输出通道。在传统卷积中,一个 $k \times k$ 的卷积核在所有输入通道上进行卷积操作,因此需要大量参数。而在深度可分离卷积中,深度卷积只在每个输入通道上独立运用卷积核,点卷积在 1×1 卷积中融合通道信息,从而有效降低了参数量和计算复杂度,特别是在输出通道数较大时,参数量差异更为显著。

(3)全局响应归一化。ConvNeXt V2引入了全局响应归一化(Global Response Normalization, GRN)。GRN的核心思想是通过归一化整个网络的响应(激活),而不仅仅是单个层的响应^[13]。GRN主要通过3个步骤完成:全局特征聚合、特征归一化和特征校准。首先,聚合输入特征中的信息,提取全局的统计特征(如全局平均值),用于表示整体特征分布;其次,对聚合后的特征进行归一化处理,限制特征值的范围,抑制过高或过低的激活值;最后,根据归一化结果动态调整特征权重,强化重要特征、抑制无效特征。这种归一化方法考虑了整个网络中每一层的输出,并根据全局统计信息对其进行规范化,从而提高训练稳定性并加速收敛。

ConvNeXt V2 Block结构如图6所示。首先,对输入特征进行特征提取,并通过批量归一化以增强梯度的稳定性。接着,使用 1×1 卷积将特征维度扩展4倍,以获取更加丰富的语义信息。激活函数选择GELU代替传统的ReLU,以提升模型的非线性表达能力。此外,引入了GRN模块,用于增强通道之间的对比度和选择性。

1.4 基于FPN的多尺度特征融合

PointPillars通过骨干网络提取不同尺度的特征图,并使用反卷积操作对这些特征图进行上采样,将其尺寸统一后在通道维度上进行拼接。然而,在特征融合过程中仍存在以下问题:①该结构未能充分利用低层特征图的空间信息,融合每层的特征信息导致检测效果不佳,尤其是小目标物体;②上采样解卷积后直接进行通道拼接,容易导致浅层特征的空间信息和深层特征的语义信息不能有效融

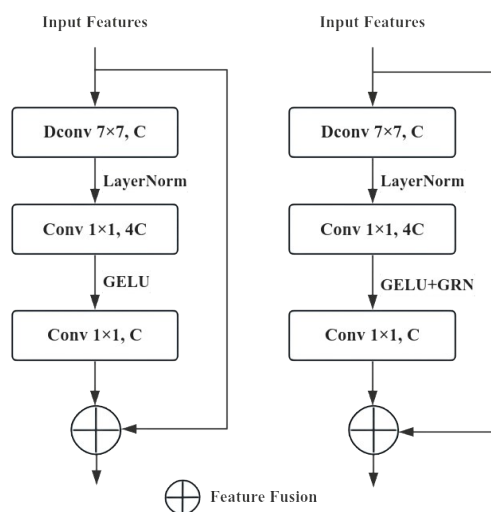


Fig. 6 The structure of ConvNeXt V1Block and ConvNeXt V2 Block

图6 ConvNeXt V1Block和ConvNeXt V2 Block结构

合。而FPN是一种常用的多尺度特征融合方法,在几乎不增加原有模型计算量的情况下,能够显著提升小物体目标的检测效果^[14]。因此,本文重新设计了特征融合过程,改进后的结构如图7所示。

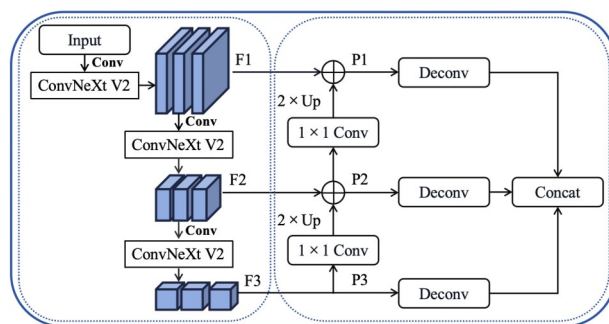


Fig. 7 The structure of FPN-based multi-scale feature fusion

图7 基于FPN的多尺度特征融合结构

网络结构由两部分组成,第一部分为自上而下,第二部分为自下而上。①自上而下:该部分对原始特征图逐级下采样,每层下采样的倍数递增,特征图尺寸缩小一半,而通道数则翻倍。FPN的输入是来自骨干网络3个阶段的特征图 $F=\{F1, F2, F3\}$,其分辨率分别是输入图像的 $(1/2, 1/4, 1/8)$;②自下而上:该部分对高层特征图进行2倍上采样,将通道数调整后与低层特征图融合,从而得到增强后的FPN特征图 $P=\{P1, P2, P3\}$ 。最后,分别对FPN每一层的输出调整通道数,并按照上采样率 $\{1, 2, 4\}$ 进行上采样,再沿通道维度拼接。

通过基于FPN的多尺度特征融合,增强了高层特征,使得每一层的特征图能够融合不同分辨率和语义强度的特征,从而显著提升了整体网络性能。

1.5 损失函数

目标检测模型结合了各种损失函数,主要包括定位(边界框坐标回归)和分类(物体类别的分类)。每个目标的边界框由 $(x, y, z, w, l, h, \theta)$ 表示,其中 x, y, z 指物体3D边

界框的中心点坐标; w 、 l 、 h 指边界框的宽度、长度和高度; θ 指边界框的偏航角。在边界框定位中,检测目标的真实框与预测框的差值定义如式(5)所示。

$$\begin{cases} \Delta x = \frac{x^{gt} - x^a}{d^a}, \Delta y = \frac{y^{gt} - y^a}{d^a}, \Delta z = \frac{z^{gt} - z^a}{h^a} \\ \Delta w = \log \frac{w^{gt}}{w^a}, \Delta l = \log \frac{l^{gt}}{l^a}, \Delta h = \log \frac{h^{gt}}{h^a} \\ \Delta \theta = \sin(\theta^{gt} - \theta^a) \end{cases} \quad (5)$$

其中, x^{gt} 、 x^a 分别代表真实框和预测框, $d^a = \sqrt{(w^a)^2 + (l^a)^2}$ 代表差值的归一化。

总损失函数由 3 部分组成:位置损失函数、分类损失函数和方向损失函数。

位置损失函数用 SmoothL1 函数表示,如式(6)所示。

$$L_{loc} = \sum_{b \in \{x, y, z, w, l, h, \theta\}} \text{SmoothL1}(\Delta b) \quad (6)$$

分类损失函数用 Focal Loss 函数表示,如式(7)所示。

$$L_{cls} = -\alpha_a (1 - p^a)^\gamma \log p^a \quad (7)$$

其中, P^a 代表 anchor 的类别概率, $\alpha=0.25$, $\gamma=2$ 。

总损失函数如式(8)所示。

$$L = \frac{1}{N_{pos}} (\beta_{loc} L_{loc} + \beta_{cls} L_{cls} + \beta_{dir} L_{dir}) \quad (8)$$

其中, N_{pos} 为正样本框的数量, $\beta_{loc} = 2, \beta_{cls} = 1, \beta_{dir} = 0.2$ 。

2 实验与结果分析

2.1 数据集

KITTI 数据集是由德国卡尔斯鲁厄理工学院和丰田美国技术研究院联合创建,用于评估计算机视觉算法在自动驾驶场景下的性能^[15]。本文使用 KITTI 数据集中的 3D 目标检测点云数据进行模型训练和验证,训练集包含 7 481 帧有标签样本,测试集包含 7 518 帧无标签样本。本文依据 PointPillars 方法对样本进行划分,将原始训练数据分为 3 712 帧训练集和 3 769 帧验证集,主要用于检测 3 种类型的目标,即汽车(Car)、行人(Pedestrian)和骑自行车者(Cyclist)。根据目标大小、截断程度以及是否被遮挡将目标分为 3 个难度等级:Easy、Moderate、Hard。其中,行人和骑行者由于目标体积较小且点云分布稀疏,通常被视为小目标。此外,困难(Hard)级别的目标通常伴随着更小的目标尺寸和更强的遮挡,结合中等(Moderate)和容易(Easy)级别的目标进行对比,可以全面评估算法在不同目标大小、截断程度和遮挡条件下的检测性能。

2.2 实验环境

实验中使用了基于 PyTorch 的深度学习框架。PyTorch 是目前深度学习领域的主流框架之一,拥有强大的 GPU 加速、动态计算图和自动微分等优点,在计算机视觉、自然语言处理等领域均有应用^[16]。实验中所使用的软件

和硬件设备配置如下:GPU:RTX 2080 Ti(11 GB)*1;CPU:12 vCPU Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50 GHz;内存:64 GB;显存:24 GB;系统:Ubuntu 20.04;环境:Python 3.7,PyTorch 1.10.1,OpenPCDet 0.6.0。同时,实验中使用 Adam 优化器,批次大小为 4,最大迭代次数为 80,学习率为 0.003,学习率衰减系数为 0.1,权重衰减为 0.01,动量为 0.9。

2.3 评价指标

根据 KITTI 数据集的评估标准,选择精度(Precision)、召回率(Recall)、平均精度(Average Precision, AP)和均值平均精度(mean Average Precision, mAP)作为模型性能的评价指标^[17]。

(1)精确率和召回率。精确率表示在所有被预测为正的样本中,实际为正的样本所占比例,如式(9)所示。召回率表示实际为正的样本中被正确识别为正的样本,体现了模型识别正类目标的能力,如式(10)所示。

$$\text{Precision} = \frac{TP}{TP + FP} \quad (9)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (10)$$

其中,TP(True Positive)表示被正确预测为正的样本数量,FP(False Positive)表示错误地将实际为负的样本预测为正,FN(False Negative)表示错误分类为负样本的正样本数量。

(2)平均精度(AP)和平均精度均值(mAP)。在目标检测任务中,最常用的评估标准为平均精度(AP)和平均精度均值(mAP)。AP是通过计算精确度—召回率(PR)曲线下的面积而得出,如式(11)所示。mAP是对所有 c 个类别的平均精度进行求平均,表示模型在各类别上的总体表现,如式(12)所示。

$$AP = \int_0^1 P(R) dR \quad (11)$$

$$mAP = \frac{1}{|c|} \sum_{c \in C} AP_c \quad (12)$$

2.4 性能评估及其他方法比较

为了直观地说明 3D 目标检测的性能,将本文方法与 MV3D^[18]、AVOD-FPN^[19]、3DSSD^[20]、VoxelNet^[21]、SECOND、Point-GNN^[22]、EPNet++^[23]、TANet^[24]、BirdNet+^[25]、IA-SSD^[26]等多种方法进行性能比较,主要针对车辆(Car)、行人(Pedestrian)和骑行者(Cyclist)3种目标在平均精度(AP)上的性能评估。本文 3D 目标检测模型在 KITTI 数据集的 3D 模式下的检测结果与对比如表 1 所示。其中,N/A 表示该方法未公开相应类别的检测结果。

从对比结果可以看出:

(1)在简单、中等、困难 3 种场景中,本文方法对汽车、行人和骑行者三类目标的平均检测精度分别是 81.84%、52.90%、71.24%。与原始的 PointPillars 方法相比,汽车、行人和骑行者这三类目标的三维目标检测准确率分别提高了 1.13%、3.80% 和 3.65%,平均精度提高了 2.86%。相较

Table 1 Results of 3D object detection models on KITTI validation set
表 1 3D 目标检测模型在 KITTI 点云验证集上的检测结果

Method	Modality	Car.(IoU≥0.7)			Ped.(IoU≥0.5)			Cyc.(IoU≥0.5)		
		Easy	Mod.	Hard	Easy	Mod.	Hard	Easy	Mod.	Hard
MV3D	LiDAR	71.09	62.35	55.12	N/A	N/A	N/A	N/A	N/A	N/A
AVOD	&Img. LiDAR	81.94	71.88	66.38	46.35	39.00	36.58	59.97	46.12	42.36
-FPN	&Img. LiDAR	91.37	81.96	76.71	52.79	44.38	41.29	76.15	59.71	53.67
EPNet++	&Img. LiDAR	77.47	65.11	57.73	39.48	33.69	31.51	61.22	48.36	44.37
VoxelNet	LiDAR	83.13	73.66	66.20	51.07	42.56	37.29	70.51	53.85	46.90
SECOND	LiDAR	88.36	79.57	74.55	54.64	44.27	40.23	82.48	64.10	56.90
3DSSD	LiDAR	88.33	79.47	72.29	51.92	43.77	40.14	78.60	63.48	57.08
Point -GNN	LiDAR	83.81	75.38	67.66	54.92	46.67	42.42	73.84	59.86	53.46
TANet	LiDAR	76.15	64.04	59.79	65.67	53.84	49.06	41.55	35.06	32.93
BirdNet+	LiDAR	88.34	80.13	75.04	46.51	39.03	35.60	78.35	61.94	55.70
IA-SSD	LiDAR	87.92	78.53	75.68	53.87	48.85	44.57	80.86	63.24	58.67
PointPil- lars	LiDAR	88.95	80.03	76.54	57.60	53.32	47.78	84.15	67.83	61.74
Ours	LiDAR									

于其他方法,本文方法取得了较优的检测精度,尤其改善了对小目标的检测效果。该结果表明,融合 CBAM 注意力机制能够调整对不同区域特征的关注,提取更具代表性的特征。同时,在二维骨干网络中将 ConvNeXt V2 模块与原有简单卷积层串联,从而增强网络特征提取能力,获得更加丰富的特征信息。最后,通过基于 FPN 的多尺度特征融合进行特征整合,从而提高了检测精度。

(2)在基于 LiDAR 与图像融合的方法中,MV3D 在汽车类别上的精度最低,中等和困难场景仅为 62.35% 和 55.12%,适应复杂场景能力不足。AVOD-FPN 在行人和骑行者检测中表现不足,困难场景精度仅为 36.58% 和 42.36%。EPNet++ 在融合方法中表现较优,但在骑行者类别上,困难场景精度仅为 53.67%,仍低于本文方法。在 LiDAR-only 方法中,VoxelNet 精度较低,车辆类别在中等和困难场景的精度仅为 65.11% 和 57.73%。SECOND 通过稀疏卷积优化,行人中等和困难场景的精度分别提升至 42.56% 和 37.29%,但仍不及本文方法的表现。3DSSD 和 Point-GNN 在各目标类别上表现较为均衡,骑行者困难场景精度分别为 56.90% 和 57.08%。相比之下,本文算法在各类别和难度场景下表现出明显优势。在汽车类别的检测中,本文算法在中等和困难场景的精度分别达到 80.03% 和 76.54%,显著高于 VoxelNet 和 SECOND。在行人检测中,本文算法在中等和困难场景的精度分别达到 53.32% 和 47.78%,较 TANet 分别提升了 6.65% 和 5.36%。在骑行者检测中,本文算法在所有难度下均取得最佳表现,困难场景精度达到 61.74%,明显优于 IA-SSD(55.70%) 和 TANet(53.46%)。这些结果表明,本文算法通过优化特征提取和融合策略,显著提升了整体检测精度。

2.5 典型场景下可视化结果分析

选取 KITTI 数据集中的 6 个典型场景,将本文模型应

用于这些场景的检测结果并进行可视化。图 8 展示了不同场景下的目标检测效果,分别突出目标密集场景、远距离稀疏小目标场景,以及光照条件变化对检测结果的影响。通过对这些场景的分析,可以清晰地看到本文算法在应对多样化挑战时的表现。上半部分展示了真实场景中的图像,下半部分则呈现了相应点云数据的 3D 目标检测结果。在点云目标检测结果中,分别用绿色检测框代表车辆类别,蓝色检测框代表行人类别,黄色检测框代表骑行者类别。

图 8(a) 和 (b) 展示了目标密集、种类较多的场景。在这种场景下,汽车数量多且分布密集,同时存在大量遮挡情况,如车辆彼此遮挡。本文方法能够准确检测目标位置和类别,即使是那些部分被遮挡的汽车也能被有效识别。图 8(c) 和 (d) 展示了远距离稀疏小目标的场景,这类场景下,远处的行人和骑行者由于距离激光雷达较远,其点云非常稀疏。在强光照和弱光照的情况下,稀疏小目标的检测极具挑战性,因为较少的点云信息使得物体特征难以提取。本文方法能够在这种稀疏场景下保持较好的检测性能,无论是在强光还是暗光环境中,都能检测到远距离的行人和骑行者。图 8(e) 和 (f) 分别展示了在强光和暗光条件下的 3D 点云目标检测结果。在暗光环境下,远距离被遮挡汽车依然能够被清晰检测出,说明本文方法具有很好的光线鲁棒性,能够在低照明条件下保持稳定的检测能力。这对于实际应用具有重要意义,尤其是在夜间或隧道等低光环境中,可靠的目标检测对于确保车辆的安全行驶至关重要。相比之下,强光下的检测结果虽然也能很好地识别车辆,但对于远处的曝光区域,检测效果稍有下降。这是因为强光环境下点云的反射特性发生了变化,影响了远距离物体的特征提取。

但是,该方法也存在一些漏检、误检情况。导致这些检测失误的主要原因有:

(1)激光雷达采集的点云密度与目标的距离成反比,近处物体的点云较密集,而远处物体的点云较稀疏。稀疏点云中的物体信息有限,难以提取足够的特征,尤其是对小目标或细长目标(如行人、骑行者)。因此,这种密度的不均匀性可能导致远处目标漏检,近处复杂场景中的物体边界模糊化则可能造成误检。

(2)当背景环境点云密集,并且部分点云的轮廓和行人、骑行者等小目标非常相似时,很容易将背景点误检成行人、骑行者等。例如:图 8(a) 场景下,将路边较矮的路桩误检测成行人;图 8(b) 场景下,将路边的树枝和墙壁误检测成行人和骑行者等。

(3)在强光环境下,激光雷达的反射特性发生变化,导致部分物体的点云缺失和稀疏,丢失部分重要的特征信息,进而影响物体检测精度,造成误检或漏检。

总体而言,这些场景展示了本文算法在处理不同复杂度任务时的适应性。无论是面对密集目标、远距离稀疏小

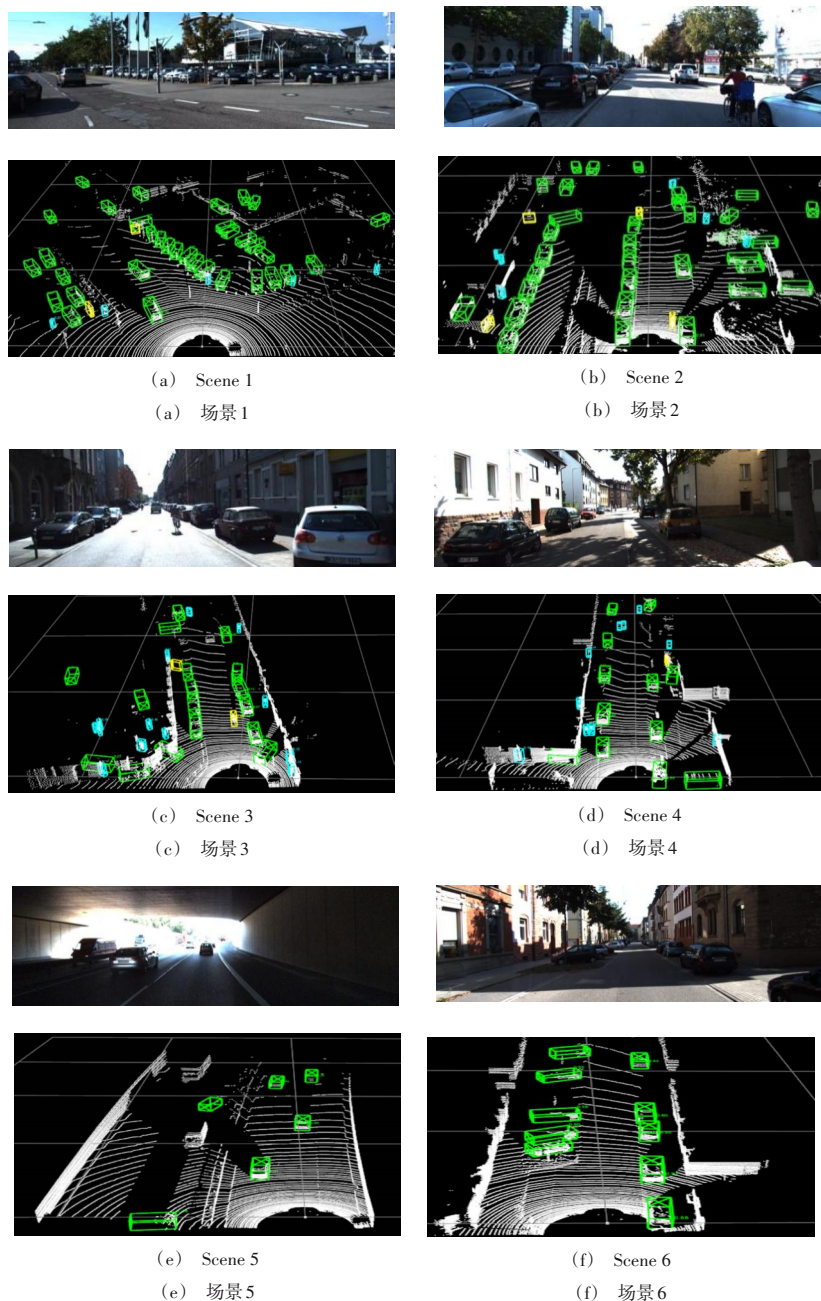


Fig. 8 Images and 3D object detection results in different scenes

图8 不同场景下图像和3D目标检测效果

目标,还是光照变化,该算法都能保持较高的检测精度和鲁棒性,展示了其在现实应用中的潜力。

2.6 消融实验

为了验证不同改进模块对性能的影响,对汽车、行人和骑行者三类目标进行了消融实验,以分析各模块对整体模型检测性能的影响。3D模式下不同模块的检测结果比较如表2所示。

消融实验共有5组对比实验,基线模型为PointPillars。实验1在骨干网络前添加CBAM卷积注意力机制,在汽车、行人、骑行者3个类别中的精度分别提高了0.12%、1.15%、2.34%,平均精度提高了1.20%。实验2在骨干网络中改进

Table 2 Comparison of detection results of different modules in 3D mode

表2 3D模式下不同模块的检测结果比较

Experiment	CBAM	FPN	ConvNeXt V2	AP/%			mAP/%
				Car	Ped.	Cyc.	
Baseline				80.71	49.10	67.59	65.80
Experiment 1	√			80.83	50.25	69.93	67.00
Experiment 2		√		80.67	49.35	68.73	66.25
Experiment 3	√	√		81.20	51.74	69.95	67.63
Ours	√	√	√	81.84	52.90	71.24	68.66

了基于FPN的多尺度特征融合,在汽车、行人、骑行者3个类别中的精度分别提高了-0.04%、0.25%、1.14%,平均精度提高了0.45%。实验3同时添加CBAM卷积注意力机制

和多尺度特征融合,在汽车、行人、骑行者3个类别中的精度分别提高了0.49%、2.64%、2.36%,平均精度提高了1.83%。实验4为本文方法,在汽车、行人、骑行者3个类别中的精度分别提高了1.13%、3.80%、3.65%,平均精度提高了2.86%。

综上所述,本文提出的添加CBAM注意力机制、使用ConvNeXt V2改进骨干网络以及基于FPN的多尺度特征融合均能够提高网络整体性能,尤其是检测小目标的性能(如:行人和骑行者)。

3 结语

本文针对点云稀疏小目标物体检测精度较低的问题,提出了CCF-PointPillars的点云3D目标检测方法,解决了二维伪图像有效特征缺乏等问题,并对点云特征提取网络和多尺度特征融合进行了改进。在KITTI数据集上进行模型训练和验证,实验结果表明,本文方法相比于原始方法,在3D目标检测汽车、行人和骑行者3个目标类别上的平均精度分别提高了1.13%、3.80%、3.65%,平均精度提高了2.86%,有效地提高了模型性能,尤其是针对行人和骑行者的小目标类别。同时,与其他多种目标检测方法相比,本文方法在检测精度方面具有明显优势。

但是,将场景中的背景点云误检测为目标的情况仍存在。后续研究中,需进一步提升和优化模型,可拓展特征的丰富性,增加视觉图像的特征信息,并尝试多模态数据融合的3D目标检测。

参考文献:

- [1] WU S W, LI Y X, ZHANG S C, et al. 3D object detection methods based on point cloud with deep learning [J]. Journal of Telemetry, Tracking and Command, 2024, 45(5): 1-18.
武淑文, 李燕娣, 张少琛, 等. 基于深度学习的3D点云目标检测研究综述[J]. 遥测遥控, 2024, 45(5): 1-18.
- [2] LU H H, CHEN X S, ZHANG G Y, et al. SCANet: spatial-channel attention network for 3D object detection [C]//Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, 2019: 1992-1996.
- [3] KU J, MOZIFIAN M, LEE J, et al. Joint 3d proposal generation and object detection from view aggregation [C]//Madrid: The 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2018.
- [4] YANG Z T, SUN Y A, LIU S, et al. STD: sparse-to-dense 3D object detector for point cloud [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1951-1960.
- [5] ENGELCKE M, RAO D, WANG D Z, et al. Vote3deep: fast object detection in 3D point clouds using efficient convolutional neural networks [C]//Proceedings of the 2017 IEEE International Conference on Robotics and Automation, 2017: 1355-1361.
- [6] CHARLES R Q, SU H, KAICHUN M, et al. PointNet: deep learning on point sets for 3D classification and segmentation [C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 77-85.
- [7] QI C R, YI L, SU H, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 5105-5114.
- [8] DENG J, SHI S, LI P, et al. Voxel R-CNN: towards high performance voxel-based 3d object detection [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 1201-1209.
- [9] YAN Y, MAO Y, LI B. Second: sparsely embedded convolutional detection [J]. Sensors, 2018, 18(10): 3337.
- [10] YANG H H, WANG W X, CHEN M H, et al. PVT-SSD: single-stage 3D object detector with point-voxel transformer [C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 13476-13487.
- [11] LANG A H, VORA S, CAESAR H, et al. PointPillars: fast encoders for object detection from point clouds [C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 12689-12697.
- [12] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module [C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 3-19.
- [13] WOO S, DEBNATH S, HU R, et al. ConvNeXt V2: co-designing and scaling convnets with masked autoencoders [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 16133-16142.
- [14] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2117-2125.
- [15] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: the kitti dataset [J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.
- [16] PASZKE A, GROSS S, MASSA F, et al. Pytorch: an imperative style, high-performance deep learning library [J]. Advances in Neural Information Processing Systems, 2019, 32: 8024-8035.
- [17] PADILLA R, PASSOS W L, DIAS T L B, et al. A comparative analysis of object detection metrics with a companion open-source toolkit [J]. Electronics, 2021, 10(3): 279.
- [18] CHEN X Z, MA H M, WAN J, et al. Multi-view 3d object detection network for autonomous driving [C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1907-1915.
- [19] KU J, PON A D, WALSH S, et al. Improving 3d object detection for pedestrians with virtual multi-view synthesis orientation estimation [C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019: 3459-3466.
- [20] YANG Z T, SUN Y N, LIU S, et al. 3DSSD: point-based 3d single stage object detector [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11037-11045.
- [21] ZHOU Y, TUZEL O. VoxelNet: end-to-end learning for point cloud based 3d object detection [C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 4490-4499.
- [22] SHI W J, RAJKUMAR R. Point-GNN: graph neural network for 3D object detection in a point cloud [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 1711-1719.
- [23] LIU Z, HUANG T, LI B, et al. EPNet++: cascade bi-directional fusion for multi-modal 3D object detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45: 8324-8341.
- [24] LIU Z, ZHAO X, HUANG T, et al. Tanet: robust 3D object detection from point clouds with triple attention [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 11677-11684.
- [25] BARRERA A, BELTRÁN J, GUINDEL C, et al. BirdNet+: two-stage 3D object detection in LiDAR through a sparsity-invariant bird's eye view [J]. IEEE Access, 2021, 9: 160299-160316.
- [26] ZHANG Y, HU Q, XU G, et al. Not all points are equal: learning highly efficient point based detectors for 3d lidar point clouds [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 18953-18962.