DOI:10.19651/j.cnki.emt.2417519

基于 ConvNeXt 和可变形交叉注意力的多模态 3D 目标检测方法*

周 鹏[1] 宋志强[2] 胡 凯[1] 宋利鹏[1] 李明阳[1]
(1.南京信息工程大学自动化院 南京 210044;2.无锡学院自动化院 无锡 214105)

摘 要:近年来,随着新能源汽车的快速发展,3D 目标检测作为自动驾驶技术的核心基础正变得愈发重要。融合雷达点云与图像等多模态信息的策略,能够显著提升目标检测的准确性与鲁棒性。受 BEVDet 启发,本研究提出了一种基于 BEV(鸟瞰图)视角的改进多模态融合 3D 目标检测方法。该方法采用 ConvNeXt 网络结合 FPN-DCN 结构高效提取图像特征,并通过可变形交叉注意力机制实现图像与点云数据的深度融合,从而进一步提升模型的检测精度。在 nuScenes 自动驾驶数据集上的实验表明,本研究模型性能优异,在测试集上的 NDS 达到了 64.9%,显著超越了大多数现有检测方法。

关键词:自动驾驶;3D 目标检测;多模态融合;可变形交叉注意力机制
中图分类号:TN958.98 **文献标识码**:A **国家标准学科分类代码**:510.4

Multimodal 3D object detection method based on ConvNeXt and deformable cross attention

Zhou Peng[1] Song Zhiqiang[2] Hu Kai[1] Song Lipeng[1] Li Mingyang[1]
(1. School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China;
2. School of Automation, Wuxi University, Wuxi 214105, China)

Abstract: In recent years, with the rapid development of new energy vehicles, 3D object detection, as a core foundation of autonomous driving technology, has become increasingly important. Strategies that integrate multimodal information, such as radar point clouds and images, can significantly enhance the accuracy and robustness of object detection. Inspired by BEVDet, this paper proposes an improved multimodal fusion 3D object detection method based on the BEV (bird's eye view) perspective. The method employs a ConvNeXt network combined with an FPN-DCN structure to efficiently extract image features and utilizes a deformable cross-attention mechanism to achieve deep fusion of image and point cloud data, thereby further enhancing the detection accuracy of the model. Experiments on the nuScenes autonomous driving dataset demonstrate the superior performance of our model, with an NDS of 64.9% on the test set, significantly outperforming most existing detection methods.

Keywords: autonomous driving;3D target detection;multimodal fusion;deformable cross-attention mechanism

0 引 言

近年来,新能源汽车高速发展,智能驾驶系统已经逐渐进入到了大众的工作和生活中。在大多数的自动驾驶场景中,使用多个传感器来收集不同类型的数据,能够利用多类型传感器捕获更加丰富的环境信息,获得更好的目标检测精度。基于多传感器的 3D 目标检测多以某一个传感器信息为主,其他传感器信息为副的方式进行特征融合,例如摄像头和测距传感器的组合,这可能导致某一传感器信息的缺失,从而影响整体 3D 目标检测性能。基于多传感器融合的目标检测不断提高其检测性能,但不能很好地统一多模态信息,如何利用深度学习的方法对数据进行处理,如何整合不同传感器的数据进行训练,提高 3D 模型训练识别的准确性,也是研究的重要内容之一。

本文受到 BEVDet[1] 的启发,在原有的架构上,提出了一个基于 BEV 视角的多模态融合 3D 目标检测方法。

收稿日期:2024-12-02
*基金项目:江苏省产学研合作项目(BY20230688)资助

BEVDet 是一种针对目标检测的深度学习网络，相较于传统的目标检测网络，BEVDet 则将目标检测转换为鸟瞰视角(从上方看)的空间问题。BEVDet 利用多视角或多传感器的数据(如相机、激光雷达、雷达等)，通过将图像特征转换成图像 BEV 特征，将图像 BEV 特征和 BEV 特征融合与变换，生成鸟瞰图并在此视角下进行目标检测。

BEVDet 以其全面的空间信息捕捉能力，在物体定位的准确性上具有显著优势。通过多传感器数据的融合，BEVDet 在复杂环境中展现出了更高的稳定性和精确度。然而，这一技术也面临着挑战：多传感器融合过程中所需的大量数据处理，使得原有模型在特征提取速度上已难以满足需求。此外，在动态且复杂的环境条件下，雷达与图像数据融合时可能会出现在鸟瞰视图中物体位置的偏差。针对这些挑战，提出了相应的改进措施。本文的主要贡献可以概括为以下几点：

1)设计了 ConvNeXt[2] 加 FPN-DCN 的网络结构提取图像特征，该结构可以对一些复杂的特征进行提取特征，提高检测精度。

2)采用了一个可变形注意力机制模块对图像和点云数据的特征进行多模态特征融合，降低点云数据和图像数据的偏离，提高模型的精确度。

3)本文设计改进的 3D 目标检测模型在 nuScenes 自动驾驶数据集上表明，该模型具有一个不错的目标检测精度，相较于现有的多模态目标检测方法，有一个不错的提升。

1　相关研究

为了充分利用现有数据以更高精度检测 3D 物体，将投入使用更多的传感器，而如何有效地融合相机图像和测距信息已成为非常关键和热门的研究课题。郑美琳等[3]改进 PointRCNN 网络，通过融入通道注意力串行空间注意力机制，极大提高 PointRCNN 网络在 KITTI 数据集的检测性能，解决三维点云数据的不规则性和密度不均匀问题。王庆林等[4]通过提出 Point-Focus 和 Point-CSPNet 结构，对 PointRCNN 算法进行了有效的改进，提高了对远距离车辆目标的检测精度和整体检测性能。基于鸟瞰图视角的 3D 目标检测包括：在 BEV 视角下融合相机图像和点特征图也是一个不错的选择，BEVFusion[5] 在共享鸟瞰图(BEV)表示空间中统一多模态特征，并通过逐元素串联融合特征图，但整体计算复杂度较高，尤其在数据融合方面。UVTR[6] 保留体素空间而不进行高度压缩，以减轻语义歧义并实现空间连接。跨模态交互可以在点云中获得有效的几何感知表达，并在图像中获得丰富的上下文特征，跨模态的交互对于视角转换，会造成交互的数据误差。BevDepth[7] 针对基于相机的鸟瞰图(BEV)3D 对象检测的深度估计。引入了现实深度估计模块，显著提升了感知精度。

针对图像、点云数据的 3D 目标检测和鸟瞰图的算法的方法各不相同，但是都在目标检测的过程中保持一定的检测精度。原始点云的算法虽然能很好保持点云的原始几何信息，但是对于图像特征和雷达点云数据特征融合，计算量很大。本文算法的改进在一定程度上可以降低融合运算难度，提高目标检测的准确性。

2　模型介绍

2.1　3D 目标检测网络

图 1 为本文的整体结构。如图 1 所示，本文模型主要包括图像特征提取、图像视角转换模块、BEV 编码模块、图像点云特征融合模块、检测头五部分组成。主要流程如下：

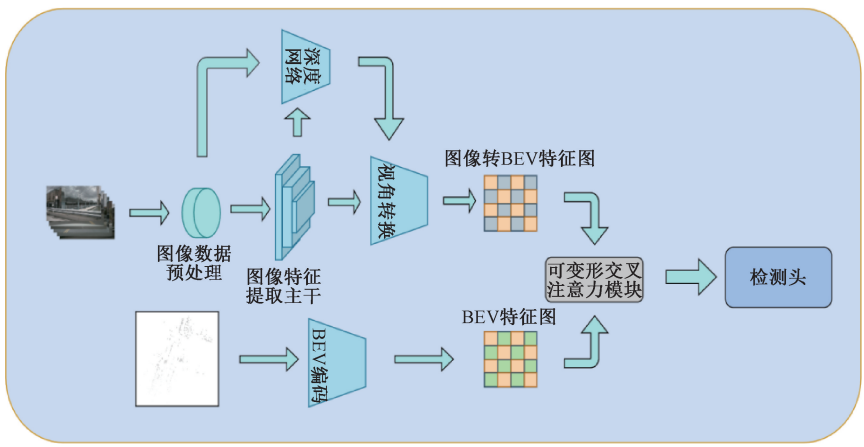


图 1　3D 目标检测流程

Fig. 1　Flowchart of the 3D target detection process

首先，对数据集数据进行预处理，通过生成随机增强参数，调整大小、裁剪、反转等操作对图像进行预处理，将预处理获得的相机参数收集到深度网络，供下一步使用。将处理好的图像转入到 ConvNeXt＋FPN-DCN 的图像特征提取模块，获得图像特征图，将部分图像特征图作为深度网络的输入，最后将另一部分图像特征图和深度网络的

输出通过 LSS(lift, splat, shoot)[8] 视角转换模块输出图像 BEV 特征图。将点云数据集输入到 BEV 编码器中,BEV 编码器由 ConvNeXt 和 FPN-LSS 构成,提取 BEV 特征,通过可变形交叉注意力模块进行多模态融合图像和 BEV 特征,动态降低多模态融合的误差,将输出结果传输到采用 CenterPoint[9] 第一阶段的 3D 物体检测头,实现对 3D 目标的检测。

本章节主要介绍图像特征提取网络和可变形交叉注意力模块。

2.2　图像特征提取网络

本文的图像特征提取网络采用的是 ConvNeXt+FPN-DCN 的网络结构。受到 BEVDet 网络的启发,BEVDet[1] 在图像提取网络使用的是基础的 ResNet+FPN 的网络结构,在现如今的环境中,检测性能较差,计算速度较慢。ConvNeXt 网络通过结合卷积和一些 Transformer 设计方法,在性能方面接近 Transformer 架构,同时保持了 CNN 网络的计算效率。相较于传统的 ResNet[10] 和 Swin Transformer[11],ConvNeXt 能够在大型数据集的图像分类任务中保持良好的准确度,避免了复杂的注意力机制,解决了 Transformer 中计算和内存开销较大的问题,推理速度更快。受到 DCN(可变形卷积)的启发,在 FPN 结构中加入 DCN 的结构能够有效处理多尺度信息,更加灵活地捕捉物体的局部特征,在处理小物体和复杂环境下物体时能够提供更精确的特征提取和更强的适应性。

本节将输入图片设定为 $N\times3\times H\times W$,其中 N 表示相机的个数,3 代表图片的通道数,H 和 W 分别代表图片的高度和宽度。作为图片特征提取模块的 ConvNeXt+FPN-DCN(可变形卷积)的网络结构如图 2 所示,包括 Block1～Block3 共 3 个主干网络结构。

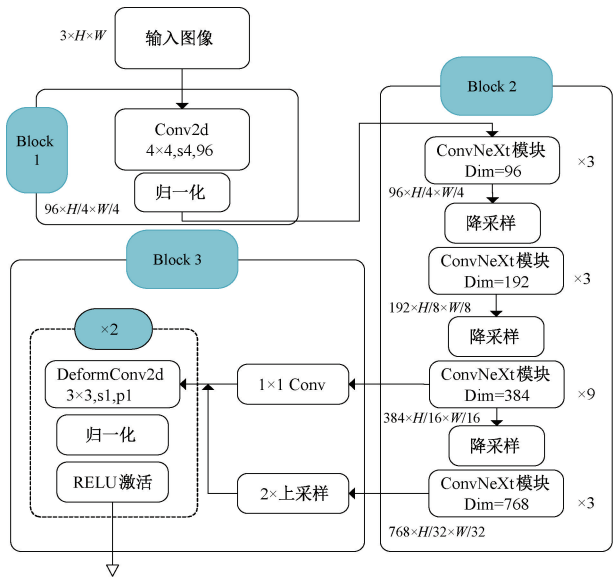


图 2　ConvNeXt+FPN-DCN 的网络结构

Fig. 2　Network architecture of ConvNeXt+FPN-DCN

Block 1 是图像处理的初始阶段,由一个具有 96 维卷积核大小为 4×4、步长为 4 的卷积层和一个归一化层的模块组成。卷积核大小和步长相同的卷积结构使得模型参数减少,整体准确率得到提升。输入图片经过 Block1 的卷积操作后,输出特征图像为 $N\times96\times H/4\times W/4$。

Block 2 是 ConvNeXt 网络的核心模块,包括 4 个阶段,特征提取过程通过卷积模块的顺序级联展开。第 1 个阶段包含 3 个卷积模块,模块的深度为 96;第 2 个阶段包含 1 个下采样层和 3 个卷积模块,每个模块的深度为 192;第 3 个阶段的迭代模式紧接着使用个卷积模块,每个模块的深度为 384;第 4 个阶段包含 1 个下采样层和 3 个卷积模块,每个模块的深度为 768。其中降采样模块由 1 个归一化层和 1 个 2×2 的卷积(步长为 2)组成,经过 Block 2 多层下采样和卷积操作,进一步提取特征后,输出特征图像为 $N\times768\times H/32\times W/32$。其中 Block 2 的 ConvNeXt 模块的具体结构如图 3 所示。

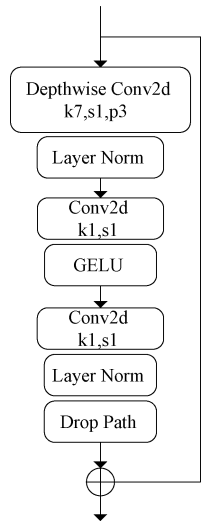


图 3　ConvNeXt 模块

Fig. 3　ConvNeXt module

ConvNeXt 模块采用类似 Transformer 多层感知机制(MLP[12])的逆瓶颈结构,通过两端通道数少、中间通道数多来避免信息流失。该模块包括 3 个主要部分:首先使用一个 7×7 深度卷积核(步长 1)进行卷积操作,并进行归一化;其次,通过 1×1 卷积核(步长 1)和 GELU 激活函数将通道数提升至原来的 4 倍;最后,使用一个 1×1 卷积核将通道数降回初始值,形成逆瓶颈结构。随后,加入可学习的缩放层(Layer Scale)和 DropPath 正则化,进一步提升模型性能和识别准确率。

ConvNeXt 模块采用 GELU[13] 激活函数代替广为使用的 RELU 激活函数,该函数具有 RELU 函数优点(如解决梯度消失问题、计算效率高),又能更好地模拟数据的自然分布特性。GELU 激活函数的数学表达式为:

$$GELU(x)=xP(X\leqslant x)=x\Phi(x)\tag{1}$$

其中，$\Phi(x)$表示标准正态分布下的累积分布函数(CDF)。GELU 函数通过对输入值使用高斯分布的累积概率，实现输入的非线性转换。实际应用中，GELU 激活函数常用的近似公式为：

$$GELU(x)=0.5x\left(1+\tanh\left(\sqrt{\frac{2}{\pi}}(x+0.044\ 715x^3)\right)\right) \tag{2}$$

Block 3 是 FPN＋DCN 的网络结构，首先对特征图像为 $N\times384\times H/16\times W/16$ 进行一个卷积核为 1×1 的横向卷积，调整其通道数不变。对特征图像为 $N\times768\times H/32\times W/32$ 进行 2 倍上采样，与经过 1×1 的横向卷积的特征图像进行拼接，通过两个由 3×3 可变形卷积层(步长 1、填充 1)、归一化层和 ReLU 激活层组成的模块，提高不同尺度的特征融合，提高图像特征提取模块的检测，将进入视角转换模块，将图像特征转换成 BEV 特征，实现 2D 对 3D 的视角转换。

Block 3 使用的可变形卷积层[14]相较于标准的卷积层，更加灵活。标准卷积通过在输入图像上滑动固定的卷积核来计算特征图，可变形卷积层则允许卷积核的采样位置在训练中可以自适应地调整，通过引入一个偏移量(offset)改变卷积核采样的位置。可变形卷积实现如图 4 所示。

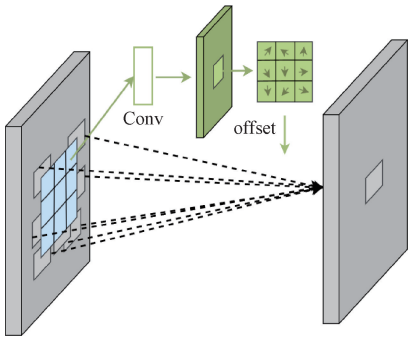


图 4　可变形卷积实现图

Fig. 4　Diagram of deformable convolution implementation

具体的标准卷积公式为：

$$y(p)=\sum_{q\in N_{k(p)}}x(q)\cdot w(q-p) \tag{3}$$

其中，$y(p)$表示输出特征图在位置 p 的值；$x(q)$表示输入特征图在位置 q 的值；$w(q-p)$代表卷积核在相对位置 $q-p$ 上的权重；$N_{k(p)}$表示表示卷积核的空间范围，通常是一个固定大小的窗口(例如，3×3，5×5 等)。在可变形卷积中，通过偏移量 $\Delta p(p,q)$ 来改变卷积核采样的位置，偏移量是通过网络学习得到的。可变形卷积公式为：

$$y(p)=\sum_{q\in N_{k(p)}}x(q+\Delta p(p,q))\cdot w(q-p) \tag{4}$$

可变形卷积可以将图像中变形最显著的地方通过卷积操作操作来获取偏移量，这种方法可以有效捕获变形前后的变化区域。在特征提取模块中引入可变形卷积，输出特征图的卷积核与偏移量同时学习图像变化前后的特征信息，通过双层的可变形卷积模块，提高模型对图像不同大小、形状的图像前后变化的物体的检测精度。

2.3　可变形交叉注意机制模块

在点云 BEV 特征和图像 BEV 特征融合时，雷达点云会受到方位角误差的影响，导致雷达传传感器检测到物体之外的其他点。由 BEV 编码器生成的 BEV 特征可能将原来物体对应相邻的 BEV 网格所占据，导致图像 BEV 特征难以对齐雷达 BEV 特征。

为了解决这个问题，本文使用可变形交叉注意机制[15]动态对齐图像 BEV 特征和雷达 BEV 特征。在实际过程中，雷达点云数据可能会与真实位置偏移一段距离，采用可变形交叉注意力机制来准确捕捉和校正雷达点云位置的偏离，创新性的可变形交叉注意力模型如图 5 所示。

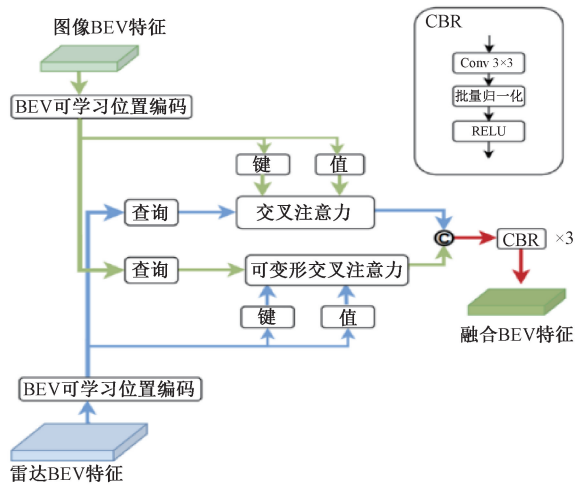


图 5　可变形交叉注意力模块图

Fig. 5　Diagram of the deformable cross-attention module

可变形交叉注意机制能有效降低交叉注意的计算复杂度，从 $O(H^2W^2C)$ 降低到 $O(HWC^2K)$，其中 H 和 W 表示 BEV 特征的高度和宽度，C 表示 BEV 特征通道，K 表示变形交叉注意中参考点的数目。给定分别由 $F_c\in R^{Cc\times H\times W}$、$F_r\in R^{Cr\times H\times W}$ 表示的摄像机和雷达 BEV 特征，F_r 添加可学习的位置嵌入。然后，F_c 被转换为查询 z_{qc} 和参考点 p_{qc}，F_r 被视为键和值。

接下来，通过以下公式计算多头可变形交叉注意力：

$$DeformAttn(z_{qc},p_{qc},F_r)=$$
$$\sum_{h=1}^{H}w_h\left[\sum_{k=1}^{K}A_{hqk}\cdot\boldsymbol{W}'_hF_r(p_{qc}+\Delta p_{hqk})\right] \tag{5}$$

其中，h 为注意力头部的索引，k 为采样的键的索引，K 表示采样的键的总数，Δp_{hqk} 表示采样偏移，A_{hqk} 表示 z_{qc} 和 F_r 计算的注意力权重，w_h 表示融合多头注意力的输出权重值，$\boldsymbol{W}'_h$ 为第 h 个头部的值投影矩阵。对于图像 F_c

部分采用基础的交叉注意力[16]就可以实现捕获和校正,具体的多头交叉注意力机制公式如下:

$$CrossAttn(z_{qr},F_c)=\sum_{h=1}^{H}w_h\left[\sum_{k=1}^{K}A_{hqk}\cdot W'_hF_c\right] \quad (6)$$

最后,本文中的可变形交叉注意模块可以公式化为:

$$\begin{cases}F_c\leftarrow CrossAttn(z_{qr},F_c)\\F_r\leftarrow DeformAttn(z_{qc},p_{qc},F_r)\end{cases} \quad (7)$$

通过交叉注意机制对齐雷达和相机 BEV 特征之后,利用通道和空间融合层来聚合多模态 BEV 特征。具体地,将两个 BEV 特征连接为 $F_{multi}=[F_c,F_r]$。然后,将 F_{multi} 发送到 3 个 CBR 模块,进一步融合多模态特征。CBR 块依次由 Conv3×3、批量归一化层和 ReLU 激活函数组成。

3　实验结果与分析

3.1　数据集与评价指标

本文使用的数据集是公开的自动驾驶数据集 nuScenes[17]进行了网络的泛化性能验证。nuScenes 基准包括来自 6 个摄像机的 1 000 个场景。它是基于视觉的 3D 对象检测和 BEV 语义分割的最新流行基准。这些场景被正式分成 700/150/150 场景,用于训练/验证/测试。多达 1.4 M 带注释的 3D 边界框,可用于 10 个类别:汽车、卡车、公共汽车、拖车、建筑车辆、行人、摩托、自行车、障碍物和交通锥。在中心点之后,本研究定义了地平面中 51.2 m 内的感兴趣区域(ROI),默认情况下分辨率(即中心点中的体素大小)为 0.8 m。

进行 3D 对象检测,nuScenes 数据集提供了一组评估指标,包括平均平均精度(mAP)和 5 个真阳性(TP)指标:ATE、ASE、AOE、AVE 和 AAE,它们分别测量位置误差、尺度、方向、速度和外观特征。整体性能通过 nuScenes 检测分数(NDS)来衡量,该分数整合了所有错误类型。

$$NDS=\frac{1}{10}\left[5\times mAP+\sum_{mTP\in TP}(1-\min(1,mTP))\right] \quad (8)$$

本文使用 AdamW[18]优化器训练模型,其中梯度以学习率 2×10^{-4} 利用。对于 ConvNext+FPN-DCN 的图像特征提取器,应用了循环策略,该策略在前 40%的时间表中将学习速率从 2×10^{-4} 线性增加到 10^{-3},并在剩余时期将学习速率从 10^{-3} 线性降低到 0。默认情况下,总计划在 20 个轮次内终止。

3.2　实验环境

本文提出的算法使用 PyTorch 框架实现,运行在 GPU 为 NVIDIA RTX 4090、CPU 为 Intel Xeon Platinum 8260、操作系统为 Ubuntu20.04 的服务器上。

输入数据为双目彩色相机和激光雷达点云数据。nuScenes 数据集中的图像分辨率为 1 600×900,通过预处理,将分辨率保持为 704×256。经过 ConvNeXt+FPN-DCN 提取特征,输出多尺度特征图,大小分别为 $H/4\times W/4$,$H/8\times W/8$,$H/16\times W/16$,$H/32\times W/32$。

3.3　实验结果与分析

本实验使用的模型在实验中处理图像和激光雷达的 BEV 特征,表 1 展示了该设置下在 NuScenes 验证集上的结果"C"和"R"分别代表摄像头和雷达。在没有使用其他数据增强和时间特征增强的情况下,所提模型优于目前一些主流相机-激光雷达多模态融合方案。相比于 CRN,所提模型在 mAP 和 NDS 上分别提升了 8.8%和 6.8%。CRN 模型中所用的图像特征提取网络是 ConvNeXt 网络,与本文的模型相比,模型中的可变形交叉注意力机制提取到了更丰富的图像语义信息,并对图像 BEV 特征和雷达点云 BEV 特征进行了偏差校正,对整体的检测效果进行了显著增强。从 mAVE 的指数来看,本文的模型相较于其他的模型,能够有效模型对动态物体的速度预测较为准确。

表 1　nuScenes 验证集上 3D 物体检测结果的比较

Table 1　Comparison of 3D object detection results on the nuScenes validation Set

方法	输入	尺寸	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
BEVDet	C	256×704	39.2	31.2	0.691	0.272	0.523	0.909	0.247
SparseBEV[19]	C	256×704	54.5	43.2	0.606	0.274	**0.387**	0.251	0.186
BEVDepth	C	256×704	47.5	35.1	0.639	0.267	0.479	0.428	0.198
RCBEV4d[20]	C+R	256×704	49.7	38.1	0.526	0.272	0.445	0.465	0.185
CRN[21]	C+R	256×704	54.3	44.8	0.518	0.283	0.552	0.279	**0.180**
Ours	C+R	256×704	61.1	53.6	**0.489**	**0.259**	0.404	**0.223**	0.192

表 2 展示了本文模型在 nuScenes 测试集上与其他以往最先进的 BEV 目标检测模型的对比结果。

在这一测试集中,除了相机和雷达的模式外,还列出了仅使用摄像头的模式。通过深入分析与其他模型的差异,可以看出多模态的数据融合可以极大的提高目标检测的精度。

本文模型相较于 SOLOFusion,在 NDS 上实现了 3 点的提升,在 mAP 上提升了 2.1 点,显著增强了整体的检测性能。在与 BEVDet 模型的对比中,可以看出,本文模型在很大程度上超越了原模型,在 NDS 上提升 8.1%,mAP

表 2　nuScenes 测试集上 3D 物体检测结果的比较

Table 2　Comparison of 3D object detection results on the nuScenes Test Set

方法	输入	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
BEVFormer[22]	C	56.9	48.1	0.582	0.256	0.375	0.378	0.126
BEVDet	C	56.8	45.1	0.511	**0.241**	0.386	0.301	0.121
PETRv2[23]	C	58.2	59.2	0.561	0.242	**0.361**	0.343	0.120
SOLOFusion[24]	C	61.9	54.0	0.453	0.257	0.376	0.276	0.148
CRN	C+R	62.4	57.5	0.416	0.264	0.456	0.365	0.130
MVFusion[25]	C+R	51.7	45.3	0.569	0.246	0.379	0.781	0.128
CRAFT[26]	C+R	52.3	41.1	0.467	0.268	0.455	0.519	0.114
Ours	C+R	**64.9**	**56.1**	**0.382**	0.225	0.353	**0.246**	**0.112**

上也有显著提高。相较于其他模型如 BEVFormer、PETRv2 等，本文模型在定位、尺度和加速度预测方面更强。

通过深入分析，发现本文模型的创新之处在于采用了 ConvNeXt+FPN-DCN 的网络结构和可变形交叉注意力机制。这种机制能够自适应地调整对不同区域的关注重点，通过对不同空间位置的跨域注意力进行变形处理，更有效地捕捉空间和上下文关系。这不仅提高了模型对复杂场景的理解和精准检测能力，而且能够根据输入的不同特点，智能调整计算资源和注意力分配策略。因此，本文模型在面对多样化的数据集和任务时，都能保持出色的性能，这一点在 NDS 和 mAP 这两个综合指标上得到了体现。尽管与某些模型存在微小差距，本文模型在大多数场景中都能提供稳定且优异的检测能力，显示出其在各种环境下的适应性和鲁棒性。

3.4　消融实验

在 Nuscenes 验证集上对本文方法进行了详细的消融实验，具体消融实验结果如表 3 所示。首先，对本文模型构成进行了消融实验，如表 3 所示，在以基线模型为 BEVDet4D 的基础上，依次将特征提取网络更换成 ConvNeXt+FPN-DCN 的网络，添加可变形交叉注意力模块，实验结果表明，通过将 ConvNeXt 与 FPN-DCN 结合，NDS 提升至 0.583，mAP 提升至 0.507，表明加入 ConvNeXt 网络和 FPN(特征金字塔网络)结合 DCN(可变形卷积网络)能够显著提高模型的性能，尤其是在更好地捕捉复杂特征方面。进一步加入可变形交叉注意力机制，NDS 提升至 0.611，mAP 提升至 0.536。这个改进显示了可变形交叉注意力机制对提升模型性能的作用，尤其是在处理空间特征和校正多模态融合偏差方面的优势。

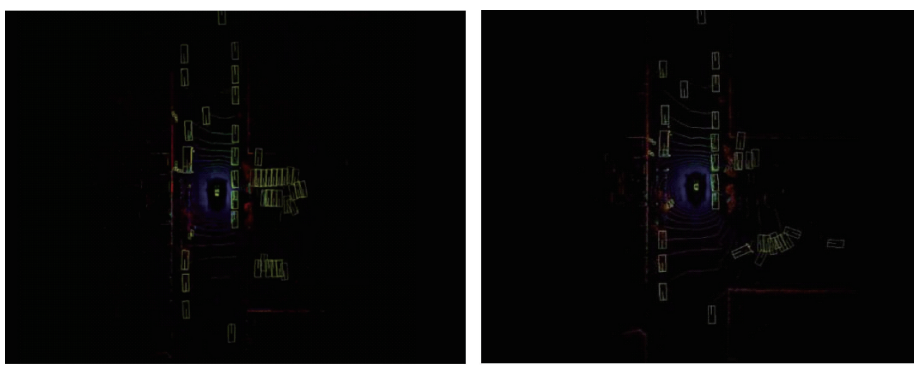
表 3　模型消融实验

Table 3　Model ablation experiment

消融实验	NDS	mAP
BEVDet4D	0.524	0.451
+ConvNeXt+FPN-DCN	0.583	0.507
+可变形交叉注意力机制	0.611	0.536

3.5　实验可视化

在 nuScenes 数据集上验证所提算法，并通过可视化窗口展现 3D 目标检测结果。其中，黄色矩形框为检测得到的车辆目标，红色点和蓝色线为包含在检测框内的点云和车辆周围的体素线，如图 6 所示。左侧为 BEVDet 可视化下的鸟瞰图，右侧为本文模型可视化下的鸟瞰图，可以看出本文模型相较于 BEVDet，目标检测上更加精确，误差更小，充分证明了本文模型改进的可行性。



(a) BEVDet可视化图　(b) 本文模型可视化图

(a) BEVDet visualization　(b) The visualization of the model presented in this paper

图 6　可视化图

Fig. 6　Visual drawing

4 结　论

本文提出了一种图像和雷达点云多模态数据融合 3D 目标检测改进算法。通过 ConvNeXt＋FPN-DCN 的网络提高模型对复杂环境下图像特征提取的能力,并通过可变形交叉注意力模块对不同模态特征进行交互,同时保留两种特征信息,能够动态地调整不同模态在融合时候的位置偏差,自适应地学习不同特征的融合范围及分布比例,减小了特征对齐误差。在 nuScenes 数据集上的对比实验和消融结果表明,相较于基线网络,本文算法在检测精度上有了显著的提高,证明了改进方法的有效性。下一步会进一步优化算法,提高模型在更多环境中的检测性能。此外,探索实际环境中行人、车辆的轨迹预测也是本文之后的重要研究方向。

参考文献

[1] HUANG J J, HUANG G, ZHU ZH, et al. Bevdet: High-performance multi-camera 3D object detection in bird-eye-view[J]. ArXiv preprint arXiv:2112.11790, 2021.

[2] LIU ZH, MAO H Z, WU CH Y, et al. A convnet for the 2020s[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 11976-11986.

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

[4] 王庆林,李辉,谢礼志,等.基于激光雷达点云的车辆目标检测算法改进研究[J].电子测量技术,2023,46(1):120-126.
WANG Q L, LI H, XIE L ZH, et al. Research on algorithm improvement for vehicle object detection based on LiDAR point cloud [J]. Electronic Measurement Technology, 2023, 46(1): 120-126.

[5] LIU ZH J, TANG H T, AMINI A, et al. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation [C]. 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023: 2774-2781.

[6] LI Y W, CHEN Y L, QI X J, et al. Unifying voxel-based representation with transformer for 3D object detection [J]. Advances in Neural Information Processing Systems, 2022, 35: 18442-18455.

[7] LI Y H, GE ZH, YU G Y, et al. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection [C]. AAAI Conference on Artificial Intelligence. 2023, 37(2): 1477-1485.

[8] PHILION J, FIDLER S. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D[C]. Computer Vision-ECCV 2020: 16th European Conference, 2020: 194-210.

[9] YIN T, ZHOU X, KRAHENBUHL P. Center-based 3D object detection and tracking [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 11784-11793.

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

[11] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. IEEE/CVF International Conference on Computer Vision, 2021: 10012-10022.

[12] TAUD H, MAS J F. Multilayer perceptron (MLP)[J]. Geomatic Approaches for Modeling Land Change Scenarios, 2018: 451-455.

[13] HENDRYCKS D, IMPEL K. Gaussian error linear units (gelus) [J]. ArXiv preprint arXiv: 1606.08415, 2016.

[14] DAI J F, QI H ZH, XIONG Y W, et al. Deformable convolutional networks [C]. IEEE International Conference on Computer Vision, 2017: 764-773.

[15] ZHU X ZH, SU W J, LU L W, et al. Deformable DETR: Deformable transformers for end-to-end object detection[J]. ArXiv preprint arXiv:2010.04159, 2020.

[16] 金宇锋,陶重犇.基于 Transformer 的融合信息增强 3D 目标检测算法[J].仪器仪表学报,2023,44(12):297-306.
JIN Y F, TAO ZH B. Transformer-based fusion information enhanced 3D object detection algorithm[J]. Chinese Journal of Scientific Instrument, 2023, 44(12): 297-306.

[17] CAESAR H, BANKITI V, LANG A H, et al. nuscenes: A multimodal dataset for auto-nomous driving [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 11621-11631.

[18] KINGMA D P. Adam: A method for stoc-hastic optimization[J]. ArXiv preprint arXiv:1412.6980, 2014.

[19] LIU H S, TENG Y, LU T, et al. Sparsebev: High-performance sparse 3d object detection from multi-camera videos [C]. IEEE/CVF International Conference on Computer Vision, 2023: 18580-18590.

[20] ZHOU T H, CHEN J J, SHI Y N, et al. Bridging the view disparity between radar and camera features

for multi-modal fusion 3D object detection[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(2): 1523-1535.

[21] KIM Y, SHIN J, KIM S, et al. CRN: Camera radar net for accurate, robust, efficient 3D perception[C]. IEEE/CVF International Conference on Computer Vision, 2023: 17615-17626.

[22] LI ZH Q, WANG W H, LI H Y, et al. BEVFormer: Learning bird′s-eye-view representation from multi-camera images via spatiotemporal transformers[C]. European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 1-18.

[23] LIU Y F, YAN J J, JIA F, et al. PETRv2: A unified framework for 3D perception from multi-camera images[C]. IEEE/CVF International Conference on Computer Vision, 2023: 3262-3272.

[24] PARK J, XU CH F, YANG SH J, et al. Time will tell: New outlooks and a baseline for temporal multi-view 3D object detection[J]. ArXiv preprint arXiv: 2210.02443, 2022.

[25] WU Z ZH, CHEN G L, GAN Y ZH, et al. MVFusion: Multi-view 3D object detection with semantic-aligned radar and camera fusion[C]. 2023 IEEE International Conference on Robotics and Automation(ICRA). IEEE, 2023: 2766-2773.

[26] KIM Y, KIM S, CHOI J W, et al. CRAFT: Camera-radar 3D object detection with spatio-contextual fusion transformer[C]. AAAI Conference on Artificial Intelligence, 2023, 37(1): 1160-1168.

作者简介

周鹏，硕士研究生，主要研究方向为无人驾驶目标检测、无人驾驶环境感知等。

E-mail：202312490378@nuist.edu.cn

宋志强（通信作者），教授，主要研究方向为智能系统控制。

E-mail：1750545330@qq.com

胡凯，副教授，主要研究方向为多智能体学习、水下图像增强等。

E-mail：nuistpanda@163.com

宋利鹏，硕士研究生，主要研究方向为多智能体系统协作与控制等。

E-mail：202312490325@nuist.edu.cn

李明阳，硕士研究生，主要研究方向为多智能体系统协作与控制等。

E-mail：202212490507@nuist.edu.cn