文章编号:2096-1472(2024)03-0067-07
DOI:10.19644/j.cnki.issn2096-1472.2024.003.014

多尺度特征结合注意力机制的室内3D点云目标检测

顾方宇,胡海洋

(杭州电子科技大学计算机学院,浙江 杭州 310018)
✉ gfy2345@163.com; huhaiyang@hdu.edu.cn

摘 要:为了实现三维点云在室内和工业环境中的实际应用,文章改进了传统的目标检测转换器(Detection Transformer,DeTR)神经网络,并提出了一种基于分层抽象的多层点云特征提取方法;同时,设计了曲面特征提取模块对三维点云进行预处理,增强了点云的附加特征。在公开数据集ScanNet V2和工业室内数据集上对本文方法进行实验验证和评估,该方法在ScanNet V2上的mAP@0.5准确率超过最先进的模型(State-of-the-Art,SOTA)CAGroup3d,达到76.0%;在ScanNet V2上的mAP@0.25准确率超过最先进的模型CAGroup3d,达到62.2%,消融实验进一步验证了所述方法的准确性和高效性。

关键词:三维点云;目标检测;工业环境;Transformer
中图分类号:TP389.1 **文献标志码**:A

Indoor 3D Point Cloud Object Detection with Multi-scale Features and Attention Mechanism

GU Fangyu, HU Haiyang

(*School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou* 310018, *China*)
✉ gfy2345@163.com; huhaiyang@hdu.edu.cn

Abstract: To facilitate the practical application of 3D point cloud in indoor and industrial environments, this paper proposes a multi-layer point cloud feature extraction method based on hierarchical abstraction with the improvement of traditional object Detection Transformer (DeTR) neural network. Additionally, a surface representation module for preprocessing 3D point cloud is designed to enhance the additional features of the 3D point cloud. Experimental validation and assessment of the proposed method are conducted on the public dataset ScanNet V2 and an industrial indoor dataset. Experiment results show that the mAP@0.5 accuracy of the proposed method exceeds the State-of-the-Art (SOTA) model CAGroup3d, reaching 76.0%; the mAP@0.25 accuracy exceeds the SOTA model CAGroup3d, reaching 62.2%. The ablation experiment further validates the accuracy and efficiency of the method.

Key words: 3D point cloud; object detection; industrial environment; Transformer

0 引言(Introduction)

制造业的升级,本质上涉及核心技术的提升、生产模式的改进以及应用场景的拓宽,所有这些都依赖于前沿技术的推动和支持,而工业视觉中的智能传感器包括机器视觉技术在内的系统,在其中起着至关重要的作用。

在三维目标检测领域,当前的目标检测算法主要集中在从三维点云到二维特征的转换上,这主要依赖于点云预处理的方法。本文提出一种新的策略,专注于使用深度学习处理三维点云数据,将三维点云目标检测和室内机器人场景融合,实现三维视觉在室内和工业场景中的真实应用。

本文的主要贡献总结如下:(1)对传统的DeTR目标检测网络进行改进,提出了一个基于分层抽象的多层点云特征提取方法,该方法不仅能通过多层网络获取多尺度点云集,而且强化了在三维点云中进行目标检测的能力;(2)采用了曲面表示

收稿日期:2023-07-03
基金项目:浙江省自然科学基金项目(LY22F020021);浙江省重点研发计划"领雁"项目(2023C01145);国家自然科学基金项目(61802095,61572162)

模块对三维点云进行预处理，增强了三维点云的附加特征，通过从点到面的方式增强了点云的附加信息；(3)引入了基于特征金字塔融合特征的转换器神经网络(Transformer)架构，在多尺度点云集的特征图上进行多尺度注意力操作，有助于增强特征提取能力，并提高目标检测的准确率。

1 相关工作(Related work)

目标检测是使用计算机视觉技术，从不同复杂程度的背景中识别运动物体，并分离背景完成目标标记的一项研究内容。近年来，随着区域卷积神经网络(Regions with Convolutional Neural Networks，RCNN)算法[1]的提出，深度学习逐渐被应用到目标检测领域。YOLO算法[2]在已有目标检测算法结构上做出了一定的创新，不仅能从端到端直接输出目标信息，同时拥有较快的反应速度，目前依旧是工业界目标检测的首选。

三维目标检测是环境感知系统中的重要技术之一，它在自动驾驶、机器人等领域发挥着重要的作用[3]。深度学习，特别是卷积神经网络(Convolutional Neural Networks，CNN)，提供了一种可以推动三维目标检测精确度提升的方法，并且随着深度学习的迅速发展，三维目标检测的性能也得到了显著的提升。

随着深度传感器和三维激光扫描仪的普及运用，基于点云(Point Cloud)的三维目标检测得到广泛关注。目前，主流的三维数据表示方法主要有深度图、三角网格、体素和点云，其中点云是最简单的一种三维数据表示方法，具有获取简单、易于存储、可视性强、结构描述精细等优点，而且能够方便地与深度图、体素等其他数据格式相互转换，已成为三维重建、三维目标检测、即时定位与地图构建(Simultaneous Localization and Mapping，SLAM)等研究领域最基本的数据格式[4-6]。在处理点云数据的领域，PointNet[7](一种用于点云数据处理的深度学习架构)被提出直接处理点云，其使用一系列全连接层和非线性激活对点云进行处理，从而聚合所有点的特征。同时，这种方法保证了在无序点云集中，对在任意点次序中的点，均能保证拟合函数对点的输入顺序不敏感。针对PointNet中不能处理点与点间局部关系的局限性，PointNet++[8]被提出，其在原有网络的基础上使用了分层抽象(Set-Abstraction)的方法，使得模型能够捕获更丰富的局部结构信息。

使用PointNet作为骨干模型叠加检测头，在两阶段的方法上实现基于点云领域的三维目标检测是一种比较常见的研究点云目标检测算法的思路。PointRCNN[9]在PointNet的基础上，首先进行区域提议，其次对提议的区域进行分类和边界框回归。VoteNet[10]使用基于霍夫投票(Hough Voting)的方法，在处理稀疏和无序的点云数据时具有较好的性能，此外它的投票机制也使得模型对噪声和遮挡具有很好的鲁棒性。VENet[11]基于VoteNet，将基于注意力的多层感知机(Attention-based Multi-Layer Perceptron，AMLP)用于特征提取，从而提升了模型在特定数据集上的分类准确度和鲁棒性。

受遮挡、光照反射、表面材质的透明度以及传感器分辨率和视角等因素的限制，使用单个点云相机采集到的点云数据往往是不完整的。基于信息补全方法，对点云进行信息扩展和补全也是一个在点云信息处理领域常用的方法。在信息补全方面，PF-Net[12]被提出，它是基于几何结构预测的精确和保真度的点云补全方法，其中多分辨率编码器(Multi-Resolution Encoder)使用了一个联合多层感知机(CMLP)，从低分辨率的点云中提取多尺度特征；此外，点金字塔解码器(Point-Pyramid Decoder)用于预测不同深度层次的点特征，并传播整体几何信息到最后被补全点的骨架中心。

2 模型设计(Design of the model)

本文提出的基于表面增强尾骨和多尺度可变形DeTR(Multi-Scaled-Deformable-DeTR，MSD-DeTR)的三维多尺度点云目标检测方法，其主体流程结构如图1所示，该结构主要由以下几个部分组成：(1)基于曲面表示的曲面特征提取模块；(2)基于注意力增强多层感知机的特征提取模块；(3)基于多尺度特征金字塔和Transformer的注意力主干网络(DeTR)。

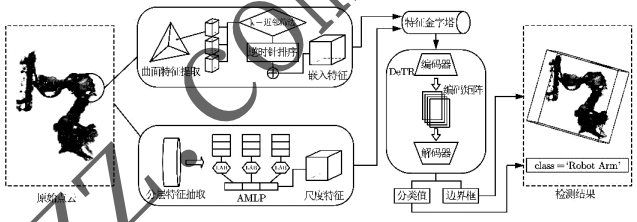


图1 多尺度可变形DeTR网络主体流程结构

Fig. 1 The process structure of multi-scaled deformable DeTR network

2.1 曲面特征提取

点云通常在三维空间中呈稀疏分布，在处理点云数据时，局部形状的表达至关重要。以往的研究通过使用额外的元素或通过不同的转换间接地从形状中学习，然而这些操作可能只能提供一些表示点云局部集合的提示，而不能明确地反映局部形状。当使用额外的信息或转换表示或处理点云数据时，可能会导致计算量显著增加，而对点云表示的贡献却很小。在某些情况下，这种做法甚至可能导致几何信息的丢失。

本文依据曲面表示的方法表示局部几何结构，并提出了一种基于伞状表面的额外信息的嵌入方式，不仅能够保留点云的几何细节，同时能够以极小的额外的计算量表述更复杂的局部形状。下面介绍一个三角信息增强的方法。

对于二维曲线上的一个点(x_i, y_i)，可以使用点法式表示过点的切线，如公式(1)所示：

$$\begin{aligned} &a_i(x-x_i)+b_i(y-y_i)=0\Rightarrow \\ &a_ix+b_iy-(a_ix_i+b_iy_i)=0 \end{aligned} \tag{1}$$

在三维曲面上，对于点集$\boldsymbol{P}=\{p_1, p_2, \cdots, p_n\}\subseteq\mathbb{R}^{(N\times 3)}$，给定点$\boldsymbol{p}_i=(x_i, y_i, z_i)$及切向量$\boldsymbol{v}_i=(a_i, b_i, c_i)$，可以将上述二维曲线的点法式的定义进行扩充，如公式(2)所示：

$$\begin{aligned} &a_i(x-x_i)+b_i(y-y_i)+c_i(z-z_i)=0\Rightarrow \\ &a_ix+b_iy+c_iz-(a_ix_i+b_iy_i+c_iz_i)=0 \end{aligned} \tag{2}$$

从点法式的定义中延伸,定义对应表面位置 $s_i=a_ix_i+b_iy_i+c_iz_i$,范围为$[-\sqrt{3}r,\sqrt{3}r]$,其中 r 表示刚好覆盖点集的立方体的边长。举例来说,如果将在$[-1,1]$范围内的已标准化点云集作为处理目标,那么这里的 $r=1$。表面位置 s_i 也可以表示原点和表面之间的有向距离。在点云空间内使用三角信息额外提取方法提取特征向量的例子如图2所示。

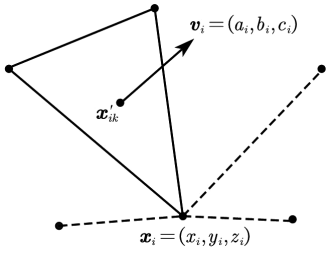


图2　三角信息的额外提取

Fig. 2 Additional extraction of triangular information

考虑到切向量 $\boldsymbol{v}_i$ 只有数值,意味着它可以指向法向量的两侧(内部或外部)。对于这个问题,保持 a_i 为正,并通过实例级随机逆运算,以50%的概率增强法线,可以得到一组三角额外信息向量 $\boldsymbol{T}=\{\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_n\}\subseteq\mathbb{R}^{(N\times4)}$,其中 $\boldsymbol{t}_i=(a_i,b_i,c_i,s_i)$ 是三角额外信息的定义。

上文讨论了有关三角额外信息向量的嵌入方法,可以观察到在伞状结构中,其中心点和周围的点可以连接成一组三角形平面。因此,对于给定点 x_i,设点的邻接三角形数量为 K,计算其邻接三角质心矩阵,定义为 $\boldsymbol{X}'_i=\{\boldsymbol{x}'_{i1},\boldsymbol{x}'_{i2},\cdots,\boldsymbol{x}'_{iK}\}\subseteq\mathbb{R}^{(K\times3)}$,使用上述三角额外信息嵌入方法,得到被嵌入额外信息的矩阵 $\boldsymbol{T}_i=\{\boldsymbol{t}_{i1},\boldsymbol{t}_{i2},\cdots,\boldsymbol{t}_{iK}\}\subseteq\mathbb{R}^{(K\times4)}$。伞状信息向量 $\boldsymbol{u}_i$ 为邻接三角质心特征和三角额外信息特征的聚合,定义为公式(3):

$$\boldsymbol{u}_i=A(\{T([\boldsymbol{x}'_{ij},\boldsymbol{t}_{ij}]),\forall j\in\{1,\cdots,K\}\}) \tag{3}$$

其中:$A(\cdot)$是聚合函数(求和函数),$T(\cdot)$是转换函数,邻接三角质心 $\boldsymbol{x}'_{ij}$ 是其相对中心点 x_i 的归一化坐标。为了计算三角额外信息 $\boldsymbol{t}_i$,需要在 xy 平面上从逆时针构建邻接三角形,所以伞状邻域中的三角形个数正好为 K。为了保证法线方向的一致性,需要通过逆时针交叉乘积计算这些法线。

对于转换函数 $T(\cdot)$,它是一个使用可学习参数的函数(线性回归器和非线性拟合器的组合),在损失的反向传播中,它能保证伞状信息随着训练逐步正确拟合。相较于使用预定的超参数进行调整,使用可学习的参数不仅能使拟合效果最优化,还能减少模型微调中的工作量。

2.2　注意力改进的分层提取模块

尽管基于分层特征提取的深层网络在点云数据处理中展现出很好的性能,但仍有一些限制和缺点,其中之一便是特征抽取能力有限。由于分层特征提取主要依赖全连接神经网络,而全连接神经网络在处理图像或序列数据时的表现弱于包括卷积神经网络在内的一部分神经网络。这使得基于分层特征提取的网络在处理复杂结构的点云数据时,可能无法提取到足够的特征信息,从而导致诸如识别准确率下降等问题。

本文方法在分层特征提取的基础上,提出了基于自注意力增强的多层感知机嵌入,这种注意力机制集成的分层特征提取模块,可以强化分层提取算法中种子点的特征描述,从而更好地提取点云数据中的特征。在分层特征提取模块(图3)中,总共包含3层采样分组模块,其中每个分层采样模块包含采样分组操作。

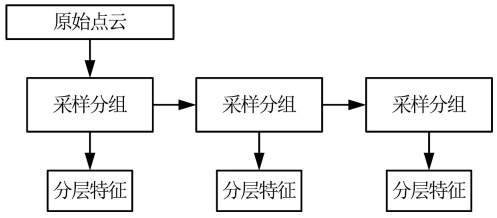


图3　分层特征提取模块

Fig. 3 Hierarchical feature extraction module

一个在工业机器人场景中所摄入的点云图往往有非常多的点,这会造成计算量过大而限制模型使用。本文的解决方案是从所有的点云数据中采样 k 个点,并且保证这 k 个点拥有足够的信息。采样操作中,主要使用最远点采样算法(Farthest Point Sampling, FPS)保证采样获取的点数量在可控范围内,主要操作步骤如下:(1)随机选择一个初始的已选择的采样点;(2)计算每个点与已选择采样点集之间的距离;(3)持续循环迭代,并将距离最远的点加入已选择采样点集,直到点集的数量符合要求。

采样分组操作如图4所示,其主要负责将采样操作所得到的点集分组特征进行聚合。假设在采样操作中得到了包含 M 个点的点集 $\boldsymbol{P}=\{\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_M\}$,$\boldsymbol{p}_i\in\mathbb{R}^d$,并使用一个线性全连接层对上述点集进行优化整合,可以得到邻域点集的分层特征 G,其中每个分层特征的子集 $\boldsymbol{G}_i=\{\boldsymbol{g}_{i1},\boldsymbol{g}_{i2},\cdots,\boldsymbol{g}_{ik}\}\in\boldsymbol{G}$,$\boldsymbol{g}_{ij}\in\mathbb{R}^d$ 都包含了 $\boldsymbol{p}_i$ 的 k 个最近邻的特征向量。

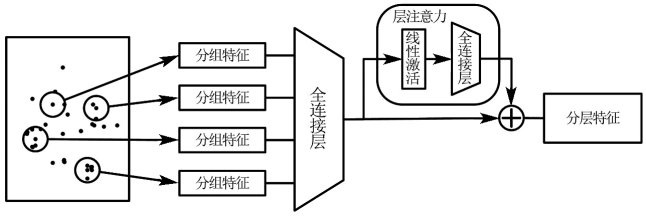


图4　采样分组操作

Fig. 4 Sampling and grouping operations

在繁杂的分层特征中,传统的基于PointNet++的特征提取模块具有一定的局限性。本文所述模型使用自注意力增强的多层感知机(Attention-based MLP, AMLP)对邻域点进行特征提取处理,可以达到提升分层特征提取能力和提升模型性能的目的。具体来说,对于每个分层特征子集的单条特征 $\boldsymbol{g}_{ij}$,从每个层池化特征,生成$(\boldsymbol{c}_1,\cdots,\boldsymbol{c}_L)$,其中 L 是多层感知机中感知器的层数。

与直接连接池化特征不同，自注意力多层感知机的设计在每一层都插入了一个级别注意力块(Level-Attention-Block，LAB)。在每个LAB中，一个池化特征向量 $\boldsymbol{c}_l \in \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_L\}$ 首先被馈入两个全连接(Fully-Connect，FC)层，输出大小为 $C/4$ 和 C。其中，第一个FC层使用线性激活单元(Rectified Linear Unit，ReLU)作为激活函数。Sigmoid(·)用于将输出权重归一化到(0,1)的范围，如公式(4)所示：

$$\text{Sigmoid}(x)=\frac{1}{1+\mathrm{e}^{-x}} \tag{4}$$

$\boldsymbol{c}_l$ 乘以学习到的权重 $\boldsymbol{W}_l$，并添加到自己的结果上，如公式(5)所示：

$$\boldsymbol{c}_l=\boldsymbol{c}_l+\boldsymbol{W}_l\times\boldsymbol{c}_l \tag{5}$$

来自所有层的特征向量被组合起来，形成组合特征向量。通过LAB中的第二个FC层输出，如公式(6)所示：

$$\boldsymbol{C}=FC(Concat(\boldsymbol{c}_1,\cdots,\boldsymbol{c}_l)) \tag{6}$$

其中：$Concat(\cdot)$ 代表拼接函数，它将多条特征向量拼合到一个向量中。$FC(\cdot)$ 代表第二个FC层中使用的全连接函数，它是多个线性连接层的组合，能够将向量整合到合适的尺寸中。通过上述方法，模型能够学习到被自注意力增强的点云特征，从而使模型能够在处理复杂点云数据时，更好地把握重要的特征信息，从而提高模型的表现。

2.3 多尺度特征融合的主干网络

两阶段的目标检测是目前常见的目标检测方法，它一般包括两个阶段：第一个阶段是生成候选区域的提议，生成方法是区域提议；第二个阶段是提议分类和提议回归，一般是对第一阶段的提议候选区域进行分类，并对提议生成的边界框进行回归。虽然两个阶段的目标检测方法在精度上均有很好的表现，但是它们也存在缺点，例如计算复杂，两个阶段的目标检测器通常需要在多个区域提议上运行，增加了计算负担。通常，两个阶段的目标检测方法需要在训练时使用复杂的采样策略，使训练过程过于烦琐。

多尺度特征融合的主干网络，主要使用特征金字塔模块融合对齐上述曲面特征提取的特征和改进分层提取尾骨提取的特征，并对融合的特征使用基于Transformer的编码器和解码器，同时使用端到端的检测骨架进行目标检测，从而简化计算能力。

特征金字塔是一种常见的特征对齐技术，它是通过对输入特征进行多次下采样构建的。举例来说，二维目标检测方法使用特征金字塔同时检测不同尺度和位置的目标。这些方法首先使用卷积神经网络对输入图像进行特征提取，然后在这些特征图的不同层级上应用区域提议网络(Region Proposal Network，RPN)和目标分类器检测不同尺度下的特征。

对于点云特征对齐的特征金字塔来说，每一次下采样都会生成一层金字塔。这些不同层次的特征代表了在原始点云集上的不同特征提取方式，不同层次的特征包含了不同尺度的信息。

在2.1小节中描述的曲面特征提取网络所得点特征，定义为点集 $\boldsymbol{X}_1 \in \mathbb{R}^{(N\times(3+F))}$，上述改进特征提取网络所得点特征，定义为点集 $\boldsymbol{X}_2 \in \mathbb{R}^{(N\times O)}$，将两个特征使用特征金字塔进行特征聚合操作得到单一聚合特征 F，如公式(7)所示：

$$\begin{aligned}\boldsymbol{F}'_1&=Conv_{1\times1}(\boldsymbol{X}_1)\\ \boldsymbol{F}'_2&=Conv_{1\times1}(UpSample(\boldsymbol{X}_2))\\ \boldsymbol{F}_{\text{fused}}&=\boldsymbol{F}'_1+\boldsymbol{F}'_2\\ \boldsymbol{F}&=Conv_{3\times3}(\boldsymbol{F}_{\text{fused}})\end{aligned} \tag{7}$$

其中：$Conv_{1\times1}(\cdot)$ 代表卷积核大小为1×1的卷积层；$Conv_{3\times3}(\cdot)$ 代表卷积核大小为3×3的卷积层；$UpSample(\cdot)$ 代表一个使用最近邻插值(Nearest-Neighbor Interpolation)的上采样网络，其主要作用为对较小尺寸的矩阵 $\boldsymbol{X}_2$ 进行插值，并放缩至 $\boldsymbol{X}_1$ 的大小。

在使用Transformer的目标检测主干网络中，它将目标检测问题看成直接从输入特征图中提取并预测物体边界框和类别的问题。端到端的主干网络可以摆脱传统目标检测框架中的一些步骤，如锚框生成以及非极大值抑制(Non-Maximum Suppression，NMS)等。

在主干网络中，Transformer模型被用来对特征进行编码和解码，以生成最终的目标检测结果。具体来说，编码器部分输入的是由特征金字塔卷积得到的聚合特征 $\boldsymbol{F}$，使用编码函数 $Encoder(\cdot)$ 对聚合特征 $\boldsymbol{F}$ 进行编码得到编码特征 $\boldsymbol{E}$，如公式(8)所示：

$$\boldsymbol{E}=Encoder(\boldsymbol{F}) \tag{8}$$

其中：$Encoder(\cdot)$ 是Transformer的编码器函数，主要包含自注意力层(Self-Attention Layer)和前馈神经网络(Feed-Forward Layer)，在两个主要的模块之后，Transformer编码器还使用了层标准化(Layer Normalization)和残差连接(Residual Connection)以稳定训练过程并加快收敛速度。

解码器部分则接收编码特征 $\boldsymbol{E}$ 和一组固定数量的查询(Query)作为输入，通过自注意力、编码器-解码器注意力(Encoder-Decoder Attention)和FC层生成最终的目标检测结果。使用编码特征 $\boldsymbol{E}$ 和查询矩阵 $\boldsymbol{Q}$，进一步使用 $Decoder(\cdot)$ 函数生成解码特征 $\boldsymbol{D}$，如公式(9)所示：

$$\boldsymbol{D}=Decoder(\boldsymbol{E},\boldsymbol{Q}) \tag{9}$$

对于每个解码器，其与编码器类似，同样包含自注意力层、前馈神经网络、层标准化和残差连接。在自注意力后，解码器还包含一个交叉注意力层。交叉注意力层类似于自注意力层，但是它使用解码器的当前输出作为查询，而用编码器的输出作为键(key)和值(value)，使得解码器在生成每个输出时，都可以考虑到编码器输出的全局信息。

解码后，所得的解码特征 $\boldsymbol{D}$ 包含目标框信息和目标分类信息。若 $\boldsymbol{D}$ 的长度为 d，物体类别数量为 C，则目标框的预测通过一个线性变换实现，如公式(10)所示：

$$\boldsymbol{box}=\boldsymbol{W}_{\text{box}}\boldsymbol{D}+\boldsymbol{b}_{\text{box}} \tag{10}$$

其中:$\boldsymbol{W}_{\text{box}}\in\mathbb{R}^{(4\times D)}$ 和 $\boldsymbol{b}_{\text{box}}\in\mathbb{R}^{4}$ 是学习到的权重和偏置;$\boldsymbol{b}_{\text{box}}\in\mathbb{R}^{4}$ 是预测的边界框,包含中心坐标(cx,cy)和宽度 w、高度 h。类别预测 $\boldsymbol{class}$ 则通过线性变换和之后的 $softmax(\cdot)$函数实现,如公式(11)所示:

$$\boldsymbol{class}=softmax(\boldsymbol{W}_{\text{class}}\boldsymbol{D}+\boldsymbol{b}_{\text{class}}) \tag{11}$$

其中:$\boldsymbol{W}_{\text{class}}\in\mathbb{R}^{(C\times D)}$ 和 $\boldsymbol{b}_{\text{class}}\in\mathbb{R}^{C}$ 是学习到的权重和偏置,$\boldsymbol{class}\in\mathbb{R}^{C}$ 是预测的类别值。$softmax(\cdot)$函数确保输出值在0和1之间,并且所有类别的概率和为1,如公式(12)所示:

$$softmax(\boldsymbol{x})=\frac{\mathrm{e}^{x_i}}{\sum_{i=1}^{K}\mathrm{e}^{x_i}} \tag{12}$$

通过这种方式,基于Transformer解码器生成的特征被变换为预测框和类别,相较于两阶段的目标检测,端到端的目标检测计算量更小且算法不依赖于极大值抑制(NMS)等步骤,这使得本文算法更加有利于工业环境下的硬件要求和检测速度要求。

模型使用的损失函数包含4个部分:主干网络中产生的多阶段损失、目标检测任务所产生的分类损失、边界框产生的回归损失和旋转损失。其中,基于Transformer的主干网络产生的损失是由每个隐藏层产生的损失的平均值,如公式(13)所示:

$$Loss_{\text{hidden}}=\frac{1}{L+1}\sum_{l=0}^{L}Loss^{(l)} \tag{13}$$

其中:$Loss^{(l)}$ 代表第 l 个隐藏层产生的损失。目标检测中产生的分类损失基于交叉熵设计,如公式(14)所示:

$$Loss_{\text{cls}}=-\sum\boldsymbol{y}_{\text{true}}\times\log(\boldsymbol{y}_{\text{pred}}) \tag{14}$$

其中:$\boldsymbol{y}_{\text{true}}$ 代表真实分类,$\boldsymbol{y}_{\text{pred}}$ 代表预测产生的分类。边界框产生的回归损失是平滑损失(Smooth Loss)$Loss_{\text{reg}}$,如公式(15)所示:

$$Loss_{\text{reg}}=SmoothL1(\boldsymbol{b}_{\text{true}},\boldsymbol{b}_{\text{pred}}) \tag{15}$$

其中:$SmoothL1(\cdot)$代表平滑函数。边界框旋转产生的损失 $Loss_{\text{rot}}$ 是相对角度间的二阶范数,如公式(16)所示:

$$Loss_{\text{rot}}=\|\boldsymbol{R}_{\text{true}}-\boldsymbol{R}_{\text{pred}}\|_{F}^{2} \tag{16}$$

其中:$\boldsymbol{R}_{\text{true}}$ 代表真实检测框的相对角度,$\boldsymbol{R}_{\text{pred}}$ 代表实际检测框的相对角度,$\boldsymbol{R}_{\text{true}}$ 和 $\boldsymbol{R}_{\text{pred}}$ 中元素的值为 π 的倍数。

3 实验结果(Experiment results)

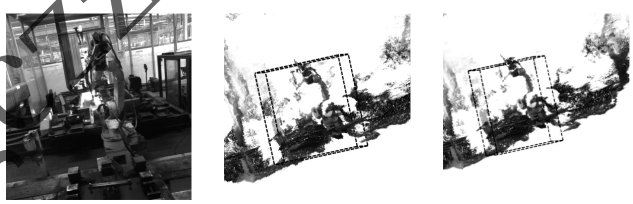
为了验证本文提出的基于表面增强尾骨和多尺度可变形DeTR的三维多尺度点云目标检测方法的有效性,将其应用在公开数据集ScanNet[13]上进行评估。此外,为了验证该方法在工业室内场景下的目标检测的有效性,在各个数据集规范的基础上对其进行训练和验证。本文遵循标准评估协议[14],并使用不同IoU阈值下的平均准确率均值(mAP)作为指标,不考虑边界框的方向。

3.1 ScanNet数据集

ScanNet V2数据集[13]是一个基于室内场景的三维重建数据集,由1 513个室内场景和18个对象类别组成,提供了每个点的实例、语义标签和三维边界框的注释。该数据集的室内场景大多为卧室、客厅、洗手间和办公室等,对象类别大多为地板、墙、椅子和沙发等,与其他公开数据集相比,ScanNet V2数据集的场景更加完整,平均覆盖的区域更大且场景更加杂乱,因此目标检测的难度更大。本文遵循标准评估协议,在设置0.25阈值(mAP@0.25)的平均精度和设置0.5阈值(mAP@0.5)的平均精度下进行评估。

3.2 工业三维目标检测数据集

工业机器人三维目标检测数据集是在杭州西奥电梯有限公司的机器人生产车间中采集的,通过zed2相机对机器人生产环境进行点云数据的提取。在符合各数据集制作标准的前提下,由人工对数据进行框选和标注,并由不同的人员对数据进行验证和调整。本数据集包含多个机器人作业场景,每个场景对象位置不同,拍摄角度不同且环境亮度不断变化。本数据集共2 234个数据,其中包含1 675个训练用例和559个测试用例,增加了4个检测对象类别,分别是机器人、推车、板材和传送带。本文在数据集上进行了实验,结果如图5所示。实验在ScanNet V2数据集上进行,使用mAP@0.25和mAP@0.5作为衡量指标。



(a)实际场景　(b)真实值　(c)本文模型预测值

图5　模型在工业室内数据集的表现

Fig. 5 The performance of the model on the industrial indoor dataset

3.3 结果分析

针对三维目标检测领域中的大规模场景室内数据集ScanNet V2,实验选用了主流的检测方法同本文模型进行对比,这些方法分别是层次几何网络(Hierarchical Geometry Network, HGNet)[14]、生成稀疏检测网络(Generative Sparse Detection Network, GSDN)[15]、三维语义实例分割的多提案聚合网络(3D Multi Proposal Aggregation, 3D-MPA)[16]、深度霍夫投票网络(VoteNet)[10]、多级上下文霍夫投票网络(Multi-Level Context VoteNet, MLCVNet)[17]、回溯代表点投票网络(Back-tracing Representative Network, BRNet)[18]、混合几何原语网络(Hybrid Geometric Primitives 3D Network, H3DNet)[19]、三维目标检测转换器(3D Detection Transformer, 3DeTR)[20]、类别感知分组网络(Class-Aware Grouping for 3D Object Detection, CAGroup3D)[21]、视锥网络(Frustum PointNet, F-PointNet)[22]、多模态令牌转换器网络(TokenFusion)[23]、全卷积无锚点网络(Fully Convolutional Anchor-Free 3D Object

Detection, FCAF3D)[24]。如表1所示,本文算法的平均准确率均值与最先进的模型CAGroup3d相近,mAP@0.25超过最先进的模型CAGroup3d,达到76.0%;mAP@0.5超过最先进的模型CAGroup3d,达到62.2%。

表1 在ScanNet V2数据集上进行的对比实验

Tab.1 Comparative experiments performed on the ScanNet V2 dataset

算法	主干网络	输入	*mAP*@0.25/%	*mAP*@0.5/%
HGNet	GU-net	point	61.3	34.4
GSDN	MinkNet	point	62.8	34.8
3D-MPA	MinkNet	point	64.2	49.2
VoteNet	PointNet++	point	62.9	39.9
MLCVNet	PointNet++	point	64.5	41.4
BRNet	PointNet++	point	66.1	50.9
H3DNet	4×PointNet++	point	67.2	48.1
3DeTR	Transformer	point	65.0	46.9
CAGroup3D	ResNet18	point	75.0	61.3
F-PointNet	PointNet	point+RGB	19.8	10.8
TokenFusion	Transformer	point+RGB	70.8	54.2
FCAF3D	HDResNet34	point	71.5	57.3
本文算法	Transformer	point	**76.0**	**62.2**

注:加粗数值为各列的最优结果。

3.4 消融实验

本文所述方法使用了3个模块,其中尾骨部分包括曲面特征提取模块、分层特征提取模块。为了证明本文方法使用的模块的有效性,设计了一个逐个拆除模块的消融实验进行验证。

实验使用分层特征提取模块和去掉特征金字塔模块的Transformer的主干网络模块进行实验,使用PointNet++作为尾骨和去掉特征金字塔模块的Transformer的主干网络模块进行对比实验,并测定在ScanNet V2和工业室内数据集上的平均表现,收集测试集上产生的精度数据,与本文方法进行对比。

如表2所示,针对曲面特征提取模块和注意力改进的分层提取模块实用性在ScanNet数据集和工业室内数据集上开展实验。基线特征提取模块仅使用PointNet++,AMLP代表该模型仅使用注意力增强的分层特征提取模块,RepSurf-U代表该模型仅使用伞状曲面特征提取模块,O代表没有使用对应模块,P代表使用了对应模块。针对尾骨模块的实用性,使用注意力增强的分层特征提取模块以及曲面特征提取模块可以明显增强模型在ScanNet V2数据集上的准确率,在仅使用注意力增强的尾骨模块中,准确率相较原基线方法提升了约2.4%,在增加使用曲面特征提取模块后,准确率相较仅使用基于注意力增强多层感知机的特征提取模块提升了约9.5%,证明本文的尾骨模块在基线方法上做出了有效改进。

表2 消融实验

Tab.2 Ablation experiments

算法	模块		ScanNet+室内数据集
	AMLP	RepSurf-U	*mAP*@0.25/%
基线	O	O	65.0
本文算法	P	O	67.4
本文算法	P	P	**76.9**

注:加粗数值为各列最优结果。

4 结论(Conclusion)

在本研究中,成功地提出并实现了一个具有创新性的三维点云目标检测框架,该框架适用于工厂的室内环境。设计的框架主要由3个部分构成:基于表面表示的曲面特征提取模块、基于注意力增强多层感知机的特征提取模块,以及基于多尺度特征金字塔和Transformer的注意力主干网络。这种结构设计在复杂的室内环境中实现了高精度的目标检测,实验结果也证明了其优越性。值得强调的是,方法中实现了一种创新的特征提取技术,即将注意力机制融入多层感知机的采样过程中,能够更有效地提取点云集的特征,在目标检测中取得了更好的效果。同时,检测网络采用特征金字塔和Transformer,简化了检测过程,也大大提高了目标检测的速度,对于实际的工业应用场景具有很大的价值。总体来说,本研究不仅提供了一种理论上的新方法,还通过实验证明了这种方法的有效性。希望该研究能引起更多研究人员的兴趣,共同推动其不断进步。同时,希望这个框架能够在未来的工业应用中发挥价值,改善和优化目标检测的效果。

参考文献(References)

[1] GIRSHICK R,DONAHUE J,DARRELL T,et al. Region-based convolutional networks for accurate object detection and segmentation[J]. IEEE transactions on pattern analysis and machine intelligence,2015,38(1):142-158.

[2] REDMON J,DIVVALA S,GIRSHICK R,et al. You only look once: unified, real-time object detection[C]//IEEE. Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016:779-788.

[3] BIMBRAW K. Autonomous cars: past, present and future a review of the developments in the last century, the present scenario and the expected future of autonomous vehicle technology[C]//IEEE. 2015 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO). Piscataway: IEEE, 2015:191-198.

[4] CADENA C, CARLONE L, CARRILLO H, et al. Past, present, and future of simultaneous localization and mapping: toward the robust-perception age[J]. IEEE, 2016, 32(6):1309-1332.

[5] BUTIME J, GUTIERREZ I, CORZO L G, et al. 3D recon-

struction methods,a survey[C]//INSTICC. Proceedings of the First International Conference on Computer Vision Theory and Applications. Setúbal:INSTICC,2006:457-463.

[6] GUO Y,WANG H,HU Q,et al. Deep learning for 3D point clouds:a survey[J]. IEEE transactions on pattern analysis and machine intelligence,2020,43(12):4338-4364.

[7] QI C R,SU H,MO K,et al. Pointnet:deep learning on point sets for 3D classification and segmentation[C]// IEEE. Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE,2017:652-660.

[8] QI C R,YI L,SU H,et al. Pointnet++:deep hierarchical feature learning on point sets in a metric space[J]. Advances in neural information processing systems,2017,30(1):5105-5114.

[9] SHI S,WANG X,LI H. Pointrcnn:3D object proposal generation and detection from point cloud[C]//IEEE. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE,2019:770-779.

[10] QI C R,LITANY O,HE K,et al. Deep hough voting for 3D object detection in point clouds[C]//IEEE. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 9277-9286.

[11] XIE Q,LAI Y K,WU J,et al. Venet:voting enhancement network for 3D object detection[C]//IEEE. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE,2021:3712-3721.

[12] HUANG Z,YU Y,XU J,et al. Pf-net:point fractal network for 3D point cloud completion[C]//IEEE. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 7662-7670.

[13] DAI A,CHANG A X,SAVVA M,et al. Scannet:richly-annotated 3D reconstructions of indoor scenes[C]//IEEE. Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2017:5828-5839.

[14] YAO T,LI Y,PAN Y,et al. Hgnet:learning hierarchical geometry from points, edges, and surfaces[C]//IEEE. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2023:21846-21855.

[15] GWAK J Y,CHOY C,SAVARESE S. Generative sparse detection networks for 3D single-shot object detection[C]// Springer. Computer Vision-ECCV 2020. Berlin: Springer, 2020:297-313.

[16] ENGLEMANN F,BOKELOH M,FATHI A,et al. 3D-mpa:multi-proposal aggregation for 3D semantic instance segmentation[C]//IEEE. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE,2020:9031-9040.

[17] XIE Q,LAI Y K,WU J,et al. Mlcvnet:multi-level context votenet for 3D object detection[C]//IEEE. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 10447-10456.

[18] CHENG B,SHENG L,SHI S,et al. Back-tracing representative points for voting-based 3D object detection in point clouds[C]//IEEE. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE,2021:8963-8972.

[19] ZHANG Z,SUN B,YANG H,et al. H3DNet:3D object detection using hybrid geometric primitives[C]//Springer. Computer Vision-ECCV 2020. Berlin:Springer,2020: 311-329.

[20] MISRA I,GIRDHAR R,JOULIN A. An end-to-end transformer model for 3D object detection[C]//IEEE. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE, 2021:2906-2917.

[21] WANG H,DONG S,SHI S,et al. Cagroup3D:class-aware grouping for 3D object detection on point clouds[J]. Advances in neural information processing systems,2022, 35(1):29975-29988.

[22] QI C R,LIU W,WU C,et al. Frustum pointnets for 3D object detection from rgb-d data[C]//IEEE. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 918-927.

[23] WANG Y,CHEN X,CAO L,et al. Multimodal token fusion for vision transformers[C]//IEEE. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE,2022:12186-12195.

[24] RUKHOVICH D,VORONTSOVA A,KONUSHIN A. Fcaf3D:fully convolutional anchor-free 3D object detection[C]//Springer. Computer Vision-ECCV 2022. Berlin: Springer,2022:477-493.

作者简介:

顾方宇(1999-),男,硕士生。研究领域:计算机视觉,深度学习。

胡海洋(1977-),男,博士,教授。研究领域:机器视觉,智能制造。