基于交叉融合自注意力的点云语义分割

舒军[1],王帅[1],杨莉[2*],陈宇[2]
(1 湖北工业大学 太阳能高效利用及储能运行控制湖北省重点实验室,武汉 430068;2 湖北第二师范学院 计算机学院,武汉 430205)

摘 要 针对目前点云语义分割算法通常采用简单的串联三维原始坐标信息方式建模几何信息,导致建模不完整问题.提出了交叉融合自注意力网络,在该网络的编码层中设计了交叉融合自注意力机制模块,通过交互学习坐标和特征信息,提高局部细粒度特征描述能力,使得几何信息建模更加完整.同时为了更好地结合浅层与高层特征,提出了一种层级特征融合模块,通过自适应地连接网络不同层,实现不同层的特征整合.在S3DIS、Semantic3D和SemanticKITTI数据集上实验表明:该算法优于RandLA-Net等先进算法.

关键词 点云;语义分割;交叉融合自注意力;层级特征融合
中图分类号 TP391 **文献标志码** A **文章编号** 1672-4321(2025)01-0096-11
doi:10.20056/j.cnki.ZNMDZK.20250730

Point cloud semantic segmentation algorithm based on cross fusion self-attention

SHU Jun[1], WANG Shuai[1], YANG Li[2*], CHEN Yu[2]
(1 Hubei Key Laboratory for High-efficiency Utilization of Solar Energy and Operation Control of Energy Storage System, Hubei University of Technology, Wuhan 430068, China; 2 College of Computer, Hubei University of Education, Wuhan 430205, China)

Abstract In response to the existing issue of incomplete geometry modeling caused by the prevalent approach of simply concatenating raw 3D coordinate information in current point cloud semantic segmentation algorithms, a Cross-Fusion Self-Attention Network is proposed. Within the encoding layers of the network, the Cross-Fusion Self-Attention Mechanism module is introduced, which leverages interactive learning between coordinate and feature information to enhance the capability of describing fine-grained local features. This leads to a more comprehensive modeling of geometric information. Additionally, to effectively integrate shallow and deep-level features, a Hierarchical Feature Fusion module that adaptively connects different layers of the network is proposed, enabling the integration of features from various levels. Experimental results on the S3DIS, Semantic3D, and SemanticKITTI datasets demonstrate the superiority of our algorithm over advanced approaches such as RandLA-Net.

Keywords point cloud; semantic segmentation; cross-fusion self-attention; hierarchical feature fusion

点云数据是由一系列三维点组成的数据集,用于表示三维空间中的几何结构或物体表面.相较于二维图像数据,三维点云是无序和非结构化的数据[1],但是点云提供了更加丰富的信息,能够更加充分的表征一些复杂场景.针对点云数据结构,设计一个神经网络对三维点云进行语义分割是一项具有挑战性的工作.

为应对点云语义分割这一挑战,近年学者们提出了越来越多的3D点云分割深度学习框架.主要可以分为基于投影、体素和点的三种方法[2-3].基于投

收稿日期 2023-07-21 ***通信作者** 杨莉,研究方向:智能控制,E-mail: hbyangli@hue.edu.cn
作者简介 舒军(1973-),男,副教授,博士,研究方向:计算机视觉,E-mail: shujun@hbut.edu.cn
基金项目 国家自然科学基金资助项目(61603127);湖北省教育科学规划项目(2022ZA41)

影的方法[4-5]是通过将三维点云数据投影到二维图像上，利用神经网络对投影图像进行特征提取，以获取三维模型表面信息．然而，基于投影的方法会降低点云维度，从而在聚合几何和结构信息方面存在一定缺陷．基于体素的方法[6-7]是将点云体素化为密集的3D网格，再通过神经网络进行处理．然而，体素方法需要不断提升体素分辨率以提高整体精度，但这会导致内存增加的问题．基于点的方法是对每个点直接进行处理，提取出点云中特征信息，并使用这些特征进行后续的处理．QI等人提出了具有里程碑意义的PointNet[8]，是第一个直接使用神经网络处理点云数据而无需额外操作的网络．受到PointNet启发，相关学者提出了一系列直接处理原始点云数据的网络．QI等人提出PointNet++[9]网络，该网络设计了一种多层次局部特征聚合模块，可以更好聚合局部特征．THOMAS等人提出了KPConv[10]，引入了一个新概念Kernel Points，即在点云中自适应选取一些点作为卷积核的模板，然后通过对这些模板进行插值来构造卷积核．HU等人提出了RandLA-Net[11]，专门针对大规模点云数据，提高分割效率．现有的算法忽略了以下两个关键问题：首先，在建模几何信息时，由于缺乏交互融合坐标特征信息的能力，难以准确捕捉点云数据中空间信息；其次，在编码层，由于缺乏结合浅层和高层语义特征的能力，难以有效地分割出特征相似样本．

输入点云数据$P=\{p_i,f_i|i=1,2,3,\cdots,N\}$是坐标和特征信息的集合，其中$N$是点云数量，$p_i\in R^{1\times 3}$为坐标信息，$f_i\in R^{1\times d}$为特征信息（例如，颜色、法向量等）．以往工作通常是将特征信息和坐标信息分开处理．在特征提取过程中仅简单的串接3维原始坐标信息对几何信息建模，可能导致模型泛化能力和鲁棒性下降．在2D图像中，合理的解决方案是使用卷积来考虑提取特征过程中相对位置关系．由于点云数据的无序性，无法使用大的卷积核进行操作．然而，如果在点云中执行相似的操作，可以有效地增强模型的特征提取能力．

随着点云神经网络研究深入，在二维图像与自然语言处理领域取得了显著成果的自注意力机制，被广泛应用于三维点云处理[12-13]．REN等人提出PA-Net[14]，该网络设计了两个并行的自注意力机制，同时关注坐标和特征信息．ZENG等人提出LEARD-Net[15]，通过自注意力机制，让网络同时关注空间几何结构、颜色信息和语义特征．然而，上述网络均未考虑坐标特征的交互融合问题．

层级特征融合在二维图像处理中同样有着广泛应用．例如，GU等人[16]使用两个并行编码器来提取不同层信息并将它们合并．ZHAO等人[17]使用不同大小的全局平均池来构建空间金字塔以融合不同层特征．在点云中，HUANG等人[18]提出了一种基于特征的多尺度网络来完成点云任务．LI等人[19]提出了一种多尺度域特征和聚合模型，增强网络特征提取能力．GENG等人[20]提出了一种多尺度注意力聚合网络，从编码器和解码器捕获全局特征．本文提出了一种新的层级特征融合方式，以自适应地连接网络的不同层．

为了解决点云数据中坐标信息和特征信息交互融合问题，本文设计了交叉融合自注意力（Cross-fusion self-attention，CFSA）机制模块，该模块能够交互地考虑点云特征和坐标信息．在CFSA模块中，所有特征和坐标都能自适应地增强彼此的表达能力，从而实现了坐标和特征的有效融合．

为了解决点云数据中浅层特征和高层语义特征结合问题，本文提出了一种层级特征融合（Hierarchical feature fusion，HFF）模块．该模块从上到下融合编码部分的特征，通过自适应地合并前一层特征来实现特征融合．总的来说，本文贡献可以总结如下：

（1）设计了CFSA模块，该模块不仅具有置换不变性，无论输入点的顺序如何变化，该模块都能给出相同的特征提取结果．而且能够交互地提取坐标和特征信息，自适应增强坐标和特征的信息，建模更加完整的几何信息．

（2）设计了HFF模块，自适应地捕获不同尺度的特征．它可以很容易地嵌入到网络的不同层中，为网络带来更丰富的尺度和梯度信息．

（3）基于CFSA和HFF模块，提出了一个新的点云语义分割网络．能够有效地处理点云的分割任务，在S3DIS、Semantic3D和SemanticKITTI数据集上获得了有竞争力的结果．

1　本文算法

模型如图1所示，使用包含N个点的点云集合，其中每个点具有xyz坐标位置信息和特征信息作为输入，采用具有跳跃连接的编解码器结构．将点云输入到五个编解码层来学习每个点的特征．在编码端，点云通过基于交叉融合自注意力的局部特征提取（Local feature extraction，LFE）模块，丰富坐标信

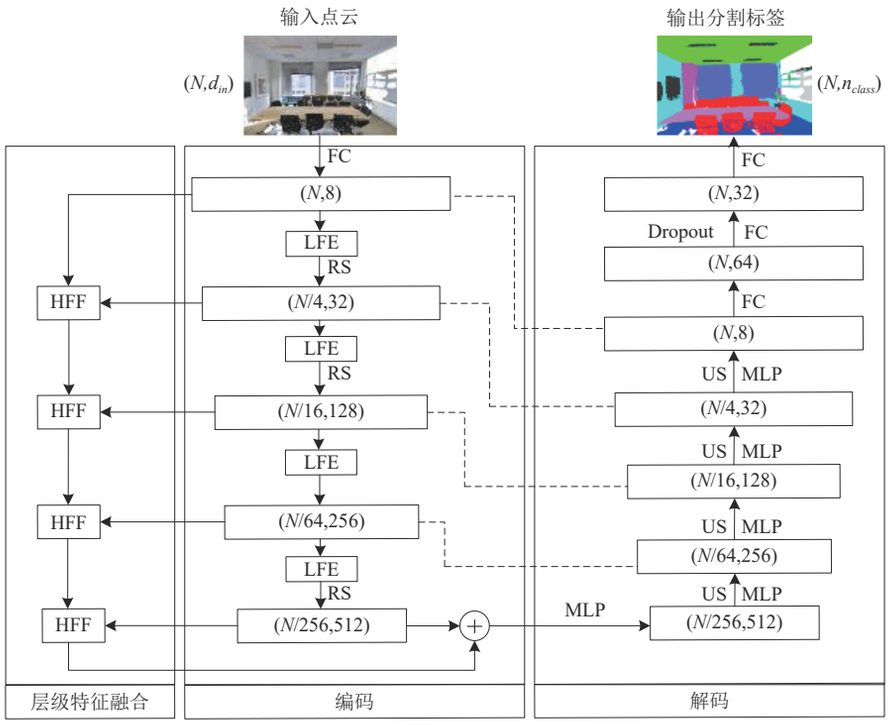

图1 网络整体结构

Fig. 1 Overall network structure

息,交互融合坐标和特征信息,扩大每个点感受野.使用随机采样(Random Sampling,RS)减小点云规模.同时通过设计层级特征融合模块,聚合每一层的语义信息,并级联到编码层的最后.解码端每个点采用(*K*-NearestNeighbor,KNN)方式获得最邻近点,通过线性插值方法[21]进行上采样(Up Sampling,US),同时与对应编码端的特征进行合并.最后通过全连接(Full Connection,FC)和Dropout操作获得最终的分割结果.

1.1 基于交叉融合注意力的局部特征提取(CFSA-LFE)

局部特征提取(LFE)是编码层的核心,主要由三个模块组成,包括局部坐标编码模块(Local coordinate encoding,LCE)、交叉融合自注意力(CFSA)池化模块和残差优化(Residual optimization,RO)模块.

(1)局部坐标编码(LCE).

为丰富增强坐标信息,LCE模块采用KNN方法获取每个点的K个最近邻点的位置信息$\{p_i^1\cdots p_i^m\cdots p_i^K\}$,并对其进行编码.具体结构如图2所示.编码过程定义如下:

$$r_i^K = G(\mathrm{MLP}(g(p_i, p_i^m, p_i - p_i^m, \| p_i - p_i^m \|))), \quad (1)$$

式中$p_i \in R^{1\times 3}$为i点坐标,$i \in \{1\cdots N\}$;$p_i^m \in R^{1\times 3}$为i点的第m个邻点坐标,$m \in \{0\cdots K\}$;$(p_i - p_i^m) \in R^{1\times 3}$为$i$点与邻居点相对坐标;$\| p_i - p_i^m \| \in R^{1\times 1}$为$i$点与其邻居点欧式距离;$g$表示连接操作,即将$p_i$和$p_i^m$,$p_i - p_i^m$,$\| p_i - p_i^m \|$进行拼接,得到维度为(1,10)的相对空间位置信息;多层感知机(Multilayer Perceptron,MLP)将连接后的相对空间位置信息扩展到和$f_i \in R^{1\times d}$一样的维度,得到i点的第m个邻点高维度的相对空间位置信息,MLP是一种基本的前馈人工神经网络模型,通过堆叠多个全连接层和非线性激活函数,允许MLP在输入数据的多个维度上进行组合和交互,以逐层逐渐拓展特征的维度.最终,将i点的K个最近邻点高维度的相对空间位置信息拼接,得到i点的局部坐标编码$r_i^K \in R^{K\times d}$,将K个最近邻点特征拼接得到i点的局部特征$f_i^K \in R^{K\times d}$.

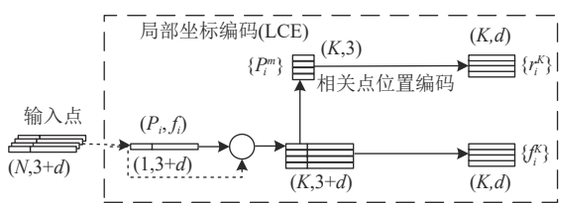

图2 LCE模块

Fig. 2 LCE structure

(2)交叉融合自注意力池化.

自注意力机制的原理是通过计算每个位置与其他位置之间的相关性来捕捉序列中元素之间的依赖关系,然后根据相关性对不同位置的元素进行加权求和,生成上下文表示.CFSA模块使用强大的

自注意力机制交互增强局部坐标和特征信息，它所接收的输入由LCE模块的输出构成，即LCE后的坐标和特征信息.该模块结构具体如图3所示，具体的计算方法如下：

上半部分输入为r_i^K，将r_i^K线性变换之后得到$r_{i-\text{query}}^K, r_{i-\text{key}}^K, r_{i-\text{value}}^K$三个新的特征描述.同理，$f_{i-\text{query}}^K, f_{i-\text{key}}^K, f_{i-\text{value}}^K$是由下半部分输入$f_i^K$线性变换之后得到，线性变换过程可以描述为：

$$\begin{cases} r_{i-\text{query}}^K, r_{i-\text{key}}^K, r_{i-\text{value}}^K = L(r_i^K) \\ f_{i-\text{query}}^K, f_{i-\text{key}}^K, f_{i-\text{value}}^K = L(f_i^K) \end{cases}, \tag{2}$$

其中L代表线性变换，query、key、value分别代表查询、键和值.在注意力机制中，将输入向量转换为查询(query)、键(key)和值(value)向量的过程通常被称为"线性变换".这个线性变换是通过矩阵乘法实现的.

具体而言，给定输入向量$\boldsymbol{X}$，可以通过以下线性变换将其转换为查询向量$\boldsymbol{Q}$、键向量$\boldsymbol{K}$和值向量$\boldsymbol{V}$：$Q = X \cdot W_Q, K = X \cdot W_K, V = X \cdot W_V$，其中$W_Q, W_K, W_V$是可学习的权重矩阵，它们分别用于对输入向量进行查询、键和值的线性变换.

然后通过交叉自注意力运算得到r_{i-o}^K, f_{i-o}^K，该过程定义如下：

$$\begin{cases} r_{i-o}^K = r_{i-\text{value}}^K \otimes f_{i-a}^K + r_i^K \\ f_{i-o}^K = f_{i-\text{value}}^K \otimes r_{i-a}^K + f_i^K \end{cases}, \tag{3}$$

其中⊗表示矩阵乘法，其中r_i^K, f_i^K表示原始的残差分支，由式(3)可以看出，CFSA模块交互增强了坐标和特征信息，上式中的r_{i-a}^K, f_{i-a}^K由查询和键加权得到，具体过程定义如下：

$$\begin{cases} r_{i-a}^K = \text{soft}\max(\text{sum}((r_{i-\text{query}}^K)^T \otimes r_{i-\text{key}}^K)) \\ f_{i-a}^K = \text{soft}\max(\text{sum}((f_{i-\text{query}}^K)^T \otimes f_{i-\text{key}}^K)) \end{cases}, \tag{4}$$

其中⊗同样也是表示矩阵乘法，sum表示将矩阵乘法的结果的每一行与第一行相加，最后通过softmax分配权重.

相较于传统自注意力机制，CFSA机制在LCE后实现了坐标和特征信息的交互作用增强.最终，通过连接和池化处理，得到查询点新特征描述F_{i-out}^K，具体定义如下：

$$F_{i-\text{out}}^K = \text{MLP}\left(\sum_{i=1}^{K} g(r_{i-o}^K, f_{i-o}^K)\right). \tag{5}$$

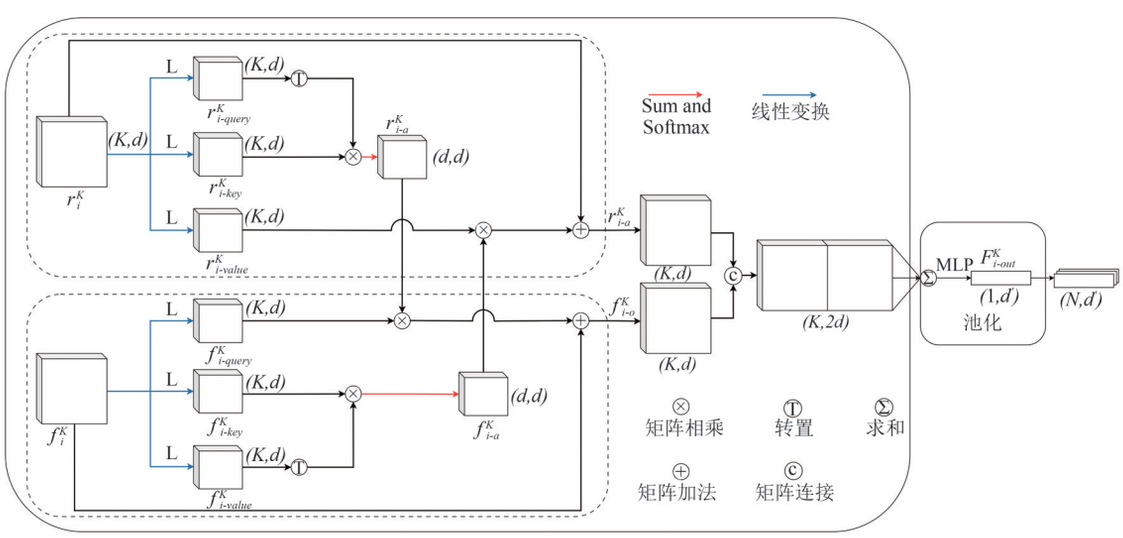

图3　交叉融合自注意力池化模块

Fig. 3　Cross fusion self-attention pooling module

(3)残差优化.

为处理大规模点云数据，编码阶段采用随机抽样下采样.随机抽样具有低时间和空间复杂度、内存占用少、高效率等优点.然而，随机采样可能会导致关键点信息丢失.为减轻关键点信息丢失的影响，本文引入残差优化模块，在本研究中堆叠LCE模块和CFSA池化模块以有效扩大每个点的感受野.根据以往的理论，LCE模块和CFSA池化模块堆叠次数越多，扩展效果越好.然而，考虑计算效率和模块可迁移性，本文通过对LCE模块和CFSA池化模块进行了两次堆叠，并添加跳跃连接以实现残差学习.具体结构见图4.

1.2　层级特征融合

鉴于点云具有广泛的范围和复杂多样的物体尺度，为了扩大模型的感受野并保留更多的局部细粒度信息，本文提出了一种层级特征融合模块，用

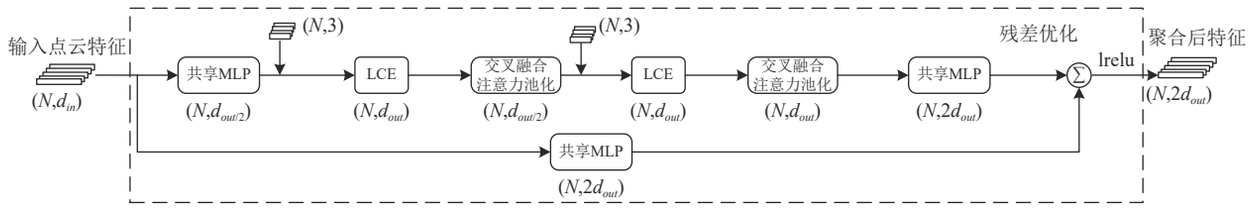

图4 残差优化模块

Fig. 4 Residual optimization module

于有效地融合不同尺度的特征.该模块将相邻尺度的特征进行组合,以实现这一目标.

HFF采用层级融合策略对不同层次特征进行融合,将邻近层特征信息采用一个上下文注意力融合模块进行融合,结合浅层语义特征和高层语义特征生成注意力权重,考虑不同尺度特征间的差异,对原始点和采样点之间进行跨级别长依赖关系建模.HFF模块详细结构如图5所示.

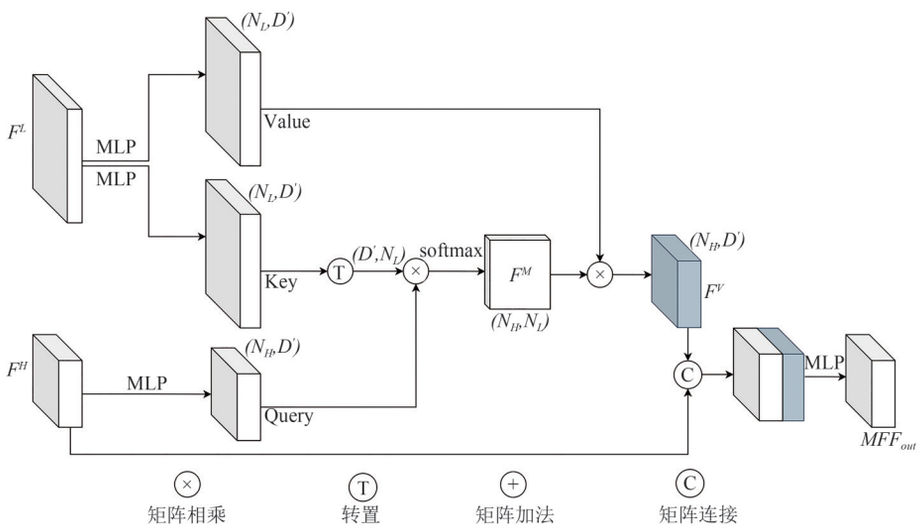

图5 层级特征融合模块结构

Fig. 5 Hierarchical feature fusion module structure

该模块采用跨层次注意力机制将低层语义信息$F^L \in R^{N^L \times D^L}$和高层语义信息$F^H \in R^{N^H \times D^H}(N^L > N^H)$进行融合,该模块以采样后的高层特征$F^H$为查询集,以低层特征$F^L$为键和值集.模块中亲和力矩阵计算过程可描述为如下:

$$F^M = \text{soft max}(\text{MLP}(F^H) \otimes (\text{MLP}(F^L))^T), \quad (6)$$

其中MLP表示多层感知器,⊗表示矩阵乘法,T表示转置,最后通过softmax分配权重.接着通过在F^M和值集之间应用矩阵乘法,建立采样之前和之后的特征映射.其表示为:

$$F^V = F^M \otimes (\text{MLP}(F^L)), \quad (7)$$

其中F^V表示跨层上下文信息,其进一步与F^H融合以用于增强信息,最后再经过MLP得到与F^H同尺寸的输出HFF_{out}.具体过程如下:

$$\text{HFF}_{\text{out}} = \text{MLP}(C(F^H, F^V)), \quad (8)$$

其中C表示通道维度中的级联.图5中,最后一层HFF_{out}表示聚合的多层上下文信息,并且其被级联到编码器最后一层后.通过这种方式,HFF可以分层补偿信息损失并丰富特征编码.

2 实验结果分析

在本节中,对提出的网络在三个主流的语义分割数据集(S3DIS,Semantic3D,SemanticKITTI)上进行了评估.此外,还进行了一些相关的消融实验,包括网络结构分析和自注意力机制方式选择,以验证所提出的各个模块.

2.1 数据集介绍和实验环境参数设定

本文主要在3个数据集上进行评估,分别是S3DIS、Semantic3D和SemantiKITTI.S3DIS是室内场景数据集,Semantic3D是室外场景数据集,SemanticKITTI是无人驾驶场景数据集,不同数据集有不同的点数和特征,每个数据集的详细介绍如下.S3DIS是大型室内场景点云数据集,该数据集包含6个教学和办公场景区域,包含13个类别,共271个房间.每个点云数据具有9个特征,即坐标信息x、y、z、颜色信息

R、G、B和3个对应的法向量.

Semantic3D数据集是一个庞大的自然场景点云数据集，包含超过40亿个点.该数据集覆盖了多个场景，包括街道、广场、村庄和城堡等.每个点云数据都包含7个特征，包括坐标信息(x、y、z)、反射强度以及颜色信息(R、G、B).

SemanticKITTI是自动驾驶领域的权威数据集.该数据集类别既包括行人、车辆等交通参与者，也包括停车场、人行道等地面设施，每个点云数据具有4个特征，即坐标信息x、y、z和反射强度.

实验参数设置如下：在Ubuntu20.04系统上基于TensorFlow2.6.0框架进行计算，使用NVIDIA Quadro P6000 GPU进行加速.采用Adam优化器，并将三个数据集的Batchsize分别设置为6、3、3.初始学习速率均设置为0.01，最大迭代次数均为100.

2.2　评价指标

整体准确率(OA)，平均准确率(mAcc)，平均交并比(mIoU)是常见定量评估点云语义分割性能的三个指标.其中OA表示分类器对所有样本的分类正确率，mAcc每个标签准确率的平均值，mIoU表示各个类别的预测标签与真实标签的交集和并集之比的平均数.这些指标具体计算公式分别如下：

$$\mathrm{OA}=\frac{\sum_{i=1}^{c}\mathrm{TP}}{\sum_{i=1}^{c}(\mathrm{TP}+\mathrm{FN})},\tag{9}$$

$$\mathrm{mAcc}=\frac{1}{c}\cdot\sum_{i=1}^{c}\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}},\tag{10}$$

$$\mathrm{mIoU}=\frac{1}{c}\cdot\sum_{i=1}^{c}\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}},\tag{11}$$

上式中，c指类别数量；TP(true positives)表示真实为真，预测也为真的数量；FP(false positives)表示真实为假，预测为真的数量；FN(false negtives)表示真实为真，预测为假的数量；TN(true negtives)表示真实为假，预测也为假的数量.

2.3　S3DIS数据集实验结果评估

本研究使用S3DIS数据集，将271个房间划分为6个区域，通过对这六个区域进行6倍交叉验证，评估所提出算法的性能.将所提出算法与其他算法在6个区域中的定量结果进行对比，结果如表1所示，其中最优的结果用加粗表示.本文算法在OA、mAcc和mIoU三个指标上均优于其他算法，分别为87.5%、82.4%和71.1%.在地板、柱子、椅子、写字板和杂物类别上的mIoU取得了最佳性能，分别比表中其他算法的最佳结果分别提升了1.0%、0.4%、1.7%、0.6%和0.3%.此外，在窗户和门等类别上，分割精度同样表现突出.

然后，本文将提出的算法与PointNet++、RandLA-Net进行比较，并将其定性结果进行对比，证明了本文算法的优势.如图6所示，第一列为会议室场景，第二列为走廊场景，第三列为办公室场景.每个场景都包含场景真实标签、PointNet++预测、RandLA-Net预测和本文算法预测.可以观察到本文的算法可以准确预测出相似度高的物体、小物体的边缘轮廓和使嵌入物体的轮廓更平滑，如柱子、梁和墙的拐角等几何形状相似的物体，摆放着书籍和杂物的书架等小物体的边缘，墙上的黑板等嵌入物体的轮廓.这归功于局部坐标编码模块和交互自注意力模块，局部坐标编码模块保留了丰富的局部几何信息，交叉自注意力模块增强了坐标和特征的交互学习.

表1　S3DIS数据集语义分割定量结果

Tab. 1　Quantitative results of semantic segmentation of S3DIS dataset　/%

模型	mIoU	OA	mAcc	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	book	Board	clutter
PointNet[8]	47.6	78.6	66.2	88.0	88.7	69.3	42.4	23.1	47.5	51.6	54.1	42.0	9.6	38.2	29.4	35.2
PointNet++(SSG)[9]	55.7	83.9	68.3	91.5	95.6	77.5	28.3	29.1	50.8	44.3	61.1	68.4	21.8	54.1	48.0	53.3
PointNet++(MSG)[9]	57.6	86.0	68.5	92.2	91.8	78.1	30.6	31.3	56.5	63.1	62.8	64.9	19.4	55.8	49.1	54.1
SPG[22]	62.1	85.5	73.0	89.9	95.1	76.4	62.8	47.1	55.3	68.4	73.5	69.2	63.2	45.9	8.7	52.9
PointWeb[23]	66.7	87.3	76.2	93.5	94.2	80.8	52.4	41.3	64.9	68.1	71.4	67.1	50.3	62.7	62.2	58.5
KPCnov[10]	70.6	—	79.1	93.6	92.4	83.1	63.9	54.3	66.1	76.6	57.8	64.0	69.3	74.9	61.3	60.3
RandLA-Net[11]	70.0	87.1	81.5	93.1	96.1	80.6	62.4	48.0	64.4	69.4	69.4	76.4	60.0	64.2	65.9	60.1
本文	71.1	87.5	82.4	93.2	97.1	80.3	62.9	54.7	64.8	71.6	67.7	78.1	62.1	65.1	66.5	60.6

2.4　Semantic3D数据集实验结果评估

采用Semantic3D数据集中的reduce-8进行实验评估，其包含15个地区的训练点云数据和4个地区的测试点云数据.实验定量结果如表2所示，本文算法在Semantic3D数据集上的平均交并比和总体准确率均优于对比算法，mIoU为78.3%，OA为95.1%.在建筑(教堂、市政厅、车站等)、硬景观(一个杂乱的类，例如加登墙，喷泉，银行等)和汽车上表现最佳，相较于本文对比算法的最佳结果分别提升了0.3%、1.3%和0.5%.此外，在人造地形和自然地形等

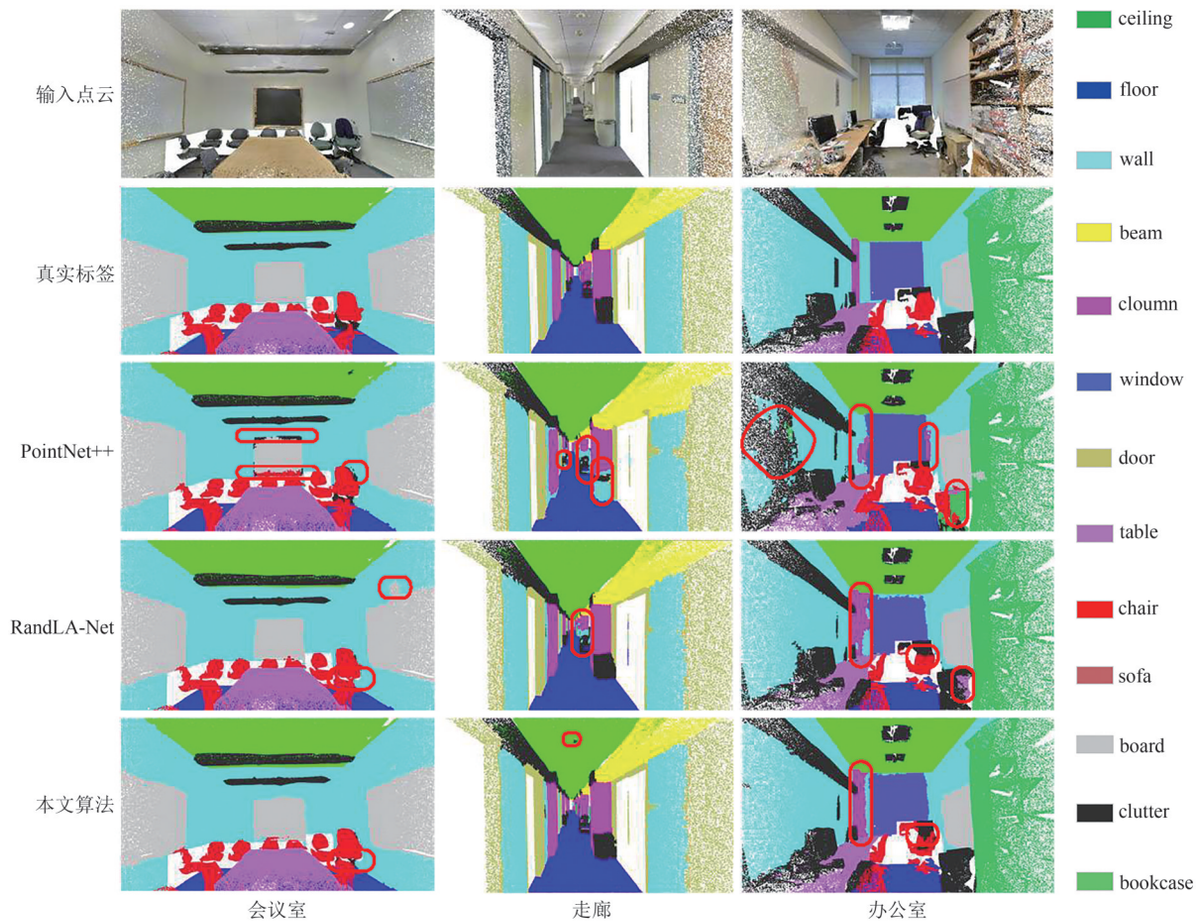

图6 S3DIS数据集语义分割可视化

Fig. 6 Visualization of semantic segmentation of S3DIS dataset

表2 Semantic3D数据集语义分割定量结果

Tab. 2 Quantitative results of semantic segmentation of Semantic3D dataset /%

模型	mIoU	OA	man-madeterrain	natural terrain	high vegetation	low vegetation	buildings	hard scope	scanning artefact	car
SnapNet[24]	59.1	88.6	82.0	77.3	79.7	22.9	91.1	18.4	37.3	64.4
ShellNet[25]	69.3	93.2	96.3	90.4	83.9	41.0	94.2	34.7	43.9	70.2
GACNet[26]	70.8	91.9	86.4	77.7	88.5	60.6	94.2	37.3	43.5	77.8
SPG[22]	73.2	94.0	97.4	92.6	87.9	44.0	83.2	31.0	63.5	76.2
RandLA-Net[11]	77.4	94.8	95.6	91.4	86.6	51.5	95.7	51.5	69.8	76.8
KPCnov[10]	74.6	92.9	90.9	82.2	84.2	47.9	94.9	40.0	77.3	79.7
本文	78.3	95.1	95.7	91.8	87.9	51.4	96.0	52.8	70.4	80.2

类别上也取得了较好的效果.

测试结果可视化图像如图7所示,由于该数据集并未公布测试集的真实标签,图中从左往右分别是输入点云数据和预测标签.从整体上看,提出算法的分割效果较好,能够有效区分建筑物、道路等目标的边界.值得注意的是,硬景观类别分布不均匀,其形状和结构变化较大,内部的几何形状、颜色和纹理特征也随场景变化而变化,但提出的算法在这种复杂情况下仍获得了最佳的分割性能.通过数据分析和结果可视化可以看出,该算法能够识别点云结构中的细节和复杂部分,有效区分不同目标的特征和细节.这表明网络具有优秀的特征提取、空间信息聚合和精确分割的能力,充分验证了特征提取模块的效果.

2.5 SemanticKITTI数据集实验结果评估

SemanticKITTI数据集是基于KITTI数据集的扩展,表3为本文算法在SemanticKITTI数据集上与现有的一些基于点、基于投影和体素的经典算法定量结果对比.从表中结果可以看出,本文的算法优于大多数的算法,mIoU为55.3%,并且在车、植被和地

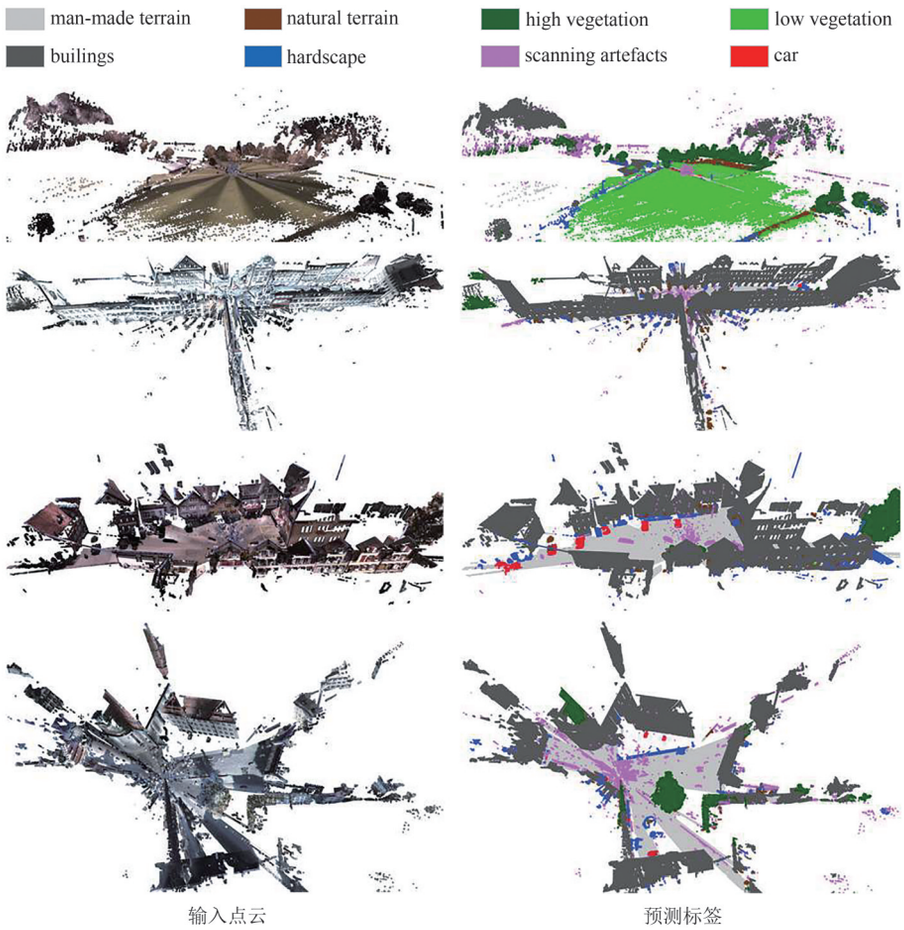

图7　Semantic3D数据集语义分割可视化

Fig. 7　Semantic3D dataset semantic segmentation visualization

表3　SemanticKITTI数据集语义分割定量结果

Tab. 3　Quantitative results of semantic segmentation of SemanticKITTI dataset　/%

方法	投影&体素							点				
模型	Squeeze Seg[27]	Squeeze SegV2[4]	S-BKI[24]	Range Net++[28]	Lattice Net[5]	Polar Net[29]	Salsa Next[30]	Point Net[8]	SPG[22]	Pointnet ++[9]	Rand LA-Net[11]	本文
参数量(M)	1	1	—	50	—	14	6.73	3	0.25	6	1.24	4.9
mIoU	29.5	39.7	51.3	52.2	52.2	54.3	59.5	14.6	17.4	20.1	53.9	55.3
car	68.8	81.8	83.8	91.4	88.6	83.8	91.9	46.3	49.3	53.7	94.2	94.6
bicycle	16	18.5	30.6	25.7	12	40.3	48.3	1.3	0.2	1.9	26	31.7
motorcycle	4.1	17.9	43	34.4	20.8	30.1	38.6	0.3	0.2	0.2	25.8	34.9
truck	3.3	13.4	26	25.7	43.3	22.9	38.9	0.1	0.1	0.9	40.1	37.1
Other-vehicle	3.6	14	19.6	23	24.8	28.5	31.9	0.8	0.8	0.2	38.9	33.5
person	12.9	20.1	8.5	38.3	34.2	43.2	60.2	0.2	0.3	0.9	49.2	46.1
bicyclist	13.1	25.1	3.4	38.8	39.9	40.2	59	0.2	2.7	1	48.2	50.2
motorcyclist	0.9	3.9	0	4.8	60.9	5.6	19.4	0	0.1	0	7.2	5.6
road	85.4	88.6	92.6	91.8	88.8	90.8	91.7	61.6	45	72	90.7	91.5
parking	26.9	45.8	65.3	65	64.6	61.7	63.7	15.8	0.6	18.7	60.3	61.4
sidewalk	54.3	67.6	77.4	75.2	73.8	74.4	75.8	35.7	28.5	41.8	73.7	74.9
Other-ground	4.5	17.7	30.1	27.8	25.6	21.7	29.1	1.4	0.6	5.6	20.4	24.7
building	57.4	73.7	89.7	87.4	86.9	90	90.2	41.4	64.3	62.3	86.9	89.5
fence	29	41.1	63.7	58.6	55.2	61.3	64.2	12.9	20.8	16.9	56.3	59.9
vegetation	60	71.8	83.4	80.5	76.4	84	81.8	31	48.9	46.5	81.4	84.3
trunk	24.3	35.8	64.3	55.1	57.9	65.5	63.6	4.6	27.2	0.9	61.3	58.7
terrain	53.7	60.2	67.4	64.6	54.7	67.8	66.5	17.6	24.6	30	66.8	68.4
pole	17.5	20.2	58.6	47.9	41.5	51.8	54.3	2.4	15.9	6	49.2	51.5
Traffic-sign	24.5	36.3	67.1	55.9	42.7	57.5	62.1	3.7	0.8	8.9	47.7	53.5

形上取得最优的分割结果.本文算法在基于点的方法中具备显著的优越性,并且在基于投影和体素的方法中也表现出一定的优势,仅次于SalsaNext算法.虽然有些对比算法在SemanticKITTI数据集上在更多标签上表现优秀,但它们会牺牲模型大小和计算效率.本文算法的目标是提供一个在自动驾驶领域具有实际应用潜力的算法,综合考虑了多个因素的平衡,在具备一定实时性的同时,具有较为先进的分割结果.

本文算法在SemanticKITTI数据集上的分割结果可视化结果具体如图8所示.从左到右分别是真实标签、SqueezeSegV2预测结果、RandLA-Net预测结果和本文算法预测结果.从图中可以看出,本文在车的预测上与真实标签最为接近,在植被范围和地形边缘上也有很好的分割效果.从可视化分析可以看出,即使在点云密度比较稀疏的大型室外场景数据集,本文仍能取得较好的分割效果,体现出本文网络特征提取能力的有效性.

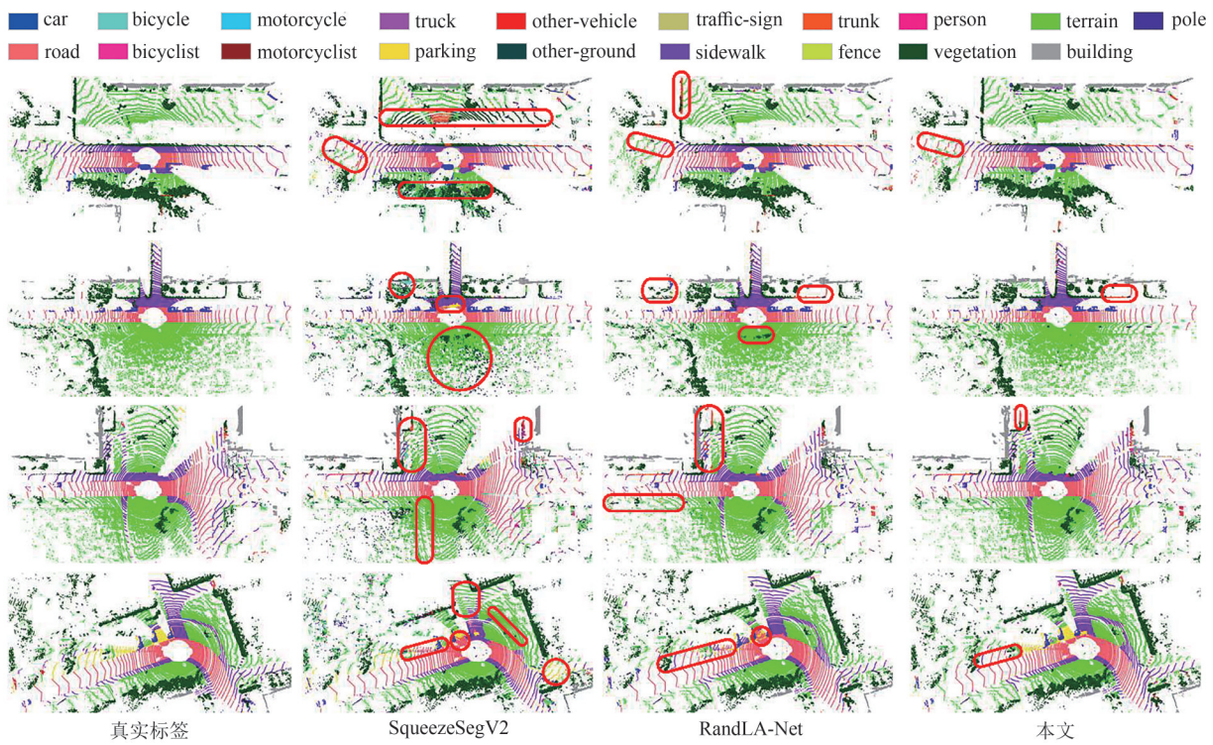

图8 SemanticKITTI数据集语义分割可视化

Fig. 8 SemanticKITTI dataset semantic segmentation visualization

2.6 消融实验

(1)网络结构分析.

为了验证提出的CFSA和HFF模块的有效性,在相同的网络框架下通过对逐个模块调整进行测试,并在S3DIS数据集上进行评估.如表4所示,在任何模块都没有添加时,mIoU仅为68.5%.当单独选用CFSA和HFF模块时,mIoU分别提高了1.8%和0.5%,达到了70.3%和69.0%.当把两个模块一起加入,在两个模块共同作用下,mIoU提高了2.6%,达到了71.1%.通过该消融实验的结果,证明了所提出模块在提取特征时的关键作用.

(2)自注意力机制方式选择.

表5展示了在S3DIS数据集上不同自注意力机制的消融实验结果,通过在构建的局部特征提取模块中分别加入通道自注意力(Channel self-attention,CSA)机制、空间自注意力(Spatial self-attention,SSA)机制、空间和通道并行作用的双通道自注意力

表4 网络结构分析

Tab. 4 Network structure analysis /%

CFSA	HFF	mIoU(S3DIS)
		68.5
√		70.3
	√	69.0
√	√	71.1

表5 自注意力机制选择

Tab. 5 Self-attention mechanism selection /%

CSA	SSA	DCSA	CFSA	mIoU(S3DIS)
√				69.5
	√			70.0
		√		70.6
			√	71.1

(Dual-channel self-attention, DCSA)机制和CFSA机制,评估了这些不同自注意力机制对点云语义分割性能的影响.从表中的结果可以看出,CFSA机制取得了最好的效果,证明了该机制的有效性.

3　结论

本文提出了一个新的算法来处理点云语义分割任务,通过改进的自注意力机制,提升网络特征提取能力.具体地,提出了CFSA模块和HFF模块对算法进行优化.CFSA模块通过在特征提取中交互融合坐标和特征信息,解决了点云几何建模不完整问题.HFF模块通过在编码部分帮助模型充分整合不同层次和尺度之间的语义信息,可以获取更加丰富的特征信息.与其它算法相比,本文所提出的算法在大规模语义分割任务上取得了更好的性能.从预测的可视化图像可以看出,提出的算法能够适应目标形状、结构和外观的变化,并在复杂场景中准确分割点云,表现出强大的适应性和泛化能力.

参　考　文　献

[1] 双丰, 黄兴文, 李勇, 等. 基于深度学习的大规模点云语义分割方法综述[J]. 测绘科学, 2023, 48(2): 195-209.

[2] 侯伟鹏, 王蕾. 基于全局上下文注意力的点云语义分割[J]. 现代电子技术, 2023, 46(9): 120-125.

[3] 于魁梧, 宋玉琴, 徐轩. 基于双注意力融合和残差优化的点云语义分割[J]. 国外电子测量技术, 2022, 41(8): 12-18.

[4] WU B C, ZHOU X Y, ZHAO S C, et al. SqueezeSegV2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a LiDAR point cloud [C]//2019 International Conference on Robotics and Automation (ICRA). Montreal: ACM, 2019: 4376-4382.

[5] ROSU R A, SCHÜTT P, QUENZEL J, et al. LatticeNet: Fast spatio-temporal point cloud segmentation using permutohedral lattices [J]. Autonomous Robots, 2022, 46(1): 45-60.

[6] GAN L, ZHANG R, GRIZZLE J W, et al. Bayesian spatial kernel smoothing for scalable dense semantic mapping[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 790-797.

[7] ZHOU W, ZHANG X D, HAO X X, et al. Multi Point-Voxel Convolution (MPVConv) for deep learning on point clouds[J]. Computers & Graphics, 2023, 112: 72-80.

[8] CHARLES R Q, HAO S, MO K C, et al. PointNet: Deep learning on point sets for 3D classification and segmentation [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 77-85.

[9] 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. Long Beach: ACM, 2017: 5105-5114.

[10] THOMAS H, QI C R, DESCHAUD J E, et al. KPConv: Flexible and deformable convolution for point clouds [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019: 6410-6419.

[11] HU Q Y, YANG B, XIE L H, et al. Learning semantic segmentation of large-scale point clouds with random sampling [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(11): 8338-8354.

[12] ENGEL N, BELAGIANNIS V, DIETMAYER K. Point transformer[J]. IEEE Access, 2005, 9: 134826-134840.

[13] GUO M H, CAI J X, LIU Z N, et al. PCT: Point cloud transformer[J]. Computational Visual Media, 2021, 7(2): 187-199.

[14] REN D Y, WU Z Y, LI J W, et al. Point attention network for point cloud semantic segmentation[J]. Science China Information Sciences, 2022, 65(9): 192104.

[15] ZENG Z Y, XU Y Y, XIE Z, et al. LEARD-Net: Semantic segmentation for large-scale point cloud scene [J]. International Journal of Applied Earth Observation and Geoinformation, 2022, 112: 102953.

[16] GU F, BURLUTSKIY N, ANDERSSON M, et al. Multi-resolution networks for semantic segmentation in whole slide images[C]//International Workshop on Ophthalmic Medical Image Analysis, International Workshop on Computational Pathology. Cham: Springer, 2018: 11-18.

[17] ZHAO H S, SHI J P, QI X J, et al. Pyramid scene parsing network [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017: 6230-6239.

[18] HUANG Z T, YU Y K, XU J W, et al. PF-net: Point fractal network for 3D point cloud completion[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 7659-7667.

[19] LI D W, SHI G L, WU Y H, et al. Multi-scale neighborhood feature extraction and aggregation for point cloud segmentation[J]. IEEE Transactions on Circuits and

Systems for Video Technology, 2021, 31(6): 2175-2191.

[20] GENG X X, JI S P, LU M, et al. Multi-scale attentive aggregation for LiDAR point cloud segmentation [J]. Remote Sensing, 2021, 13(4): 691.

[21] 朱芬芬, 王蕾, 刘华. 特征自适应融合插值的点云语义分割算法[J]. 现代电子技术, 2023, 46(12): 175-181.

[22] LANDRIEU L, SIMONOVSKY M. Large-scale point cloud semantic segmentation with superpoint graphs[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4558-4567.

[23] ZHAO H S, JIANG L, FU C W, et al. PointWeb: Enhancing local neighborhood features for point cloud processing [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 5560-5568.

[24] BOULCH A, LE SAUX B, AUDEBERT N. Unstructured point cloud semantic labeling using deep segmentation networks [C]//Proceedings of the Workshop on 3D Object Retrieval. ACM, 2017: 17 - 24.

[25] ZHANG Z Y, HUA B S, YEUNG S K. ShellNet: Efficient point cloud convolutional neural networks using concentric shells statistics [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019: 1607-1616.

[26] WANG L, HUANG Y C, HOU Y L, et al. Graph attention convolution for point cloud semantic segmentation[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach: IEEE, 2019: 10288-10297.

[27] WU B C, WAN A, YUE X Y, et al. SqueezeSeg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LiDAR point cloud [C]//2018 IEEE International Conference on Robotics and Automation (ICRA). Brisbane: IEEE, 2018: 1887-1893.

[28] MILIOTO A, VIZZO I, BEHLEY J, et al. RangeNet: Fast and accurate LiDAR semantic segmentation[C]// 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Macau: IEEE, 2019: 4213-4220.

[29] ZHANG Y, ZHOU Z X, DAVID P, et al. PolarNet: An improved grid representation for online LiDAR point clouds semantic segmentation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020: 9598-9607.

[30] CORTINHAL T, TZELEPIS G, AKSOY E. SalsaNext: Fast semantic segmentation of LiDAR point clouds for autonomous driving[J]. arXiv: 2020, 2003.03653.

(责编&校对 雷建云)