中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2023)02-0483-12

论文引用格式:Liu S, Cao Y F, Huang W H and Li D D. 2023. LiDAR point cloud semantic segmentation combined with sparse attention and instance enhancement. Journal of Image and Graphics,28(02):0483-0494(刘盛,曹益烽,黄文豪,李丁达. 2023. 融合稀疏注意力和实例增强的雷达点云分割. 中国图象图形学报,28(02):0483-0494)[DOI:10.11834/jig.210787]

融合稀疏注意力和实例增强的雷达点云分割

刘盛*,曹益烽,黄文豪,李丁达

浙江工业大学计算机科学与技术学院,杭州　310023

摘　要:**目的**　雷达点云语义分割是3维环境感知的重要环节,准确分割雷达点云对象对无人驾驶汽车和自主移动机器人等应用具有重要意义。由于雷达点云数据具有非结构化特征,为提取有效的语义信息,通常将不规则的点云数据投影成结构化的2维图像,但会造成点云数据中几何信息丢失,不能得到高精度分割效果。此外,真实数据集中存在数据分布不均匀问题,导致小样本物体分割效果较差。为解决这些问题,本文提出一种基于稀疏注意力和实例增强的雷达点云分割方法,有效提高了激光雷达点云语义分割精度。**方法**　针对数据集中数据分布不平衡问题,采用实例注入方式增强点云数据。首先,通过提取数据集中的点云实例数据,并在训练中将实例数据注入到每一帧点云中,实现实例增强的效果。由于稀疏卷积网络不能获得较大的感受野,提出Transformer模块扩大网络的感受野。为了提取特征图的关键信息,使用基于稀疏卷积的空间注意力机制,显著提高了网络性能。另外,对不同类别点云对象的边缘,提出新的TVloss用于增强网络的监督能力。**结果**　本文提出的模型在SemanticKITTI和nuScenes数据集上进行测试。在SemanticKITTI数据集上,本文方法在线单帧精度在平均交并比(mean intersection over union,mIoU)指标上为64.6%,在nuScenes数据集上为75.6%。消融实验表明,本文方法的精度在baseline的基础上提高了3.1%。**结论**　实验结果表明,本文提出的基于稀疏注意力和实例增强的雷达点云分割方法在SemanticKITTI和nuScenes数据集上都取得了较好表现,提高了网络对点云细节的分割能力,使点云分割结果更加准确。

关键词:激光雷达(LiDAR);语义分割;空间注意力机制;Transformer;深度学习(DL);实例增强

LiDAR point cloud semantic segmentation combined with sparse attention and instance enhancement

Liu Sheng*, Cao Yifeng, Huang Wenhao, Li Dingda

College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou　310023, China

Abstract: **Objective** Outdoor-perceptive recognition is essential for robots-mobile and autonomous driving vehicles applications. LiDAR-based point cloud semantic segmentation has been developing for that. Three-dimensional image-relevant (3D image-relevant) LiDAR can be focused on the range of information quickly and accurately for outdoor-related perception with no illumination effects. To get feasible effects for autonomous driving vehicles, LiDAR point cloud-related semantic segmentation can be predicted in terms of point cloud analysis for overall scene factors like roads, vehicles, pedestrians, and plants. Recent deep learning-based (DL-based) two-dimensional image-relevant (2D image-relevant) computer vision

收稿日期:2021-09-08;**修回日期**:2022-03-10;**预印本日期**:2022-03-17

*通信作者:刘盛　edliu@zjut.edu.cn

基金项目:国家重点研发计划资助(2018YFB1305200);浙江省科技厅公益技术应用研究计划项目(LGG19F020010)

Supported by: National Key R&D Program of China (2018YFB1305200); Science Technology Department of Zhejiang Province (LGG19F020010)

has been developing intensively. Nevertheless, LiDAR point cloud data is featured of unstructured, disorder, sparse and non-uniform densities beyond 2D image-relevant structured data. The challenging issue is to extract semantic information from LiDAR data effectively. DL-based methods can be divided into three categories: 1) point-based, 2) projection-relevant, and 3) voxel-related. To extract effective semantic information, the existing methods are often used to project irregular point cloud data into 2D images-structured because of the unstructured characteristics of LiDAR point cloud data. However, geometric information loss-derived high-precision segmentation results cannot be obtained well. In addition, lower segmentation effect for small sample objects has restricted by uneven data distribution. To resolve these problems, we develop a sparse attention and instance enhancement-based LiDAR point cloud segmentation method, which can improve the accuracy of semantic segmentation of LiDAR point cloud effectively. **Method** An end-to-end sparse convolution-based network is demonstrated for LiDAR point cloud semantic segmentation. To optimize uneven data distribution in the training data set, instance-injected is used to enhance the point cloud data. Instance-injected can be employed to extract its points cloud data factors like pedestrians, vehicles, and bicycles. Instance-related data is injected into an appropriate position of each frame during the training process. Recently, the receptive field-strengthened and attention mechanism-aware visual semantic segmentation tasks are mainly focused on. But, a wider receptive field cannot be realized due to the encoder-decoder-based network ability. A lightweight Transformer module is then illustrated to widen the receptive field of the network. To get global information better, the Transformer module can be used to build up the interconnection between each non-empty voxel. The Transformer module is used in the bottleneck layer of the network for memory optimization. To extract the key positions of the feature map, a sparse convolution-based spatial attention module is proposed as well. Additionally, to clarify the edges of multiple types of point cloud objects, a new TVloss is adopted to identify the semantic boundaries and alleviate the noise within each region-predicted. **Result** Our model is proposed and evaluated on SemanticKITTI dataset and nuScenes dataset both. It achieves 64.6% mean intersection over union (mIoU) in the single-frame accuracy evaluation of SemanticKITTI, and 75.6% mIoU on the nuScenes dataset. The ablation experiments show that the mIoU is improved by 1.2% in terms of instance-injected, and the spatial attention module has an improvement of 1.0% and 0.7% each based on sparse convolution and the transformer module. The efficiency of these two modules is improved a total of 1.5%, the mIoU-based TVloss achieves 0.2% final gain. The integrated analysis of all modules is increased by 3.1% in comparison with the benchmark. **Conclusion** A new sparse convolution-based end-to-end network is developed for LiDAR point cloud semantic segmentation. We use instance-injected to resolve the problem of the unbalanced distribution of data profiling. A wider range of receptive field is achieved in terms of the proposed Transformer module. To extract the key location of the feature map, a sparse convolution-based spatial attention mechanism is melted into. A new TVloss loss function is added and the edge of the objects in point clouds is clarified. The comparative experiments are designed in comparison with recent SOTA (state of the art) methods, including projection and point-based methods. Our proposed method has its potentials for the improved segmentation ability of the network to point cloud details and the effectiveness for point cloud segmentation further.

Key words: LiDAR; semantic segmentation; spatial attention mechanism; Transformer; deep learning (DL); instance enhancement

0 引 言

环境感知是移动机器人和无人驾驶汽车应用的首要任务。主流的环境感知传感器包括相机和激光雷达,相比于视觉传感器,3 维激光雷达能够快速准确地获取周围环境的距离信息并且不受光照影响。3 维点云语义分割算法能够预测出场景中物体的类别,如道路、车辆、行人和植物等。3 维点云分割在无人驾驶、移动机器人以及 VR/AR(virtual reality/augmented reality)等领域都有广泛应用,是计算机视觉的重要研究方向。对周围环境的准确分割是可靠无人驾驶的先决条件。

随着图像分割技术的发展,基于视觉的环境感知方法(Wang 等,2019)取得了优异成绩。但是与结构化的 2 维图像数据不同,雷达点云数据是非结构化的、无序性的、密度不一致的。因为这些特性,神经网络很难从雷达数据中提取有效的语义信息。

深度学习的发展不断推动图像语义分割方法的进步。有研究者(Milioto 等,2019)将图像语义分割

方法应用于点云语义分割。一些方法(Milioto 等,2019;Cortinhal 等,2020;Zhang 等,2020b)将不规则、无序分布的 3 维雷达点云投影成规则的 2 维图像,常用的投影方法有球状投影和鸟瞰图投影,再通过成熟的 2 维卷积方式对投影图像进行语义分割,最后将完成分割的投影图像反投影回 3 维空间。但是这些方法在 3 维到 2 维投影过程中会导致几何信息丢失,不能得到较高的分割精度。基于点的方法(Charles 等,2017a;Hu 等,2020)能通过较少的网络参数实现点云语义分割,但是在处理大场景点云过程中会消耗过多的计算资源,也不能得到较高精度的分割结果。Graham 等人(2018)针对传统稠密体素中存在的稀疏性问题,提出基于子流形稀疏卷积网络(submanifold sparse convolutional networks),通过将非空体素特征值和空间坐标建立哈希关系,仅对非空体素进行卷积操作,大幅降低了内存和计算资源的消耗。随着稀疏卷积(Choy 等,2019)的提出和应用,极大提高了基于体素的点云语义分割方法的效率。

视觉语义分割精度的提高主要依靠扩大感受野和使用注意力机制。感受野是语义分割任务中非常重要的一个因素,然而基于编码器的网络无法提供更大的感受野。SDRNet(spatial depthwise residual network)(Liu 等,2021)通过扩张特征整合模块(dilate feature aggregation,DFA)扩大网络感受野以提升全局特征提取能力。同时,SalsaNext(Cortinhal 等,2020)通过引入空洞卷积来扩大感受野。目前,注意力机制已广泛应用于图像分类、目标识别(Li 等,2020)和语义分割(Yu 和 Wang,2021)等领域,可以有效提取局部和全局特征。在点云上,DGCSA(dynamic graph convolution with spatial attention)(Song 等,2021)结合空间注意力模块与动态图卷积模块,取得更加精确的点云分类分割效果。Du 和 Cai(2021)提出一种基于多特征融合与残差优化的点云语义分割方法,并且引入注意力机制来提高点云聚合能力。但这些引入注意力的方法都是在小规模点云数据中应用,针对室外大场景雷达点云工作相对较少。Transformer 也是注意力机制的一种,最先提出用于自然语言处理(Vaswani 等,2017)。Transformer 在图像处理上(Liu 等,2021)取得巨大成功,但在点云上的应用(Guo 等,2021)相对较少。Transformer 位置排列不变的特点非常适合处理无序的点云数据。

此外,现有方法主要集中在对网络的改进,很少关注输入数据本身。但真实场景数据集中,存在点云数据分布不平衡现象,导致样本数量稀少的类别特征被抑制,使点云数量稀少的类别不能得到较好的预测结果。

为解决上述问题,本文提出一种基于稀疏注意力和实例增强的激光雷达点云分割的方法。本文主要贡献如下:1)采用点云实例注入的方式,提取训练数据数量稀少的类别的实例点云信息,并将其在训练过程中注入到每一帧点云的合适位置,减少了点云数量不平衡带来的精度下降问题。2)在卷积网络的瓶颈层(bottleneck layer)加入 Transformer 模块,扩大网络的感受野,通过建立点云远、近距离的上下文特征关联,有效提高了提取点云局部和全局信息的能力。3)提出一种基于稀疏卷积的空间注意力机制,通过提取特征图中代表性的局部关键信息,增强网络对关键特征的关注。4)为了增加语义分割的精度,提出一种新的 TVLoss 来增强网络对不同类别点云对象边缘的监督。

1　国内外研究现状

基于数据驱动的深度学习算法成为计算机视觉领域的重要研究方向。深度学习在 2 维图像的分类、检测和分割等领域取得优异成果。但是 3 维点云标注比 2 维图像费时费力,起步相对较晚。得益于 SemanticKITTI 数据集(Behley 等,2019)的出现,雷达点云语义分割工作相继涌现,分为基于点的方法、基于投影的方法和基于体素的方法。

基于点的方法不需要对点云进行前期处理,而是直接对点云进行操作。PointNet(Charles 等,2017a)最先提出用 MLP(multilayer perceptron)的方法对每个点进行处理,用于点云的分类和分割任务,但不能有效提取局部信息。PointNet++(Charles 等,2017b)在 PointNet(Charles 等,2017a)上加入局部特征提取模块。RandLA-Net(Hu 等,2020)针对大规模点云数据,用随机采样代替最远点采样降低计算量,提高语义分割效率,同时使用局部特征聚合方法减少随机采样带来的信息损失。SDRNet(Liu 等,2021)提出结合 SDR(spatial depthwise residual)模块和 DFA 模块的点云分割网络。SDRNet 针对点

云旋转不变性设计了改进的 SDR 模块,用于提取局部特征,消除雷达点云数据 Z 轴旋转对分割结果的影响;DFA 模块提高了网络感受野。基于点的方法能通过较少的网络参数实现点云语义分割,但在处理大场景点云过程中会消耗过多的计算资源,也不能得到较高精度的分割结果。

基于投影的方法在点云分割中受到广泛研究,主要投影方式有球状投影和鸟瞰图投影。RangeNet ++(Milioto 等,2019)最先提出通过球状投影的方法将 3 维点云投影到 2 维空间,利用 2 维卷积对投影图像进行语义分割,并在反投影过程中利用 KNN(k-nearest neighbor)进行空间邻域搜索,提高语义分割精度。SalsaNext(Cortinhal 等,2020)在 SalsaNet(Aksoy 等,2020)基础上引入空洞卷积和新的局部特征提取模块,获得更好的分割效果。Zheng 等人(2021)根据人类观察机制提出一种场景视点偏移方法,改善了因投影过程中信息丢失带来的精度下降问题。PolarNet(Zhang 等,2020b)提出一种鸟瞰图投影方法,根据激光雷达本身特点建立基于极坐标的空间雷达划分方法。相比传统基于欧氏空间的体素划分方法,基于投影的方法能减少空体素出现的概率,不同类别的点云出现在同一个体素内的概率也相应减少,但在 3 维到 2 维的投影过程中会损失场景对象的几何信息,不能得到较高的分割精度。

基于体素的方法(Zhou 和 Tuzel,2018;Wang 等,2019)将点云划分到不同体素中,并用传统卷积进行语义分割。但是因为激光雷达的稀疏性,导致大部分体素冗余,这会增加内存和计算资源的消耗。Tao 等人(2021)提出一种基于稀疏体素金字塔的多尺度点云特征提取方法,提高了点云特征提取的效率。随着稀疏卷积(Choy 等,2019)的提出和应用,极大提高了基于体素的点云语义分割方法的效率。Cylinder3d(Zhu 等,2022)采用稀疏卷积和柱状划分网格的方法在点云语义分割精度上取得了显著效果,提出的不对称残差(asymmetrical residual)卷积模块能够更好地提取类似于车辆、行人周围的语义信息,最终提高此类物体的语义分割精度。但上述方法为提高精度需要不断增加体素的分辨率,导致内存增加和计算资源消耗。

2 基于稀疏注意力和实例增强的雷达点云分割算法流程

本文提出的基于稀疏注意力和实例增强的雷达点云分割方法主要由实例注入模块、点特征提取模块和稀疏卷积模块 3 部分组成,总体架构如图 1 所示。

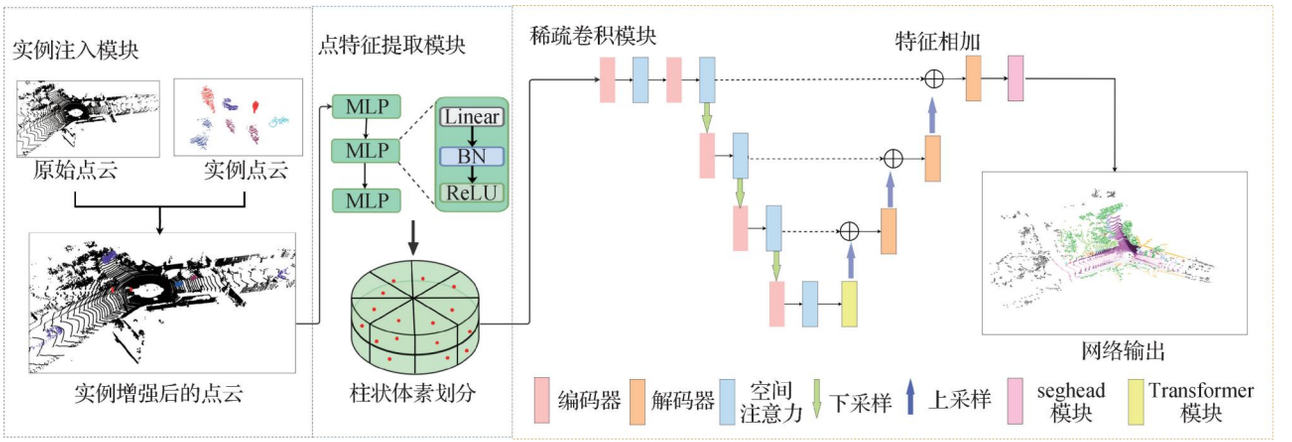


图 1 网络结构总览

Fig. 1 The overview of proposed network

2.1 网络总体架构

1)实例注入模块。网络输入数据由原始点云数据与从数据集中筛选的实例数据组合得到,并通过柱状坐标对点云进行体素划分。其中,每个点的特征包括点云欧氏空间坐标 $\{x,y,z\}$、极坐标 $\{\rho,\theta\}$、每个点云到体素中心的偏移量 $\{\Delta\rho,\Delta\theta,\Delta z\}$ 和点云的反射率 $\{i\}$,共 9 维特征。

2)点特征提取模块。在本模块中,通过多个 MLP 层从输入数据提取出每个点的特征。在每个体素中,只保留特征值最大的数据作为当前体素的

代表特征。由特征信息 $\boldsymbol{F}=\{f_1, f_2, f_3, f_4, f_5, \cdots, f_N\}$ 和对应的柱状体素坐标 $\boldsymbol{V}=\{v_1, v_2, v_3, v_4, v_5, \cdots, v_N\}$ 构建稀疏卷积张量，其中 N 表示非空体素数量，f_i 和 v_i 分别表示第 i 个特征点的特征值和柱状体素坐标。

3）稀疏卷积模块。主干网络采用类似 UNet 的设计，通过下采样不断降低特征图尺寸，获得网络更深层特征，同时通过上采样将特征图恢复到原有尺寸。通过跳跃连接将上采样的特征图与相同大小的浅层特征图相融合。在本文所提网络结构中，每一层的特征图尺寸和通道数分别为[480 × 360 × 32, 32]，[240 × 180 × 16, 64]，[120 × 90 × 8, 128]，[60 × 45 × 8, 256]和[30 × 23 × 8, 512]。最终，利用 seghead 子模块预测每个体素中每个类别的概率值，选择概率值最大的类别作为该体素预测的结果，将体素的语义信息转化为点云的语义信息。

2.2　实例注入

真实场景采集的数据集中存在类别数量不平衡现象。图 2 为 SemanticKITTI 训练数据集中不同类别点云的数量。

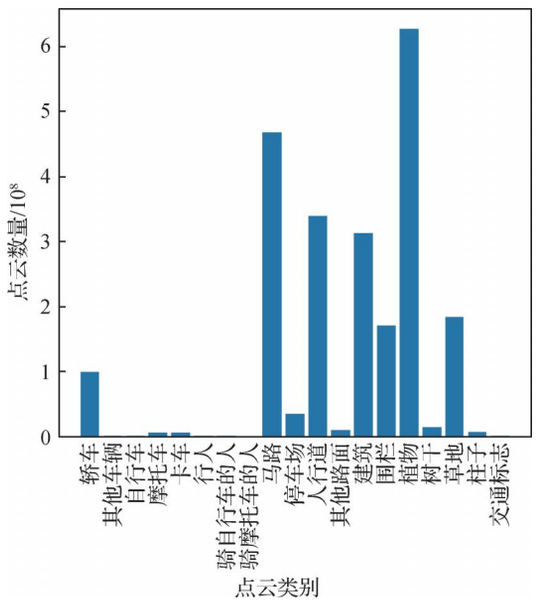


图 2　SemanticKITTI 训练数据集中不同类别点云的数量

Fig. 2　The number of point clouds of different categories on the SemanticKITTI training dataset

从图 2 可以看出，一些类别如自行车、摩托车、骑自行车的人和骑摩托车的人等的点云样本数量占整个数据集的比重非常少。这些类别的特征在网络训练过程中会被其他类别的特征淹没，通常不能得到较好的分割结果。但是在无人驾驶场景中，行人和车辆正确的语义分割极其重要，直接影响驾驶安全。本文采用实例注入的方法来减少这个问题。SemanticKITTI 数据集提供了点云的实例和语义信息。根据提供的实例 ID（identity），可以从训练集的点云数据中筛选出对应的实例点云，随后将筛选的实例数据保存到文件，并在文件名中标记语义类别。实例数据包括自行车、摩托车、卡车、其他车辆、行人、骑自行车的人和骑摩托车的人，共 6 类。

实例注入策略如下：在训练过程中，首先随机选择道路类别（包括马路、停车场和人行道）中的一个点作为基准点，并计算以基准点为圆心，半径为 R 的基准范围内所有点云的高度差。若高度差小于阈值 *threshold*，说明道路上不存在其他障碍物，此时随机选择一个实例数据插入基准位置，并且使基准范围最高点与实例数据最低点对齐；若高度差大于阈值 *threshold*，则重新选择基准点。

本文实验参数半径 R 设置为 1.5 m，高度阈值 *threshold* 设置为 0.3 m，每一帧点云注入 7 个实例数据。最终实例注入效果如图 3 所示，图中黑色的为原始点云，彩色的点云区域为注入的多个实例，颜色与实例类别对应。

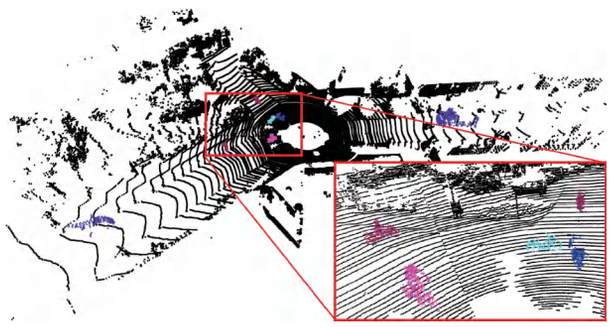


图 3　实例注入

Fig. 3　Instance injection

2.3　稀疏卷积模块和空间注意力机制

空间注意力机制能区别对待特征图的不同信息，使网络增加对关键特征的关注，提高网络性能。在不对称残差模块（asymmetrical residual block）后添加基于稀疏卷积的空间注意力机制。网络结构如图 4 所示。不对称残差模块通过交叉进行 1 × 3 × 3 和 3 × 1 × 3 的稀疏卷积提高了提取特征的效率。稀疏空间注意力机制通过 sigmoid 函数筛选

特征值较大的点作为注意力掩膜,最后将学习到的注意力掩膜与原始的特征图元素相乘,以达到提取关键特征的目的。解码器模块有两部分输入,一部分是上一层网络上采样得到的特征图,另一部分是通过跳跃连接得到的浅层网络特征,两者有相同的特征图尺寸和通道数。本文通过元素相加的方式将两部分特征相融合。

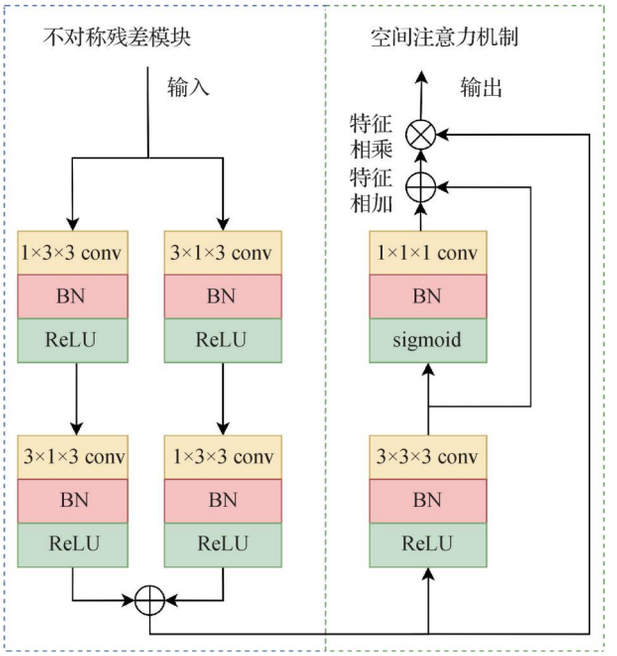


图 4　不对称残差卷积模块和空间注意力机制

Fig. 4　Asymmetrical residual block and spatial attention mechanism

2.4　Transformer 模块

Transformer 模块网络结构如图 5 所示,该模块输入为体素的位置和特征信息,其中 N 为非空体素的数量。首先对位置信息进行位置编码(position embedding),计算得到点与点之间的位置关系特征,然后将得到的位置信息与特征信息相融合。融合后的信息作为自注意力机制的输入,计算得到自注意力掩膜。自注意力机制可以表示为

$$f_{\text{Attention}}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = f_{\text{softmax}}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d}} + \boldsymbol{B}\right)\boldsymbol{V} \tag{1}$$

式中,$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V} \in \mathbf{R}^{M^2\times d}$ 分别表示 query,key 和 value 矩阵,M^2 为 patches 的数量,d 表示 query 或 key 矩阵的维度,$\boldsymbol{B}$ 为偏差矩阵(bias matrix)。将 query、key 和 value 矩阵分别经过线性网络,再用 query 和 key 矩阵的转置矩阵相乘得到 attention map。最后将 attention map 经过 softmax 后与 value 矩阵相乘作为模块的输出。

Transformer 模块采用自注意力机制的方式,通过 attention map 对点与点之间的连接赋予不同权重,增强了每个点之间的联系。Transformer 模块可以扩大网络的感受野,使网络有效提取局部和全局特征。

2.5　损失函数

本文损失函数将权重交叉熵 Lovasz-softmax 和 TVloss 线性融合得到最终的 L_{total},表示方式为

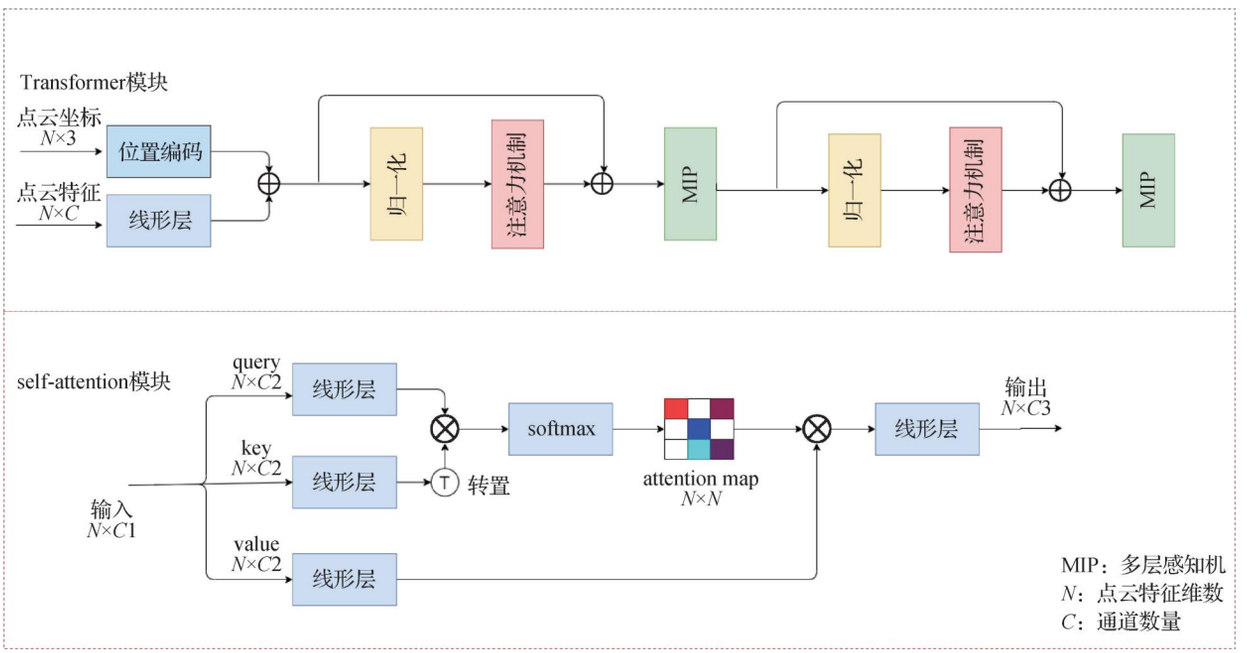


图 5　Transformer 模块

Fig. 5　Transformer module

$$L_{\text{total}} = L_{\text{wce}} + L_{\text{ls}} + L_{\text{tv}} \tag{2}$$

式中，权重交叉熵损失函数（He 等，2016）可以表示为

$$L_{\text{wce}}(y,\hat{y}) = -\sum_i \alpha_i p(y_i)\log(p(\hat{y}_i)) \tag{3}$$

式中，$\alpha_i = 1/\sqrt{f_i}$，y_i 和 $\hat{y}_i$ 分别表示每个类别的真值和预测值，f_i 表示第 i 类别出现的频率，在点云的语义分割中表示第 i 类点云的数量占总点云数量的比重。

Lovasz-softmax 损失函数（Berman 等，2018）已广泛应用在语义分割领域（Zhu 等，2022；Cortinhal 等，2020），其损失函数可以表示为

$$L_{\text{ls}} = \frac{1}{|C|}\sum_{c\in C} J(m(c)) \tag{4}$$

$$m_i(c) = \begin{cases} 1 - x_i(c) & c = y_i(c) \\ x_i(c) & \text{其他} \end{cases} \tag{5}$$

式中，$|C|$ 表示类别数量，J 是一个具有全局最小值分段线性函数。$\boldsymbol{m}(c)$ 表示 c 类别的误差向量。$x_i(c)$ 和 $y_i(c)$ 分别表示 c 类别的第 i 个点云的预测值和真值。Lovasz-softmax loss 是一个有效的附加损失函数，可用于不同的深度学习任务，例如目标检测和语义分割。因此，在训练模型过程中，Lovasz-softmax loss 与其他损失函数相结合可以实现更好的模型训练效果。

受 Gerdzhev 等人（2021）的启发，为增加网络对点云边缘的监督，本文添加了 TVLoss。原方法中直接对真值边缘和预测边缘取绝对值。这会由于边缘类别的不同导致不同的损失值，为避免这种情况，本文采用异或操作。具体函数表示为

$$\begin{aligned} L_{\text{tv}}(y,\hat{y}) = \sum_{i,j,k} & \left(\left(|y_{i+1,j,k} - y_{i,j,k}| \otimes |\hat{y}_{i+1,j,k} - \hat{y}_{i,j,k}| \right) + \right. \\ & \left(|y_{i,j+1,k} - y_{i,j,k}| \otimes |\hat{y}_{i,j+1,k} - \hat{y}_{i,j,k}| \right) + \\ & \left. \left(|y_{i,j,k+1} - y_{i,j,k}| \otimes |\hat{y}_{i,j,k+1} - \hat{y}_{i,j,k}| \right) \right) \end{aligned} \tag{6}$$

式中，$\otimes$ 表示异或操作，y 和 $\hat{y}$ 分别表示点云的真值和预测值，i、j 和 k 表示体素的 3 个维度。

3　实验和结果分析

实验在 SemanticKITTI（Behley 等，2019）和 nuScenes（Caesar 等，2020）数据集上进行，将本文算法与近年有代表性的方法比较，进行详细的精度评估，并对提出的各模块进行消融实验。

3.1　数据集及评价方式

3.1.1　SemanticKITTI 数据集

SemanticKITTI 数据集（Behley 等，2019）是一个无人驾驶场景的雷达点云数据集，面向点云语义分割和实例分割等多类任务。数据集通过 Velodyne-HDL64E 激光雷达在德国进行数据采集，每一帧点云大约有 12 万个点，共 22 个序列。实验时，将 00—07、09—10 序列共 19 130 帧点云数据作为训练集，将 08 序列共 4 071 帧点云数据作为验证集，将 11—21 序列共 20 351 帧点云数据作为测试集。点云的语义分割任务共 19 个类别。

3.1.2　nuScenes 数据集

nuScenes 数据集（Caesar 等，2020）共 1 000 个场景，每个场景持续 20 s。数据集使用 Velodyne-HDL32E 激光雷达在美国波士顿和新加坡采集数据，采样周期为 20 Hz。每个场景包含 40 个关键帧，共包含 40 000 个关键帧的点云数据。官方将点云数据划分成测试集、验证集和训练集，其中 850 个场景用于训练和验证，150 个场景作为测试。点云的语义分割任务共 16 个类别。

3.1.3　评价指标

实验采用官方提供的评价方法（Behley 等，2019；Caesar 等，2020），以 mIoU（mean intersection over union）作为精度的评价指标。mIoU 可表示为

$$f_{\text{mIoU}} = \frac{TP}{TP + FP + FN} \tag{7}$$

式中，TP 表示真阳性，FP 表示假阳性，FN 表示假阴性。

3.1.4　实验参数设置

实验中，两个数据集柱状体素的划分尺寸都是 [480 × 360 × 32]，分别表示距离、角度和高度 3 个维度。实验环境为 Ubuntu16.04、Pytorch1.4、Intel i7 - 8700K、NVIDIA RTX。使用 Adam 优化器，学习率为 0.001，batch size 为 2，共训练 40 个 epochs。由于 nuScenes 数据集未提供实例信息，实验没有使用实例注入。

3.2　SemanticKITTI 数据集实验结果

在 SemantKITTI 数据集的测试集上，将本文方法与代表性方法进行比较。其中，基于点的方法包括 TangentConv（Tatarchenko 等，2018）、RandLA-Net（Hu 等，2020）、KPConv（Thomas 等，2019）和 SDR-Net（Liu 等，2021），基于投影的方法包括 Darknet53

(Behley 等, 2019)、SqueezeSegv3 (Xu 等, 2020)、RangeNet + + (Milioto 等,2019)、Salsanext(Cortinhal 等,2020)、KPRNet(Kochanov 等,2020)和 PolarNet(Zhang 等,2020b),基于体素的方法包括 FusionNet(Zhang 等, 2020a)、TORANDONet (Gerdzhev 等, 2021)和 Cylinder3d(Zhu 等,2022)。定量实验结果如表 1 所示。可以看出,本文方法在 mIoU 指标上取得较好成绩,特别是在行人和汽车类别上取得了优秀的表现,相比 Cylinder3d,卡车(truck)提高了 8.6% ,其他车辆(other-vehicle)提高了 4.4% 。原因是本文提出的空间注意力机制和 TVloss 能很好地处理边缘细节信息。

在 SemanticKITTI 验证集上的点云语义分割结果的可视化如图 6 所示,从上到下分别是真值、预测图和细节图。从细节图可以看出,本文方法可以准确预测行人和自行车类别(图中红框)。

表 1　SemanticKITTI 测试集精度结果对比

Table 1　Accuracies results of different methods on the SemanticKITTI test set

/%

方法	mIoU	车 1	车 2	车 3	车 4	车 5	人 1	人 2	人 3	地 1	地 2	地 3	地 4	建筑	围栏	植物	树干	草地	柱子	交通标志
TangentConv	35.9	86.8	1.3	12.7	11.6	10.2	17.1	20.2	0.5	82.9	15.2	61.7	9.0	82.8	44.2	75.5	42.5	55.5	30.2	22.2
Darknet53	49.9	86.4	24.5	32.7	25.5	22.6	36.2	33.6	4.7	91.8	64.8	74.6	27.9	84.1	55.0	78.3	50.1	64.0	38.9	52.2
RandLA-Net	50.3	94.0	19.8	21.4	42.7	38.7	47.5	48.8	4.6	90.4	56.9	67.9	15.5	81.1	49.7	78.3	60.3	59.0	44.2	38.1
RangeNet ++	52.2	91.4	25.7	34.4	25.7	23.0	38.3	38.8	4.8	91.8	65.0	75.2	27.8	87.4	58.6	80.5	55.1	64.6	47.9	55.9
PolarNet	54.3	93.8	40.3	30.1	22.9	28.5	43.2	40.2	5.6	90.8	61.7	74.4	21.7	90.0	61.3	84.0	65.5	67.8	51.8	57.5
SqueezeSegv3	55.9	92.5	38.7	36.5	29.6	33.0	45.6	46.2	20.1	91.7	63.4	74.8	26.4	89.0	59.4	82.0	58.7	65.4	49.6	58.9
Salsanext	59.5	91.9	48.3	38.6	38.9	31.9	60.2	59.0	19.4	91.7	63.7	75.8	29.1	90.2	64.2	81.8	63.6	66.5	54.3	62.1
KPConv	58.8	96.0	32.0	42.5	33.4	44.3	61.5	61.6	11.8	88.8	61.3	72.7	31.6	**95.0**	64.2	84.8	69.2	69.1	56.4	47.4
SDRNet	59.1	95.4	42.3	46.0	43.2	41.0	61.4	55.5	11.5	91.1	64.5	75.7	23.8	90.7	61.6	80.8	63.8	65.0	54.7	54.6
FusionNet	61.3	95.3	47.5	37.7	41.8	34.5	59.5	56.8	11.9	91.8	68.8	77.1	**30.8**	92.5	**69.4**	84.5	69.8	68.5	60.4	**66.5**
KPRNet	63.1	95.5	54.1	47.9	23.6	42.6	65.9	65.0	16.5	**93.2**	**73.9**	**80.6**	30.2	91.7	68.4	**85.7**	69.8	71.2	58.7	64.1
TORANDONet	63.1	94.2	55.7	48.1	40.0	38.2	63.6	60.1	**34.9**	89.7	66.3	74.5	28.7	91.3	65.6	85.6	67.0	71.5	58.0	65.9
Cylinder3d	63.9	96.7	**60.3**	**57.4**	43.2	49.6	70.0	65.1	12.0	91.6	64.6	76.0	24.3	90.0	63.4	84.8	**70.7**	67.5	62.1	64.0
本文	**64.6**	**96.8**	59.3	57.2	**51.8**	**54.0**	**70.2**	**66.5**	11.9	91.7	63.9	76.1	25.8	89.6	62.3	85.0	70.2	**67.9**	**62.4**	64.5

注:加粗字体表示各列最优结果。车 1—车 5 分别表示轿车、自行车、摩托车、卡车以及其他车辆;人 1—人 3 分别表示行人、骑自行车者和骑摩托车者;地 1—地 4 分别表示马路、停车场、人行道以及其他地面。

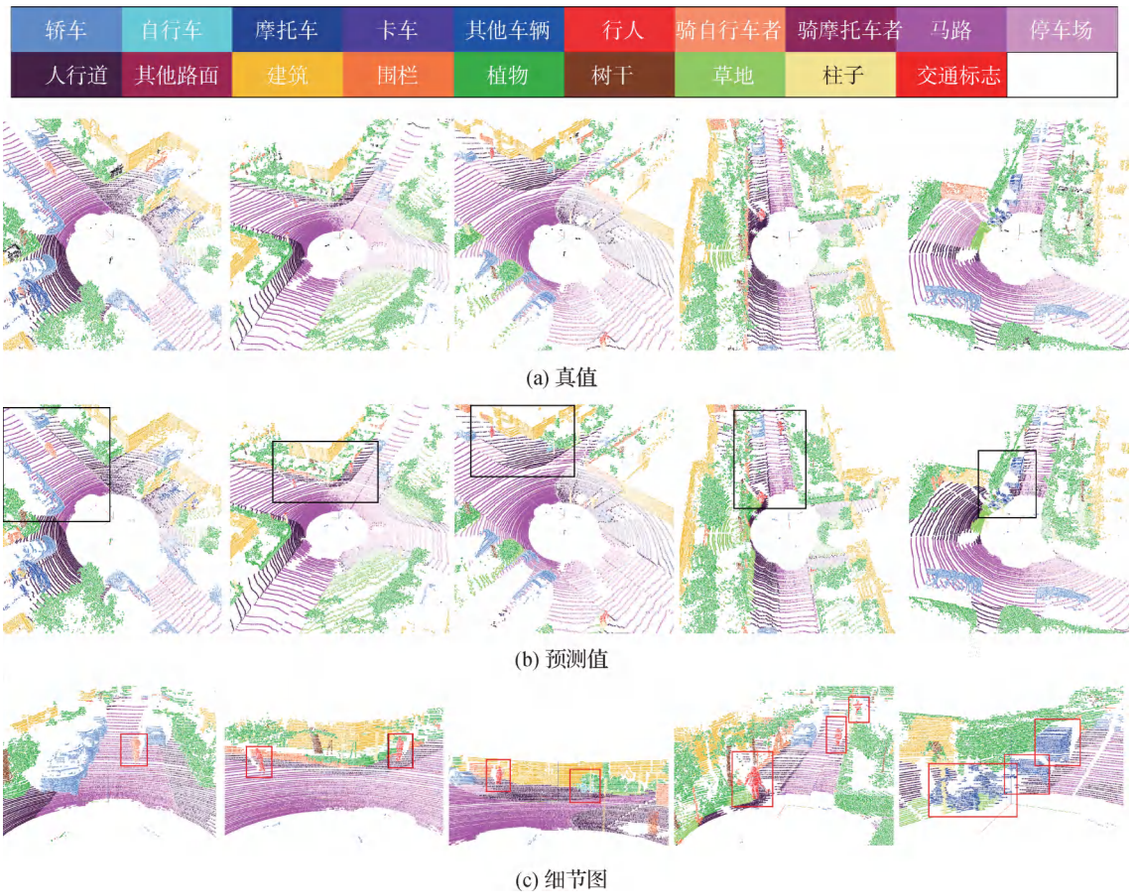
3.3　nuScenes 数据集实验结果

由于 nuScenes 数据集没有发布验证测试集精度的方法,因此在验证集上进行精度的定量分析。将本文方法与 RangeNet ++ (Milioto 等,2019)、PolarNet(Zhang 等, 2020b)、Salsanext (Cortinhal 等, 2020)和 Cylinder3d(Zhu 等,2022)进行定量比较,结果如表 2 所示。可以看出,本文方法在 mIoU 上取得了非常优秀的成绩,特别是在汽车和行人类别中,语义分割性能有显著提升。得益于稀疏卷积网络有效提取全局和局部特征,本文方法相比于基于投影的方法,mIoU 提高了 3% ～10% 。相比于基于体素的方法,本文提出的空间注意力机制可以对关键特征进行关注,Transformer 模块增强网络提取全局信息的能力。

nuScenes 验证集语义分割结果的可视化如图 7 所示,可以看出,本文方法可以准确预测激光雷达获取的点云数据。

3.4　消融实验

为评估实例注入、空间注意力机制、Transformer 和 Tvloss 等模块对网络的贡献,在 SemanticKITTI 的验证集上进行消融实验,其中,体素尺寸大小为[240 × 240 × 32],结果如表 3 所示。可以看出,



(a) 真值

(b) 预测值

(c) 细节图

图 6　SemanticKITTI 验证集语义分割结果

Fig. 6　Semantic segmentation results on the SemanticKITTI validation set((a) ground truth; (b) prediction; (c) detail images)

表 2　nuScenes 验证集精度结果对比

Table 2　Accuracies results of different methods on the nuScenes validation set

/%

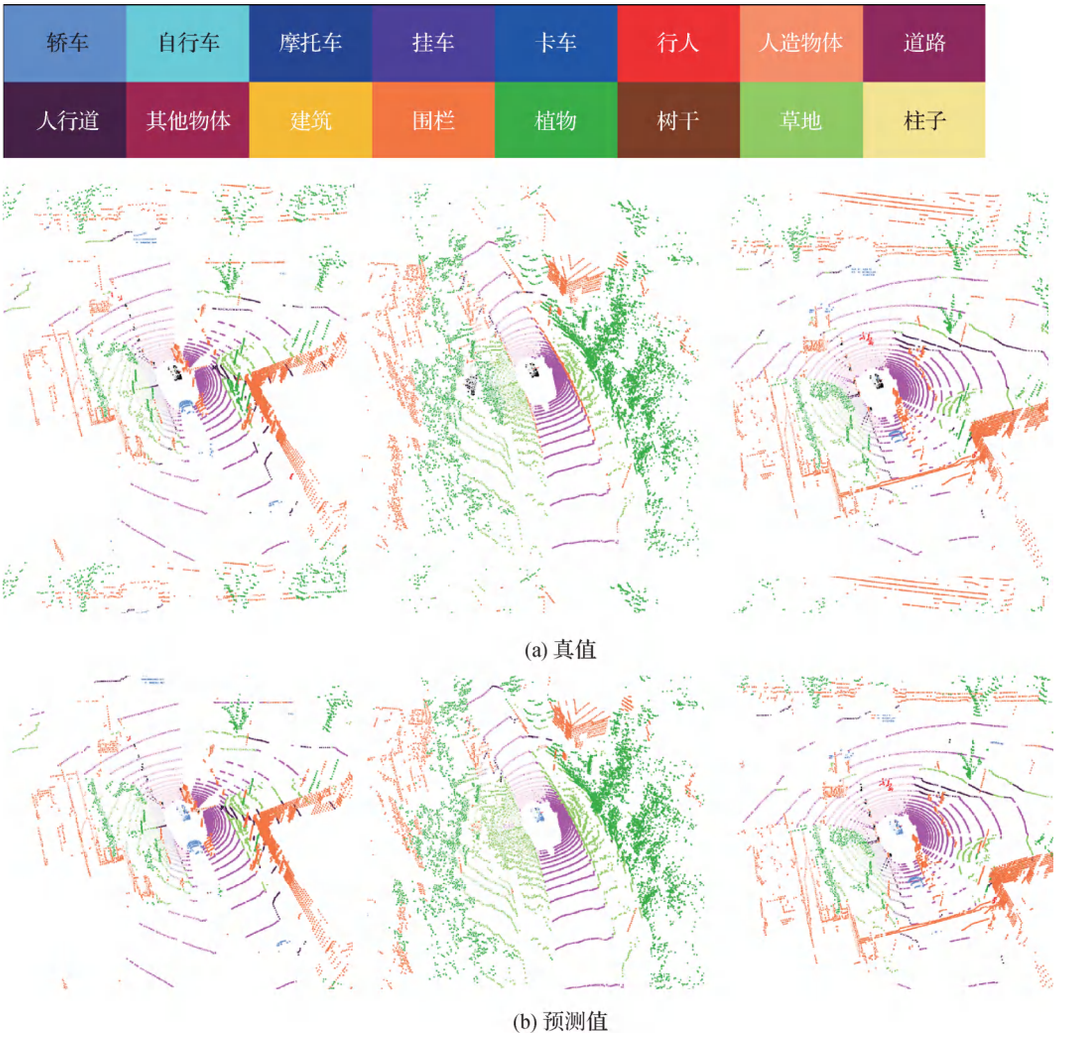
方法	mIoU	路障	自行车	客车	轿车	建筑物	摩托车	行人	圆锥桶	挂车	卡车	道路	其他物体	人行道	草地	人造物体	植物
RangeNet ++	65.5	66.0	21.3	77.2	80.9	30.2	66.8	69.6	52.1	54.2	72.3	94.1	66.6	63.5	70.1	83.1	79.8
PolarNet	71.0	74.7	28.2	85.3	**90.9**	35.1	77.5	71.3	58.8	57.4	76.1	96.5	71.1	74.7	74	87.3	85.7
Salsanext	72.2	74.8	34.1	85.9	88.4	42.2	72.4	72.2	63.1	61.3	76.5	96.0	70.8	71.2	71.5	86.7	84.4
Cylinder3d	74.8	74.5	43.1	87.4	85.9	45.1	**80.2**	79.7	65.4	61.5	80.6	96.5	71.2	74.9	75.3	87.7	87.1
本文	**75.6**	**76.2**	**43.3**	**90.3**	86.7	**49.0**	78.7	**80.5**	**66.1**	**63.7**	**80.8**	**96.6**	**71.6**	**75.3**	**75.6**	**88.6**	**87.2**

注：加粗字体表示各列最优结果。

实例注入对 mIoU 有 1.2% 的提升，空间注意力机制和 Transformer 模块分别有 1.0% 和 0.7% 的提升，两者组合共提升 1.5%，TVloss 对网络提升了 0.2%。所有模块组合在 baseline 的基础上提高了 3.1%。实验结果表明，本文提出的网络能提高激光点云的语义分割任务的精度，网络的各模块均能有效提高网络的性能。

4　结　论

本文提出一种端到端的稀疏卷积网络用于雷达点云的语义分割，使用实例注入的方法减轻数据集类别分布不平衡问题，同时为了增强网络提取全局和局部特征的能力，使用稀疏卷积的空间注意力机



(a) 真值

(b) 预测值

图 7　nuScenes 验证集语义分割结果

Fig. 7　Semantic segmentation results on the nuScenes validation set((a) ground truth; (b) prediction)

表 3　SemanticKITTI 验证集上的消融实验

Table 3　Ablation studies for network components on the SemanticKITTI validation set

/%

Baseline	实例注入	空间注意力	Transformer	TVLoss	mIoU
√	–	–	–	–	62.6
√	√	–	–	–	63.8
√	√	√	–	–	64.8
√	√	–	√	–	64.5
√	√	√	√	–	65.3
√	√	√	√	√	65.7

注：√表示采用，–表示未采用。

制提取特征图关键信息，添加 Transformer 模块提高网络提取全局信息的能力，扩大网络的感受野，提高了网络性能，并且提出新的 TVloss 加强网络对点云边缘进行监督。实验证明，本文提出的方法具有良好效果。在 SemanticKITTI 在线单帧精度评估中，mIoU 指标为 64.6%，在 nuScenes 数据集上的 mIoU 为 75.6%。消融实验结果表明，本文方法的 mIoU 在 Baseline 的基础上提高了 3.1%。对比近年出现的代表性方法，本文方法能取得较好的表现，特别是行人和车辆的分割结果取得了显著提升。

但是，通过观察实验结果可以发现，在远离雷达中心的点云数据分割误差相对较大，主要原因是远离雷达中心的点云数据相对稀疏，对于提取语义信息较为困难。

将来的工作主要有以下两个方面：1）考虑到雷达数据是一个时间序列，并且可以通过帧间匹配的方式获取两帧之间的位置信息，在网络设计上可以

利用时序和空间位置的信息;2)基于体素的方法和基于投影的方法都会产生不同的信息损失,可以通过融合基于体素的方法和基于投影的方法,减少特征损失,提高分割精度。

参考文献(References)

Aksoy E E, Baci S and Cavdar S. 2020. SalsaNet: fast road and vehicle segmentation in LiDAR point clouds for autonomous driving//2020 IEEE Intelligent Vehicles Symposium (IV). Las Vegas, USA: IEEE: 926-932 [DOI: 10.1109/iv47402.2020.9304694]

Behley J, Garbade M, Milioto A, Quenzel J, Behnke S, Stachniss C and Gall J. 2019. SemanticKITTI: a dataset for semantic scene understanding of LiDAR sequences//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9297-9307 [DOI: 10.1109/ICCV.2019.00939]

Berman M, Triki A R and Blaschko M B. 2018. The Lovasz-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4413-4421 [DOI: 10.1109/cvpr.2018.00464]

Caesar H, Bankiti V, Lang A H, Vora S, Liong V E, Xu Q, Krishnan A, Pan Y, Baldan G and Beijbom O. 2020. nuScenes: a multimodal dataset for autonomous driving//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 11618-11628 [DOI: 10.1109/cvpr42600.2020.01164]

Charles R Q, Su H, Kaichun M and Guibas L J. 2017a. PointNet: deep learning on point sets for 3D classification and segmentation//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 652-660 [DOI: 10.1109/CVPR.2017.16]

Charles R Q, Yi L, Su H and Guibas L J. 2017b. PointNet++: deep hierarchical feature learning on point sets in a metric space//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 5105-5114

Choy C, Gwak J and Savarese S. 2019. 4D spatio-temporal ConvNets: minkowski convolutional neural networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 3070-3079 [DOI: 10.1109/cvpr.2019.00319]

Cortinhal T, Tzelepis G and Aksoy E E. 2020. SalsaNext: fast, uncertainty-aware semantic segmentation of LiDAR point clouds//The 15th International Symposium on Visual Computing. San Diego, USA: Springer: 207-222 [DOI: 10.1007/978-3-030-64559-5_16]

Du J and Cai G R. 2021. Point cloud semantic segmentation method based on multi-feature fusion and residual optimization. Journal of Image and Graphics, 26(5): 1105-1116 (杜静, 蔡国榕. 2021. 多特征融合与残差优化的点云语义分割方法. 中国图象图形学报, 26(5): 1105-1116) [DOI: 10.11834/jig.200374]

Gerdzhev M, Razani R, Taghavi E and Liu B B. 2021. TORNADO-Net: mulTiview tOtal vaRiatioN semAntic segmentation with Diamond inceptiOn module//Proceedings of 2021 IEEE International Conference on Robotics and Automation. Xi'an, China: IEEE: #9562041 [DOI: 10.1109/ICRA48506.2021.9562041]

Graham B, Engelcke M and van der Maaten L. 2018. 3D semantic segmentation with submanifold sparse convolutional networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9224-9232 [DOI: 10.1109/CVPR.2018.00961]

Guo M H, Cai J X, Liu Z N, Mu T J, Martin R R and Hu S M. 2021. PCT: point cloud transformer. Computational Visual Media, 7(2): 187-199 [DOI: 10.1007/s41095-021-0229-5]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE: 1063-6919 [DOI: 10.1109/cvpr.2016.90]

Hu Q Y, Yang B, Xie L H, Rosa S, Guo Y L, Wang Z H, Trigoni N and Markham A. 2020. RandLA-Net: efficient semantic segmentation of large-scale point clouds//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11105-11114 [DOI: 10.1109/CVPR42600.2020.01112]

Kochanov D, Nejadasl F K and Booij O. 2020. KPRNet: improving projection-based LiDAR semantic segmentation [EB/OL]. [2021-07-21]. https://arxiv.org/pdf/2007.12668.pdf

Li G H, Yuan Y F, Ben X Y and Zhang J P. 2020. Spatiotemporal attention network for microexpression recognition. Journal of Image and Graphics, 25(11): 2380-2390 (李国豪, 袁一帆, 贲晛烨, 张军平. 2020. 采用时空注意力机制的人脸微表情识别. 中国图象图形学报, 25(11): 2380-2390) [DOI: 10.11834/jig.200325]

Liu S, Huang S Y, Cheng H H, Shen J Y and Chen S Y. 2021. A deep residual network with spatial depthwise convolution for large-scale point cloud semantic segmentation. Journal of Image and Graphics, 26(12): 2848-2859 (刘盛, 黄圣跃, 程豪豪, 沈家瑜, 陈胜勇. 2021. 结合空间深度卷积和残差的大尺度点云场景分割. 中国图象图形学报, 26(12): 2848-2859) [DOI: 10.11834/jig.200477]

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin transformer: hierarchical vision transformer using shifted windows [EB/OL]. [2021-03-21]. https://arxiv.org/pdf/2103.14030.pdf

Milioto A, Vizzo I, Behley J and Stachniss C. 2019. RangeNet++: fast and accurate LiDAR semantic segmentation//Proceedings of 2019 IEEE/RSJ International Conference on Intelligent Robots and

Systems (IROS). Macau, China: IEEE: 4213-4220 [DOI: 10.1109/IROS40897.2019.8967762]

Song W, Cai W Y, He S Q and Li W J. 2021. Dynamic graph convolution with spatial attention for point cloud classification and segmentation. Journal of Image and Graphics, 26(11): 2691-2702 (宋巍, 蔡万源, 何盛琪, 李文俊. 2021. 结合动态图卷积和空间注意力的点云分类与分割. 中国图象图形学报, 26(11): 2691-2702) [DOI: 10.11834/jig.200550]

Tao S B, Liang C, Jiang T P, Yang Y J and Wang Y J. 2021. Sparse voxel pyramid neighborhood construction and classification of LiDAR point cloud. Journal of Image and Graphics, 26(11): 2703-2712 (陶帅兵, 梁冲, 蒋腾平, 杨玉娇, 王永君. 2021. 激光点云的稀疏体素金字塔邻域构建与分类. 中国图象图形学报, 26(11): 2703-2712) [DOI: 10.11834/jig.200262]

Tatarchenko M, Park J, Koltun V and Zhou Q Y. 2018. Tangent convolutions for dense prediction in 3D//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3887-3896 [DOI: 10.1109/cvpr.2018.00409]

Thomas H, Qi C R, Deschaud J E, Marcotegui B, Goulette F and Guibas L. 2019. KPConv: flexible and deformable convolution for point clouds//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 6411-6420 [DOI: 10.1109/ICCV.2019.00651]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang Y Y, Luo L K and Zhou Z G. 2019. Road scene segmentation based on KSW and FCNN. Journal of Image and Graphics, 24(4): 583-591 (王云艳, 罗冷坤, 周志刚. 2019. 结合 KSW 和 FCNN 的道路场景分割. 中国图象图形学报, 24(4): 583-591) [DOI: 10.11834/jig.180467]

Xu C F, Wu B C, Wang Z N, Zhan W, Vajda P, Keutzer K and Tomizuka M. 2020. SqueezeSegV3: spatially-adaptive convolution for efficient point-cloud segmentation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 1-19 [DOI: 10.1007/978-3-030-58604-1_1]

Yu S and Wang X L. 2021. Remote sensing building segmentation by CGAN with multilevel channel attention mechanism. Journal of Image and Graphics, 26(3): 686-699 (余帅, 汪西莉. 2021. 含多级通道注意力机制的 CGAN 遥感图像建筑物分割. 中国图象图形学报, 26(3): 686-699) [DOI: 10.11834/jig.200059]

Zhang F H, Fang J, Wah B and Torr P. 2020a. Deep FusionNet for point cloud semantic segmentation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 644-663 [DOI: 10.1007/978-3-030-58586-0_38]

Zhang Y, Zhou Z X, David P, Yue X Y, Xi Z R, Gong B Q and Foroosh H. 2020b. PolarNet: an improved grid representation for online LiDAR point clouds semantic segmentation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 9601-9610 [DOI: 10.1109/CVPR42600.2020.00962]

Zheng Y, Lin C Y, Liao K, Zhao Y and Xue S. 2021. LiDAR point cloud segmentation through scene viewpoint offset. Journal of Image and Graphics, 26(10): 2514-2523 (郑阳, 林春雨, 廖康, 赵耀, 薛松. 2021. 场景视点偏移的激光雷达点云分割. 中国图象图形学报, 26(10): 2514-2523) [DOI: 10.11834/jig.200424]

Zhou Y and Tuzel O. 2018. VoxelNet: end-to-end learning for point cloud based 3D object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4490-4499 [DOI: 10.1109/cvpr.2018.00472]

Zhu X G, Zhou H, Wang T, Hong F Z, Li W, Ma Y X, Li H S, Yang R G and Lin D H. 2022. Cylindrical and asymmetrical 3D convolution networks for LiDAR-based perception. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10): 6807-6822 [DOI: 10.1109/tpami.2021.3098789]

作者简介

刘盛,男,副教授,主要研究方向为数字图像处理和计算机视觉。E-mail: edliu@zjut.edu.cn

曹益烽,男,硕士研究生,主要研究方向为计算机视觉和点云分割。E-mail: 2111912100@zjut.edu.cn

黄文豪,男,硕士研究生,主要研究方向为计算机视觉和点云检测。E-mail: 2111912111@zjut.edu.cn

李丁达,男,硕士研究生,主要研究方向为计算机视觉和点云补全。E-mail: 2111912136@zjut.edu.cn