第 60 卷 第 20 期/2023 年 10 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于稀疏卷积和注意力机制的点云语义分割方法

左蒙[1,2,3,4]，刘意杨[1,2,3*]，崔好[1,2,3]，白洪飞[2]

[1]中国科学院网络化控制系统重点实验室，辽宁 沈阳 110016；

[2]中国科学院沈阳自动化研究所，辽宁 沈阳 110016；

[3]中国科学院机器人与智能制造创新研究院，辽宁 沈阳 110169；

[4]中国科学院大学，北京 100049

摘要 近年来，三维点云语义分割方法取得了很大的进展，代表性的方法为基于稀疏卷积的方法，但是稀疏卷积会带来全局上下文信息丢失的问题。基于此，提出一种基于稀疏卷积和注意力机制的点云语义分割方法。将注意力机制引入稀疏卷积网络，增强网络对全局上下文信息的获取能力。但是注意力机制计算量较大，限制了所提方法的适用场景。进一步将空间金字塔采样引入注意力机制中，在减少计算量的同时扩展其使用场景。实验结果表明，所提方法在Scannet V2数据集上取得了71.825%的平均交并比(MIOU)，在S3DIS数据集上的MIOU达到70.5%，优于对比方法，验证了其有效性。

关键词 机器视觉；点云语义分割；稀疏卷积；注意力机制；空间金字塔采样

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP222819

Semantic Segmentation Method of Point Cloud Based on Sparse Convolution and Attention Mechanism

Zuo Meng[1,2,3,4], Liu Yiyang[1,2,3*], Cui Hao[1,2,3], Bai Hongfei[2]

[1]*Key Laboratory Networked Control Systems, Chinese Academy of Sciences, Shenyang* 110016, *Liaoning, China*;

[2]*Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang* 110016, *Liaoning, China*;

[3]*Institutes for Robotics & Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang* 110169, *Liaoning, China*;

[4]*University of Chinese Academy of Sciences, Beijing* 100049, *China*

Abstract Recently, three-dimensional point cloud semantic segmentation techniques based on sparse convolution have made great progress. However, sparse convolution causes a loss of global context information. In this study, a point cloud semantic segmentation method based on a sparse convolution and attention mechanism is proposed. Here, the attention mechanism is introduced into a sparse convolutional network to improve the network's ability to achieve global context information. However, extensive computation of the attention mechanism limits the applicability of the proposed method. Hence, to expand its usage while decreasing the amount of computation, spatial pyramid sampling is further introduced in the attention mechanism. Experimental results demonstrate that the proposed method achieves 71.825% of the average intersection over union (MIOU) on the Scannet V2 dataset and 70.5% on the S3DIS dataset, suggesting the proposed method's effectiveness and its superiority to the comparison method.

Key words machine vision; semantic segmentation of point cloud; sparse convolution; attention mechanism; spatial pyramid sampling

1 引 言

近年来，随着深度学习方法的突破性进展，许多领域结合深度学习方法取得了很大的进步。语义分割就是深度学习应用较为广泛的领域之一。语义分割可以分成二维图像语义分割与三维点云语义分割。其中，

收稿日期：2022-10-18；**修回日期**：2022-12-01；**录用日期**：2022-12-12；**网络首发日期**：2023-01-05

基金项目：国家重点研发计划项目(2021YFB3301400)、辽宁省兴辽英才计划项目(XLYC1907057)

通信作者：*sialiuyiyang@sia.cn

二维图像语义分割已经在学术和工程上取得了较大的成果。而三维语义分割由于起步较晚，成果并不丰硕。但是由于智能制造、三维重建、自动驾驶等领域的需求，三维点云语义分割近年来获得了较大的关注[1]。

三维点云语义分割与二维图像的语义分割不同。二维图像由于具有结构化的信息，可以直接使用卷积神经网络（CNN）等方法处理。但是点云数据由于具有无序性、相互作用性、变换无关性的特性，并不能直接使用CNN方法对其进行处理。为了将深度学习方法扩展到点云处理中，许多方法被提出。其中，典型的方法有基于多视图的方法、基于点的方法和基于体素的方法。

基于多视图的方法思想很朴素，将三维数据投影到二维平面，使无序的点云数据转换为二维的图像，从而方便CNN等方法处理。典型的方法有MVD[2]与VMF[3]。

基于点的方法直接从点云的无序结构中学习点云的特征，典型方法是PointNet[4]。PointNet使用多层感知机（MLP）和最大池化等操作克服点云的无序性，从而学习点云的全局特征，但是PointNet在学习点云局部特征方面存在缺陷。因此，后续的PointNet＋＋[5]、SpiderCNN[6]、GCA-Conv[7]等方法在学习点云全局特征的同时增强了对局部特征的获取能力。

基于体素的方法将无序的点云数据转换为规则的体素表示形式，从而建立结构化的数据形式，典型的方法有VoxNet[8]、MV-ASPP[9]等。最近的研究表明[10]，基于体素的方法在点云语义分割中表现出最好的性能，相比于其他方法更受关注。但上述方法在体素化之后使用常规的CNN卷积，没有利用点云体素化后的稀疏结构，因此带来了高计算量和高内存消耗。

为了进一步适应数据的稀疏性，稀疏卷积的概念被提出来[11]。稀疏卷积与常规CNN不同，在进行卷积时只对有效数据进行计算，忽略无效的数据区域，从而减小计算量。基于这种思想，许多针对点云这种稀疏性较强数据的方法被提出来。OctNets[12]将稀疏体素存储在八叉树中，使网络减少对无效数据的处理。Vote3Deep[13]通过基于中心点对称的投票策略实现稀疏卷积，并通过激活函数维持中间层的稀疏性。SSCN[14]采用新的稀疏卷积运算符，在不需要增加活跃数据数量的前提下，仅对有效数据进行计算，进一步优化了点云的处理效率。

稀疏卷积虽然能够减小计算量，但是和普通卷积一样，存在局部感受野受限的问题，导致网络对数据的远距离信息获取能力有限，限制了分割的准确率。为了增加感受野，注意力机制被引入网络中。其中，典型的方法有空间注意力方法、通道注意力方法以及将两者结合的方法。Non Local[15]是一种典型的空间注意力方法，通过计算输入注意力模块的所有位置的特征的加权和从而计算点云全局数据的内部相关性，解决数据的长距离依赖问题。SENet[16]是一种应用通道注意力的网络，通过挤压模块和激励模块来捕捉数据通道间的关系，从而决定数据不同维度间的权重。CBAM[17]将空间注意力和通道注意力串联起来，并利用全局池化来利用空间全局信息，从而自适应关注重要的对象和区域。

但是由于注意力机制过大的计算量和内存占用率，其在点云分割这类数据较为庞大的任务中受到很大的限制。为了拓展注意力机制的使用场景，将空间金字塔采样应用到注意力机制中，在不影响其性能的情况下减小计算量和内存占用率，增强语义分割网络的分割性能。

针对卷积缺乏远距离信息获取能力和注意力机制计算量过大的问题，本文提出一种基于稀疏卷积和注意力机制的点云语义分割方法。使用注意力机制解决稀疏卷积对于远距离信息获取能力不足的问题，在保证点云稀疏性的同时提高网络的分割精度。使用空间金字塔采样解决注意力机制计算量大的问题，将池化后的特征相互连接减少采样的损失。将经过空间金字塔采样后的注意力机制融合进网络的低维特征层，并与编码层的特征连接，在不显著提高计算量的同时提高网络对细节的感知能力。

2　基于改进注意力机制的网络设计

所提网络模型如图1所示，主要由体素化、特征提取网络、解体素化等3部分组成。环境感知到的点云首先需要进行体素化，从而将无序的数据转化为结构化的数据。再将体素化后的点云数据输入特征提取网络中，特征提取网络是一个由编码器和解码器组成的

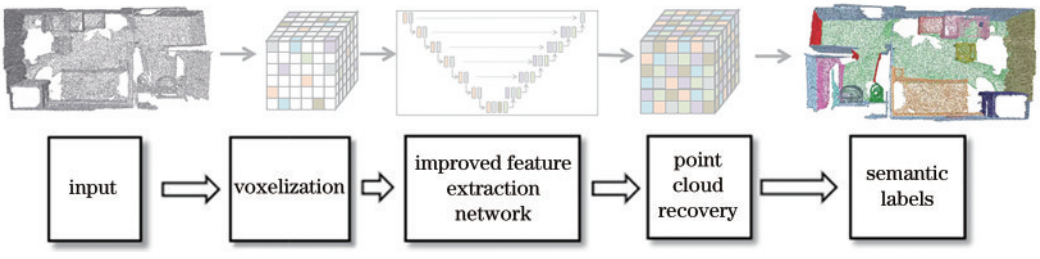

图1　点云语义分割网络模型

Fig. 1　Point cloud semantic segmentation network model

U-Net[18]结构。经过特征提取后可以输出场景的语义信息，最后进行解体素化将体素的语义恢复到每个点的语义，从而得到每个点的语义分割结果。

2.1　体素化与解体素化

输入网络的是N个三维点的集合$\{(\boldsymbol{p}_i, f_i)\}$，$\boldsymbol{p}_i$表示每个点的坐标信息，$f_i$表示每个点的颜色信息，$i\in\{1,\cdots,N\}$。点云在三维空间中$x$轴、$y$轴、$z$轴大小分别为$W$、$D$、$H$，若体素采样的分辨率为$r$，则第$i$个点体素化后在体素网格中的坐标$\boldsymbol{p}_i'$的表达式为

$$\begin{cases}\boldsymbol{p}_i'=(a,b,c)\\ a=\left\lfloor\dfrac{x_i}{r}\right\rfloor\\ b=\left\lfloor\dfrac{y_i}{r}\right\rfloor\\ c=\left\lfloor\dfrac{z_i}{r}\right\rfloor\end{cases},\tag{1}$$

式中：$\lfloor\cdot\rfloor$表示向下取整。

每个体素输入网络的特征向量如式(2)所示：

$$\boldsymbol{F}_{w',d',h'}=\frac{1}{N}\sum_{i=1}^{N}\chi_{w',d',h'}(\boldsymbol{p}_i')f_i,\tag{2}$$

$$\chi_{w',d',h'}[\boldsymbol{p}_i'(a,b,c)]=\begin{cases}1,(a,b,c)=(w',d',h')\\0,\text{else}\end{cases},\tag{3}$$

$$\begin{cases}w'\in\left(0,\left\lfloor\dfrac{W}{r}\right\rfloor\right)\\ d'\in\left(0,\left\lfloor\dfrac{D}{r}\right\rfloor\right)\\ h'\in\left(0,\left\lfloor\dfrac{H}{r}\right\rfloor\right)\end{cases}。\tag{4}$$

最后可以根据每个体素的坐标$\boldsymbol{p}_i'$和特征向量$\boldsymbol{F}_{w',d',h'}$建立哈希表，哈希表的键为非空体素坐标，哈希表的值为非空体素的特征向量，便于后续稀疏卷积的计算，加快计算速度。

由于最后需要保证输出数据与输入数据一致，因此需要将网络输出的体素语义标签转换为点的体素标签。网络解体素化使用的是三线性插值法，每个需要恢复的三维点通过周围的8个体素的值进行加权求和得到。

2.2　基于稀疏卷积和改进注意力机制的特征提取网络

特征提取的网络主体为U-Net结构，U-Net使用编码器-解码器结构，通过编码器对原始特征进行降采样，再通过解码器对降采样的特征进行上采样，得到分割结果，最后根据预测结果与真实结果的差异进行反向传播训练网络参数。特征提取网络细节如图2所示。

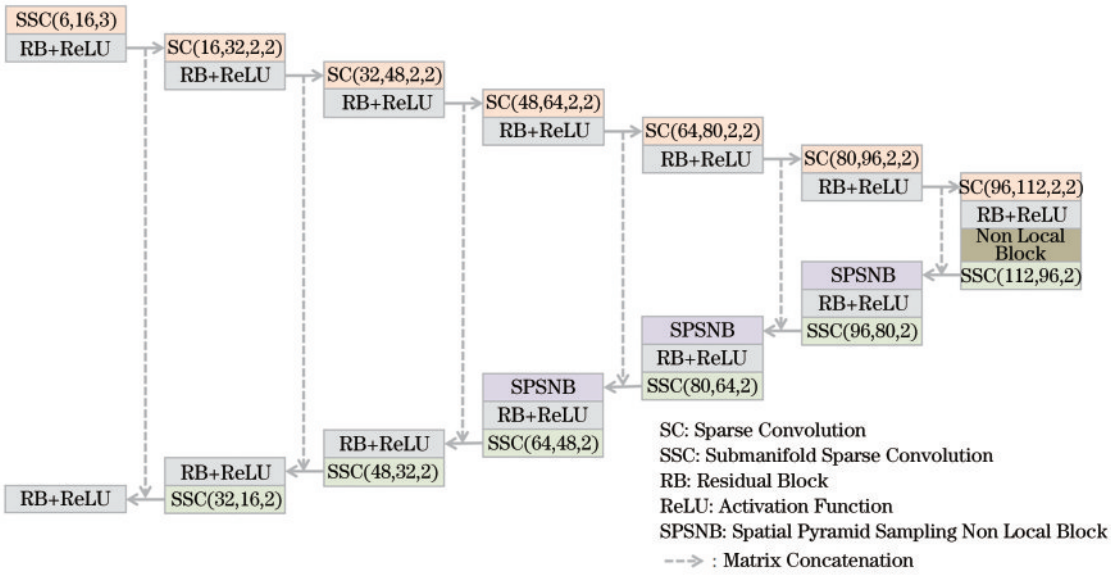

图2　基于稀疏卷积和改进注意力机制的特征提取网络

Fig. 2　Feature extraction network based on sparse convolution and improved attention mechanism

受SSCN的启发，特征提取网络的卷积模块使用的不是常规的CNN卷积，而是稀疏卷积。这是由于点云是一种稀疏性比较强的数据，特别是在进行体素化之后，大部分的体素中并不存在数据，若使用常规卷积对体素化的点云数据进行处理，势必会破坏点云稀疏的表示形式，从而增加不必要的计算量。因此，为了保持点云数据的稀疏性并减小计算量，特征提取网络的编码器和解码器均由稀疏卷积块(SC)与子流形稀疏卷积块(SSC)组成。

编码器由多层卷积层组成，每层卷积层有一个预激活的残差块，特征经过残差块之后经过一个ReLU函数，再由一个稀疏卷积块进行计算，最后输入下一层编码器。

所用残差块同样使用了稀疏卷积，其结构如图3所示。

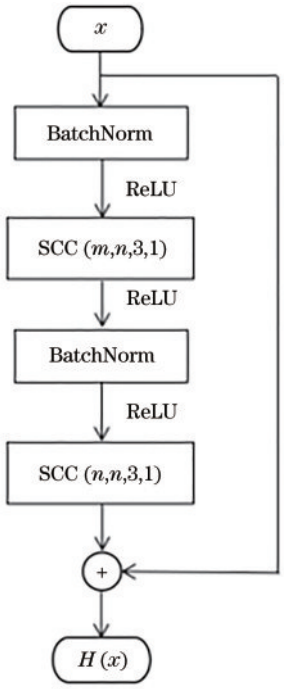

图 3 基于稀疏卷积的残差块

Fig. 3 Residual block based on sparse convolution

残差块中使用的ReLU激活函数为

$$f(x)=\max(0,x)。\tag{5}$$

解码器同样由多层卷积层组成。每层卷积层包含预激活的残差块和ReLU函数，并且经过反卷积恢复特征维度，经过多层解码器后，数据会恢复到输入特征提取网络之前的大小，并且输出数据的语义信息。

受Dyco-Net[19]的启发，所提网络将Non Local Block的注意力机制模块应用到编码层与解码层的最底层，从而增强网络对全局信息的获取能力。

为进一步扩展Non Local Block的应用范围，将空间金字塔池化应用到Non Local Block中，使其能够拓展到网络的第4层进一步增强网络的特征提取能力。经过空间金字塔采样后的Non Local Block被称为Spatial Pyramid Sampling Non Local Block(SPSNB)，将SPSNB应用到解码器的卷积层中将上一层的反稀疏卷积的输出作为输入，再输出到下一层的残差块中，从而提升网络的全局信息获取能力。

特征提取网络还将编码器特征与解码器特征通过跳跃连接结合起来，从而把对应尺度上的特征信息引入解码的过程中，为分割任务提供多层次的信息，达到更精细的分割效果。

特征提取网络输出的特征解体素化后，经过全连接层将特征维度变换到n类，n是分割的类别量。最后采用交叉熵损失函数进行代价计算，优化网络的参数。交叉熵损失函数的表达式为

$$L=\frac{1}{N}\sum_{i}L_i=-\frac{1}{N}\sum_{i}\sum_{c=1}^{M}y_{ic}\log p_{ic},\tag{6}$$

式中：N表示输入点云的点的个数；M表示分割的类别数；y_{ic}在点i的标签为c时取1，不为c时取0；p_{ic}是特征经过全连接层后预测的结果，表示点i预测属于类别c的概率。

2.3 稀疏卷积

针对点云体素化数据的稀疏性，使用稀疏卷积作为特征提取的卷积方式，从而减小计算量。稀疏卷积本质上还是卷积操作，但是其对于输入输出数据的保存形式和具体的运算过程和普通的卷积方式有较大的不同。

对于大小为$W\times H\times D$、特征维度为m的体素化点云。经过一个大小为$f\times f\times f$、卷积步长为1的稀疏卷积核，每次卷积核同时与所有特征通道窗口内的有效数据进行卷积计算，从而得到大小为$W\times H\times D$、特征维度为1的单通道数据，经过n个卷积核之后，将n个卷积核得到的结果作为不同特征维度上的数据，则可以得到大小为$W\times H\times D$、特征维度为n的结果，并送入下一阶段进行卷积。基于此，将稀疏卷积定义为$\mathrm{SC}(m,n,f,s)$，m为输入的特征维度，n为经过卷积后的输出特征维度，f为稀疏卷积的核的尺寸，s为卷积的步长。为了保证输出的张量的尺寸与输入的张量的尺寸一样，引入稀疏卷积的改进SSC，它是修正过的稀疏卷积$\mathrm{SC}(m,n,f,s=1)$。

为了直观地显示稀疏卷积的运算过程，并与普通卷积进行对比，图4显示了稀疏卷积和普通卷积的运算过程。

由图4可以看出，经过普通卷积之后，原本不存在有效数据的地方也出现了数据，且卷积块经过每一个位置时都需要计算，因此不仅破坏了原有数据的稀疏性，还带来了不必要的计算量。

与普通卷积不同的是，稀疏卷积引入空值补零、强制清零等必要操作。空值补零是为了能进行正常的卷积，由于体素化后会有较大部分存在空洞，因此在卷积的时候会进行空值补零。强制清零是维持数据稀疏性的核心操作。由于原本零值的地方会在卷积之后输出非零值，造成数据结构的膨胀，因此会在卷积之后将卷积之前为零值的地方重新置零。由于后续零值的乘法运算复杂度为零，因此稀疏卷积既可以维持数据的稀疏性，又减小了计算量。

2.4 结合空间金字塔采样的改进注意力机制

不管是正常卷积还是稀疏卷积，都会带来有限的局部感受野，导致网络过于关注局部信息，而对全局信息的获取能力不足，影响最后的分割精度。为了解决这个问题，所提网络将Non Local Block的注意力机制模块应用到编码层与解码层的最底层，从而增强网络对全局信息的获取能力。所使用的Non Local Block的结构如图5所示。

Non Local Block之所以能够建立数据的远距离依赖关系，是因为输入模块的每一个点的响应是其他

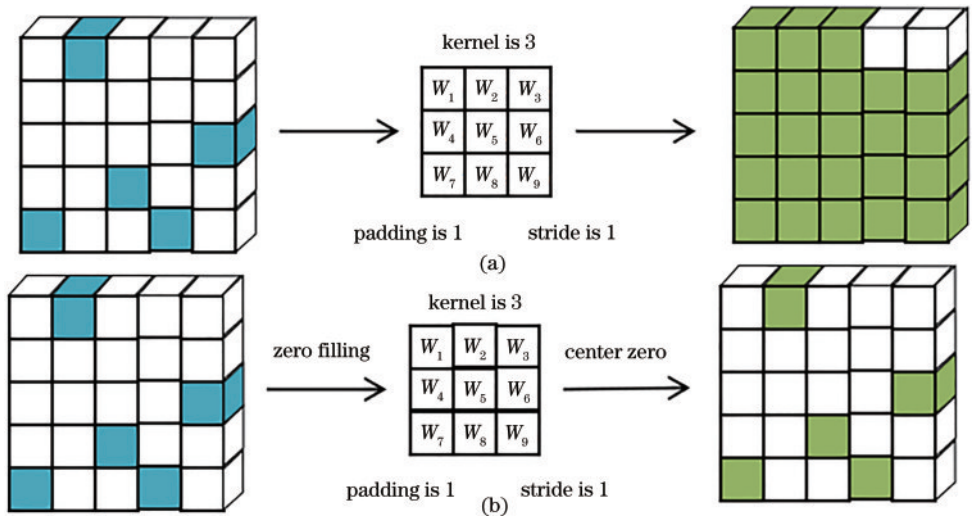

图 4　普通卷积与稀疏卷积对比。(a)普通卷积;(b)稀疏卷积

Fig. 4　Comparison between ordinary convolution and sparse convolution. (a) Ordinary convolution; (b) sparse convolution

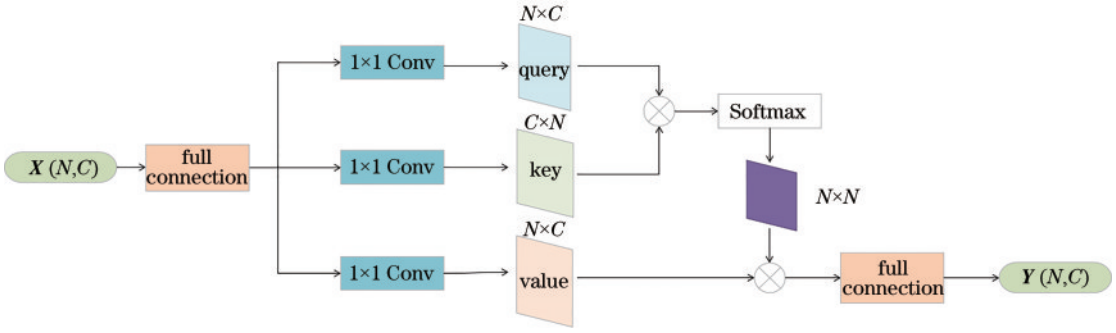

图 5　Non Local Block 结构

Fig. 5　Non Local Block structure

所有点的特征权重和。输入 Non Local Block 的数据为 $\boldsymbol{X}$,$\boldsymbol{X}$ 的维度为 $N\times C$,N 为经过编码器之后的数据的大小,C 是经过编码器之后的特征的维数。输入经过全连接网络后再通过 3 次 1×1 的卷积操作可以得到 3 种向量,查询向量 $\boldsymbol{w}_{\mathrm{q}}$、键向量 $\boldsymbol{w}_{\mathrm{k}}$ 和值向量 $\boldsymbol{w}_{\mathrm{v}}$。Non Local Block 的输出 $\boldsymbol{Y}$ 为

$$\boldsymbol{Y}=\mathrm{Softmax}\left(\boldsymbol{w}_{\mathrm{q}}\boldsymbol{w}_{\mathrm{k}}^{\mathrm{T}}\right)\boldsymbol{w}_{\mathrm{v}}。\tag{7}$$

分析式(7)的计算复杂度可以发现,$\boldsymbol{w}_{\mathrm{q}}$ 的维度为 $N\times C$,$\boldsymbol{w}_{\mathrm{k}}^{\mathrm{T}}$ 的维度为 $C\times N$,它们相乘再经过 Softmax 函数的计算之后的维度为 $N\times N$,而 $\boldsymbol{w}_{\mathrm{v}}$ 的维度为 $N\times C$,则最后进行矩阵相乘时会有一个 $N\times N$ 的计算:

$$\mathbf{R}^{N\times C}\times\mathbf{R}^{C\times N}\rightarrow\mathbf{R}^{N\times N}\times\mathbf{R}^{N\times C}\rightarrow\mathbf{R}^{N\times C}。\tag{8}$$

在点云分割中,就算经过了体素化,但是点的数量依然很庞大,这就意味着 N 会很大,因此会带来极大的计算量,导致计算资源的极大占用甚至崩溃。如何解决 N 值过大的问题是减小计算量的关键。一种很朴素的思想是将 N 替换成一个足够小的量 $S(S\ll N)$,从而大大减小矩阵计算的维度[20]。那么如何将 N 降采样到 S,且在降采样之后不影响 Non Local Block 的效果是至关重要的。所提网络引入空间金字塔池化操作,如图 6 所示。

输入 Non Local Block 模块的数据大小为 $N\times C$,体素或点的规模为 N,每个体素或点的特征大小为 C。空间金字塔采样对每一维的 N 个体素或点进行 4 种分

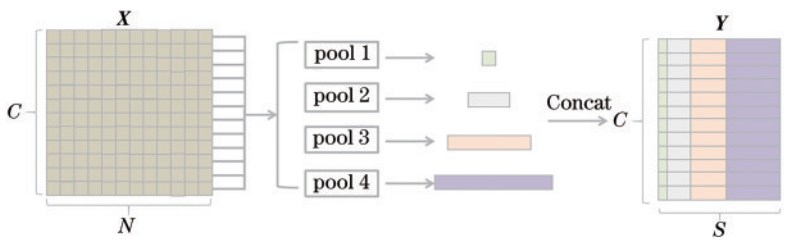

图 6　空间金字塔采样

Fig. 6　Spatial pyramid sampling

辨率的降采样，再将不同分辨率采样的特征连接起来作为新的输入，从而建立起全局和多尺度的表示方法。除了能有效表示初始数据，空间金字塔池化减小了数据的大小，从而大大降低计算开销。计算过程如下：

$$\begin{cases} \boldsymbol{X}\{N,C\} \xrightarrow{\text{pool 1}} \boldsymbol{Y}_1\{S_1,C\} \\ \boldsymbol{X}\{N,C\} \xrightarrow{\text{pool 2}} \boldsymbol{Y}_2\{S_2,C\} \\ \boldsymbol{X}\{N,C\} \xrightarrow{\text{pool 3}} \boldsymbol{Y}_3\{S_3,C\} \\ \boldsymbol{X}\{N,C\} \xrightarrow{\text{pool 4}} \boldsymbol{Y}_4\{S_4,C\} \\ \boldsymbol{Y}_1,\boldsymbol{Y}_2,\boldsymbol{Y}_3,\boldsymbol{Y}_4 \xrightarrow{\text{Concat}} \boldsymbol{Y}\{S_1+S_2+S_3+S_4,C\} \end{cases} 。 \quad (9)$$

经过空间金字塔采样后的 Non Local Block 如图 7 所示。对键向量和值向量在卷积之后进行空间金字塔采样。式(10)为采样后的计算量，从采样前的 $N\times N$ 的矩阵计算变为采样后的 $N\times S$ 的矩阵计算，计算量大大降低。因此 Non Local Block 能够应用到特征提取网络的任意层，提高网络对全局上下文信息的获取能力。

$$\mathbf{R}^{N\times C}\times\mathbf{R}^{C\times S}\rightarrow\mathbf{R}^{N\times S}\times\mathbf{R}^{S\times C}\rightarrow\mathbf{R}^{N\times C}。 \quad (10)$$

将经过空间金字塔采样后的 Non Local Block 拓展到网络的第 4 层并与编码层的特征连接，在不显著提高计算量的同时提高了网络对细节的感知能力。

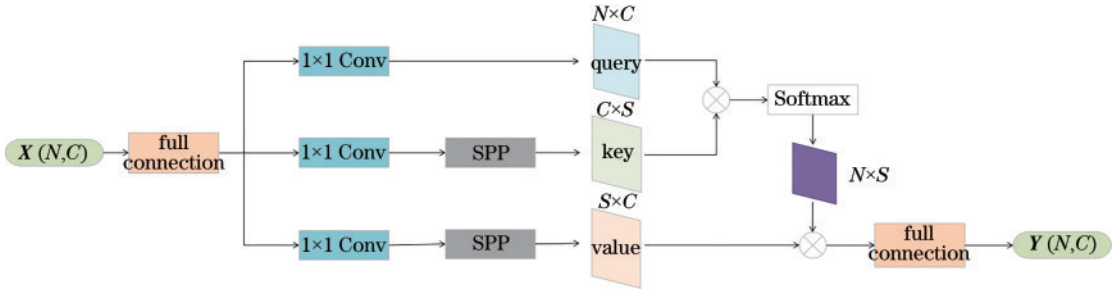

图 7　结合空间金字塔采样的 Non Local Block

Fig. 7　Non Local Block combined with spatial pyramid sampling

3　实验结果与分析

3.1　数据集

为了验证所提网络的有效性，在公开数据集 Scannet V2[21] 和 S3DIS[22] 上进行了实验。Scannet V2 包含 1613 个有 3D 实例的室内场景。数据集被分为训练集、验证集和测试集，分别包含 1201、312 和 100 个数据。数据集内包含 18 类对象。

S3DIS 也是一个室内数据集，相对于 Scannet V2 数据集来说，S3DIS 数据规模更大，场景里包含的点云也更多。数据集一共包含 6 个大型的室内区域，有 271 个房间，房间类型为会议室、大厅等数据较大的场景。每个房间内有 13 类不同物体。网络在第 5 个区域测试，在其他的区域训练。

3.2　实验环境与评价指标

实验在 Ubuntu 20.04 系统上进行，系统为 64 位。实验采用的 CPU 型号为 intel i7-7700，GPU 型号为 NVIDIA GeForce RTX3060。实验训练 400 轮，每轮的 batch 为 4，优化器采用 Adam，学习率的表达式为

$$R_{\text{lr}}=\max\left(R_{\text{lr,base}}\times 0.1^{\frac{N_{\text{epoch}}}{\frac{n_t}{n_s}}},10^{-6}\right), \quad (11)$$

式中：N_{epoch} 表示当前训练的轮数；n_t 表示总样本数；n_s 表示 batch 数；$R_{\text{lr,base}}$ 表示基础学习率。

采用平均交并比(MIOU)作为评价指标，MIOU 的表达式为

$$R_{\text{MIOU}}=\frac{1}{k}\sum_{i=1}^{k}\frac{P_{ii}}{\sum_{j=1}^{k}P_{ij}+\sum_{j=1}^{k}P_{ji}-P_{ii}}, \quad (12)$$

式中：k 代表数据集中所有待分类的类别数；P_{ij} 表示应当属于类 j 但被错误分成第 i 类的点的数量，P_{ii} 和 P_{ji} 同理。从式(12)可以看出，MIOU 表示的是待分割场景中对所有类的分割准确率，而不是对逐点的分类正确率。因此，MIOU 能够较好评价场景的分割准确率。

3.3　体素分辨率设置

输入网络的数据为点云，为了使数据结构化，需要对点云进行体素化。为了探究点云体素化分辨率对网络性能的影响，使用 Scannet V2 数据集进行了不同体素分辨率的实验。实验中取会议室场景为应用场景计算体素数量，会议室大小为 60 m×60 m×4 m。表 1 为实验结果。

从表 1 可以看出，当体素分辨率为 2 cm 时，网络的

表 1　不同体素分辨率参数对比

Table 1　Comparison of different voxel resolution parameters

Voxel resolution /cm	MIOU /%	Number of totle voxels	Number of active voxels
1	68.214	1.440×10^{10}	112541
2	70.821	1.800×10^{9}	95145
3	70.154	5.333×10^{8}	87165
4	68.657	2.250×10^{8}	81104
5	67.241	1.152×10^{8}	79514
6	66.142	6.667×10^{7}	72015

分割性能最好。体素分辨率越大，采样后的信息丢失越多，因此体素分辨率大于 2 cm 后，网络的分割性能下降。但是体素分辨率过小会导致体素数量的指数级膨胀，同样导致网络的分割性能下降。因此，所提网络使用 2 cm 的体素分辨率进行采样。进一步分析总体素数量和有效体素数量的关系可以发现，在体素分辨率为 2 cm 时，总体素数量是有效数量的 18918 倍。因此使用稀疏卷积只对有效数据区域进行计算，可以大大缩短计算时间，验证了稀疏卷积在处理点云数据上的优势。

3.4 空间金字塔采样参数设置

在改进的注意力机制中，空间金字塔采样的方法和采样大小决定了改进方法的性能，从而影响网络的分割性能。为了研究这种影响，通过对比不同的采样方法和采样大小的网络在 Scannet V2 数据集上的 MIOU 来确定参数，实验结果如表 2 所示。

表 2 不同空间金字塔采样参数对比

Table 2 Comparison of sampling parameters in different spatial pyramids

Sampling method	Sample size	Size of S	MIOU /%
Pyramid random	1,4,9,36	50	70.461
Pyramid max	1,4,9,36	50	71.324
Pyramid average	1,4,9,36	50	71.640
Pyramid average	1,9,36,64	110	71.825
Pyramid average	1,16,64,144	225	71.833

采样方法如表 2 所示，分为金字塔随机采样、金字塔最大池化和金字塔平均池化。金字塔随机采样指随机选取每个特征维度上的数据，使用 numpy.random.choice 函数随机选取 4 次，第 1 次选择 1 个数据，第 2 次选择 4 个数据，第 3 次选择 9 个数据，第 4 次选择 36 个数据，再将其拼接起来作为最后的数据。可以发现，金字塔随机采样后的 MIOU 最低，甚至低于原始网络的性能。这是由于随机采样具有随机性，很难完全表示数据的全局特点。金字塔最大池化的性能强于金字塔随机采样，但是低于金字塔平均池化，这是由于最大池化只能描述数据的全局特性，但是对于数据的局部特性描述不足。金字塔平均池化具有最好的性能，平均池化在金字塔多层次采样的结构中能更好地描述数据的特性。因此，所提网络采用金字塔平均池化作为空间金字塔采样的方法。

还对比了不同的采样大小对于性能的影响。在采样方法为金字塔平均池化的前提下，可以发现，随着采样大小的增加，网络的分割性能会增强，但是也会增加采样后 S 的大小，导致计算量的增加。尤其是采样大小从(1,9,36,64)提高到(1,16,64,144)时，网络的性能只提升了 0.008 个百分点，但是采样后的大小变为原来的 2.05 倍。因此，为了平衡性能和效率，所提网络采用的采样大小为(1,9,36,64)。

3.5 Scannet V2 数据集实验结果

为了评价所提算法的有效性，在 Scannet v2 的测试集上进行了点云分割对比实验。表 3 为实验结果，可以看出，所提算法的 MIOU 达到了 71.825%。

表 3 Scannet V2 测试集实验结果对比

Table 3 Comparison of experimental results of Scannet V2 test set unit:%

Class	PointNet++	FPConv	SSCN	Minkowski	Proposed algorithm
Wall	52.3	79.9	83.6	84.5	83.8
Floor	67.7	94.8	95.1	95.9	94.9
Cabinet	25.6	60.3	65.3	63.9	**68.4**
Bed	47.8	76.0	80.7	80.8	80.4
Chair	36.0	79.8	90.4	90.1	**91.2**
Sofa	34.6	69.6	82.0	81.5	80.2
Table	23.2	61.4	72.2	70.9	**73.5**
Door	26.1	52.4	64.3	59.8	**67.2**
Window	25.2	56.7	60.5	60.6	**64.1**
Bookshelf	45.8	71.3	78.0	75.4	76.0
Picture	11.7	25.0	31.3	31.5	**35.1**
Counter	25.0	39.2	62.5	66.0	61.2
Desk	27.8	6.3	58.7	60.5	**63.9**
Curtain	24.7	53.4	75.8	71.3	**76.2**
Refrigerator	21.2	53.8	49.4	55.6	**56.2**
Shower curtain	58.4	72.3	70.8	66.5	**72.2**
toilet	14.5	87.2	93.0	90.3	84.2
Sink	54.8	59.8	63.9	65.2	62.5
Bathtub	36.4	78.5	87.4	93.5	88.1
other	18.3	45.7	51.4	56.6	**57.2**
MIOU	33.9	63.9	70.8	70.6	**71.8**

为了直观展示算法的分割效果，对分割效果进行了可视化，不同网络在 Scannet V2 测试集上的可视化效果如图 8 所示。

对比所提网络与 PointNet++可以发现，所提网络的 MIOU 提高了 37.9 个百分点，且在墙壁、桌子、床、椅子等类别的分割精度均有所提升。这是由于 PointNet++将点云划分为不同的子区域，并采用最远点采样维持区域内点云分布，但由于点云密度存在差异，采样会破坏数据的局部关系，因此在室内的复杂场景内 PointNet++不具有良好的分割性能。而所提网络基于体素建立数据的结构化信息并使用稀疏卷积维护数据的稀疏性，有效地利用数据的局部特性，因此在各个类别上的分割精度均比 PointNet++高。

对比所提网络与 FPConv[23] 可以发现，所提网络的 MIOU 提高了 7.9 个百分点。FPConv 将局部点云通过插值的方法映射为一个二维平面，从而用二维卷积的方法计算特征。由于这种方法使用局部平面化的

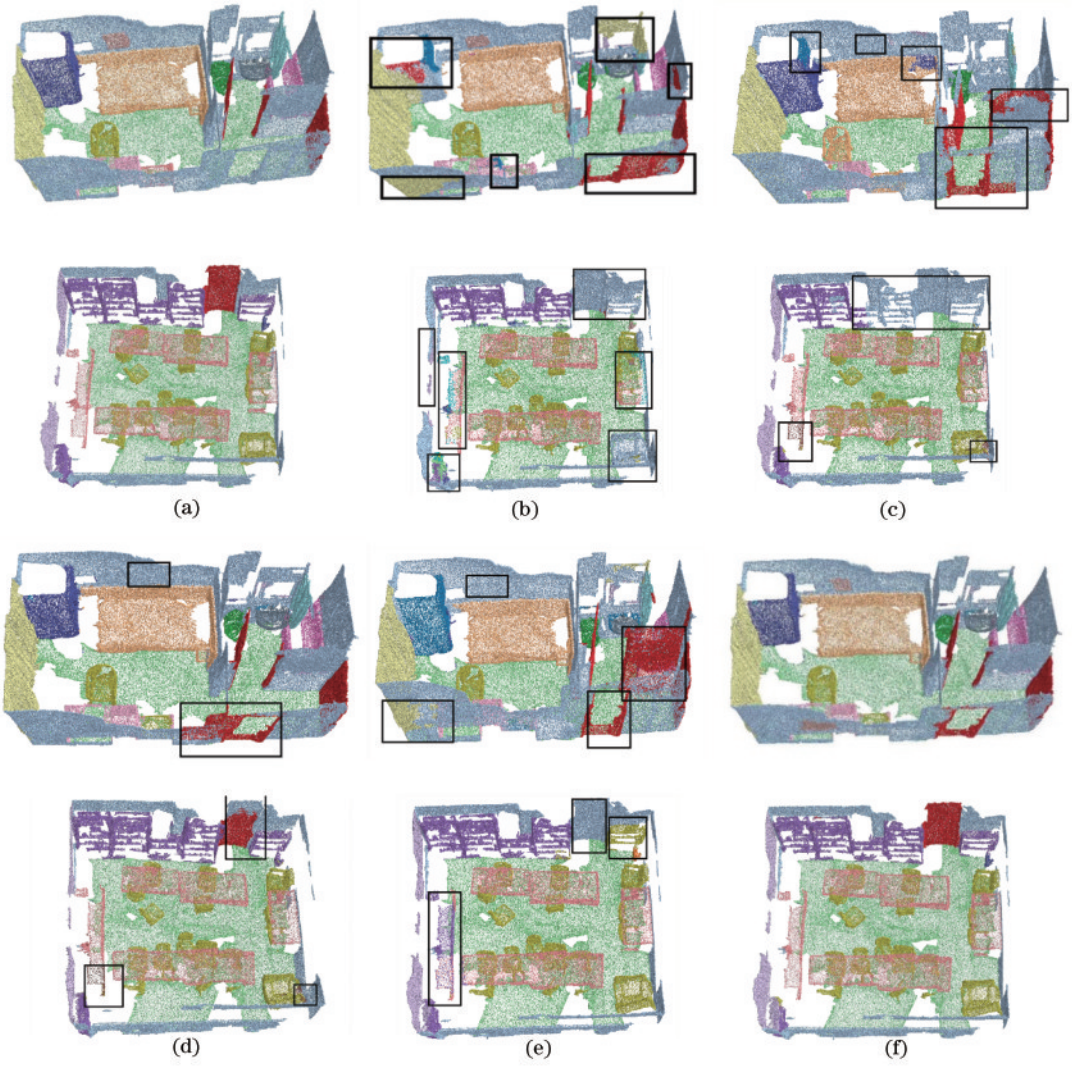

图 8　Scannet V2数据集分割可视化。(a)真值标签;(b) PointNet++; (c) FPConv; (d) SSCN; (e) Minkowski;(f)所提网络

Fig. 8　Scannet V2 dataset segmentation visualization. (a) True value label; (b) PointNet++; (c) FPConv; (d) SSCN; (e) Minkowski; (f) proposed network

操作,因此该方法在平坦区域表现良好,比如地板、床、沙发、墙壁等区域。但是对于墙壁上的门容易错分,由图 8可看出,门有3处过分割。而且在曲率大的区域分割精度较差,比如椅子、洗手池、浴缸和橱柜等场景。所提网络直接通过体素建立场景的结构化信息,有效利用场景的信息,因此在平坦区域和大曲率区域表现均比FPConv更好,尤其在椅子、洗手池、浴缸和橱柜等大曲率类别表现更好。

所提网络的MIOU相比于原始网络SSCN提升了1.0个百分点。SSCN由于只关注数据的稀疏性,而没有关注数据的全局上下文的联系,在门、画、窗户、橱柜和冰箱等依赖于场景的类别表现较差,会出现过分割或者分类错误的情况。而所提网络增加了多层自注意力机制,对全局信息的获取能力和对整体结构的判断能力增强,因此在复杂环境中的类别分割精度提升较大。由图 8可以看出,在会议室场景,SCCN对门存在过分割的情况,对橱柜存在欠分割的情况,而所提网络不存在。在卧室场景,SCCN将墙壁错误分割为门,且存在过分割的情况,而且并没有分割出墙壁上的画,而所提网络虽然存在对门的类别分割错误的情况,但是分割错误的区域更少、精度更高,而且成功分割出画的类别。这进一步证明了所提改进网络提高了对远距离信息的捕捉能力,提升了对复杂场景的分割精度。

所提网络相比于Minkowski[24],MIOU提升了1.2个百分点。Minkowski将卷积从二维扩展到四维,且使用高维条件随机场保障各个类别的一致性,从而提高各个类别的分割准确率。因此,可以看到,Minkowski对于洗手池、马桶、浴池和其他具有独特的形状,且在空间中独立的类别的分割精度较高,而对于门、窗户、画等在空间中与其他物体重合的类别的分割精度较低。从图 8可以看出,在会议室场景和卧室场景,所提网络在门、画、橱柜、窗帘等区域分割效果均比Minkowski好。

综上所述,所提网络虽然在空间相对独立和形状

独特的类别比如洗手池、马桶、浴池上的表现略差于Minkowski。但由于使用体素化建立了场景的结构化信息，且使用稀疏卷积维持数据的稀疏性，最后通过注意力机制加强了网络对全局上下文信息的获取能力，因此所提网络在容易分割的类别比如墙壁、地板、床等达到了与对比网络相似或更好的分割准确率，且在橱柜、椅子、桌子、门、窗户、画、窗帘等复杂场景下的类别达到了最高的分割准确率，MIOU也在比较的网络中最高。

3.6 S3DIS数据集实验结果

还在S3DIS数据集上的AREA 5上进行了实验。由于S3DIS规模点云规模更大、也更稀疏，可以进一步验证所提网络采用的稀疏卷积和注意力机制的有效性。表4是所提网络与其他网络在S3DIS数据集上的比较。不同网络在S3DIS数据集上的可视化效果如图9所示。

可以看出，相比于其他网络，所提网络在大场景S3DIS数据集上的MIOU最高，达到了70.5%，相比于PointNet提升了29.4个百分点，相比于KPConv提升了3.4个百分点。KPConv[25]需要手工设定核心点，在核心点加入位置偏移训练从而拟合点云的局部几何结构，这种方法无法从根本上解决卷积缺乏关联局部特征之间关系能力的问题。从表4和图9可以看到，所提网络在复杂区域表现比KPConv更好，例如大厅场景墙壁上的黑板和办公室场景书架后的墙壁。所提网络的MIOU比Minkowski提升了5.1个百分点。从图9可以看出，相比于Minkowski，所提网络在大厅场

表4 S3DIS AREA 5实验结果对比

Table 4 Comparison of experimental results of the S3DIS AREA 5 unit: %

Class	PointNet	KPConv	Minkowski	Proposed network
Calling	88.8	92.8	91.8	92.5
Floor	97.3	97.3	98.7	98.4
Wall	69.8	82.4	86.2	**89.4**
Beam	0.1	0.0	0.0	0.0
Column	3.9	23.9	34.1	**54.2**
Window	46.3	58.0	48.9	**61.2**
Door	10.8	69.0	62.4	65.1
Table	59.0	81.5	81.6	**82.1**
Chair	52.6	91.0	89.8	**92.0**
Sofa	5.9	75.4	47.2	**78.2**
Bookcase	40.3	75.3	74.9	74.2
Board	26.4	66.7	74.4	**75.2**
Clutter	33.2	58.9	58.6	54.4
MIOU	41.1	67.1	65.4	**70.5**

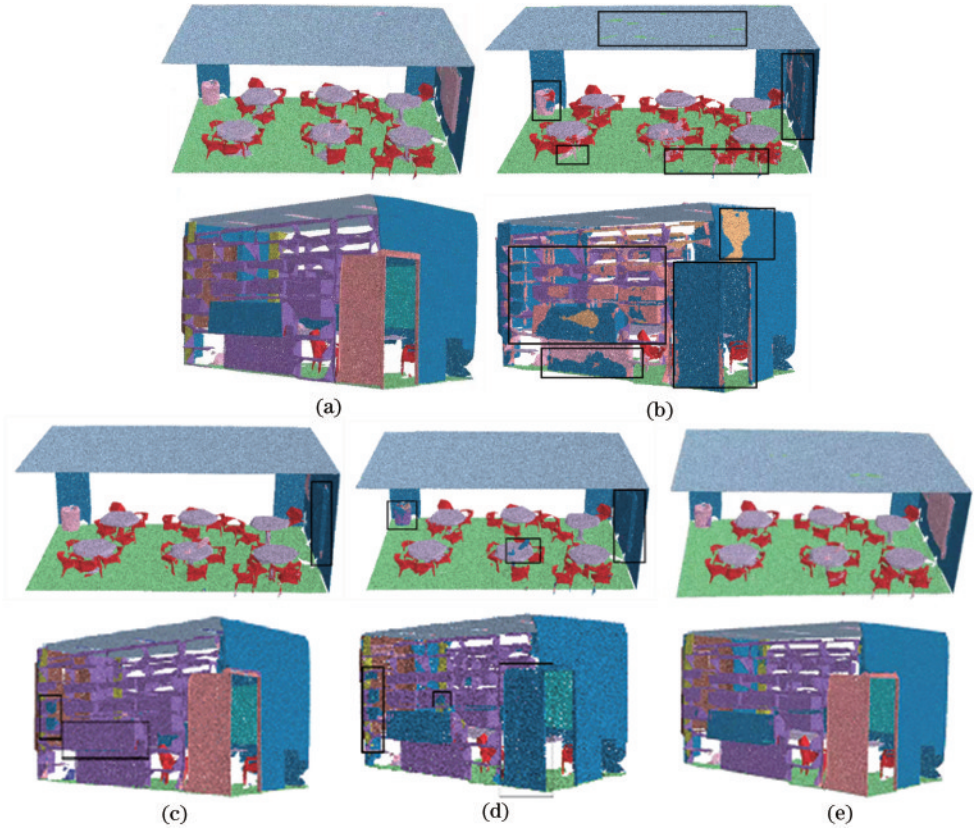

图9 S3DIS AREA 5数据集分割可视化。(a)真值标签；(b) PointNet；(c) KPConv；(d) Minkowski；(e)所提网络

Fig. 9 S3DIS AREA 5 segmentation visualization. (a) True value label; (b) PointNet; (c) KPConv; (d) Minkowski; (e) proposed network

景中的垃圾桶和黑板、办公室场景中的门和墙壁书架重合的复杂场景上表现更好。

综上所述，所提网络在大场景的S3DIS数据集上分割精度表现良好，在容易分割的类别比如墙壁、地板、吊顶等类别达到了与对比网络相似或更好的分割准确率，且在窗户、黑板和重合程度较高的书架和墙壁等与周围环境不易分割的类别上达到了最高的分割准确率，MIOU也在比较的网络中最高。

3.7 消融实验

为了对比初始主干网络SSCN、未采样前的Non Local Block网络和空间金字塔采样后的网络的性能，将采样后的Non Local Block插入网络的不同层，并在Scannet V2数据集上进行分割精度和前向推理时间的消融实验，分割精度的实验结果如表5所示。可以看出，加入空间金字塔采样注意力机制后的网络有效地缓解了卷积有限的局部感受野问题，增强了对全局上下文信息的获取能力。

表5 空间金字塔采样后Non Local Block插入不同层分割精度对比

Table 5 Comparison of segmentation accuracy of Non Local Block inserted into different layers after spatial pyramid sampling unit: %

Layer	SSCN	SSCN+Non Local Block	SSCN+SPSNB
1	70.821	71.034	71.034
2	70.821	71.342	71.214
3	70.821		71.421
4	70.821		71.825
5	70.821		71.641
6	70.821		71.322

从表5可以看出：直接在第1层添加Non Local Block的注意力机制模块后，MIOU提升了0.213个百分点；进一步将其扩展到原始稀疏卷积网络的高层，发现应用到第2层的MIOU增加0.521个百分点。这证明了稀疏卷积有限的局部感受野限制了网络的分割性能，添加注意力机制模块后提升了对远距离信息的捕捉能力，使分割精度提高。但是继续向上扩展之后计算量过大，占用计算资源过高导致无法训练。因此在本实验中不使用空间金字塔采样的Non Local Block模块仅能拓展到第2层。

通过空间金字塔采样之后的Non Local Block模块插入第2层之后MIOU相比于原网络提升了0.339个百分点，但是相比于未采样之前的有所下降，这是由于采样不可避免地导致信息丢失。但是继续向上扩展到第3层时，采样后的已经比采样前扩展到第2层的MIOU更高。扩展到第4层之后MIOU最高，达到了71.825%。但是继续向上扩展可以发现，MIOU反而下降了。这是因为特征提取网络层数越往上，数据维度越低，数据量越大，空间金字塔采样之后不足以描述原始数据的特征，从而导致分割性能下降。网络的前向推理时间消融实验结果如表6所示。

表6 空间金字塔采样后Non Local Block插入不同层前向推理时间对比

Table 6 Comparison of time for Non Local Block insertion into different layers of forward reasoning after a spatial pyramid sampling unit: ms

Layer	SSCN	SSCN+Non Local Block	SSCN+SPSNB
1	73.10	73.32	73.32
2	73.10	77.34	73.85
3	73.10		74.42
4	73.10		75.35
5	73.10		76.52
6	73.10		77.82

网络的前向推理时间是在Scannet V2数据集的验证集上取平均得到的。从表6可以发现，Non Local Block模块加入网络的第2层之后，相比于原始网络的前向推理时间增加了4.24 ms，从第3层开始由于计算量过大无法训练，因此为空值。在加入金字塔采样之后的SPSNB模块中，注意力机制模块可以顺利拓展到网络的第6层，且网络的前向推理时间相比于采样前明显下降，在分割性能最高的第4层，采样后网络的前向推理时间相比于采样前第2层还低了1.99 ms。

4 结 论

提出一种融合空间金字塔采样的Non Local Block模块和稀疏卷积的三维点云分割方法。首先对输入网络的点云进行体素化，建立起点云的结构化信息，再输入U-Net通过稀疏卷积保持体素的稀疏结构，并通过改进之后的Non Local Block模块加强全局和局部几何特征之间的联系，提高网络对信息的远距离依赖关系的获取能力。最后解体素化，从体素恢复到每个点的语义信息。通过扩展Non Local Block模块的应用层数，可以用较少的计算资源提升网络对特征的提取能力，实现点云的高效分割。

但是，网络依然存在待改进的地方，对在空间上接近的、具有相似的几何结构的类别会出现过分割或者欠分割的现象。另外网络模型训练到收敛需要一周的时间，因此后续会针对如何提高网络的分割能力和加快网络训练的收敛时间进行相关工作。

参 考 文 献

[1] 赵亮，胡杰，刘汉，等. 基于语义分割的深度学习激光点云三维目标检测[J]. 中国激光，2021，48(17)：1710004.
Zhao L，Hu J，Liu H，et al. Deep learning was based on

semantic segmentation for three-dimensional object detection from point clouds[J]. Chinese Journal of Lasers, 2021, 48(17): 1710004.

[2] Wang W, Xu Y, Ren Y C, et al. Parsing of urban facades from 3D point clouds based on a novel multi-view domain[J]. Photogrammetric Engineering & Remote Sensing, 2021, 87(4): 283-293.

[3] Kundu A, Yin X Q, Fathi A, et al. Virtual multi-view fusion for 3D semantic segmentation[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12369: 518-535.

[4] 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, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 77-85.

[5] Qi C R, Yi L, Su H, et al. PointNet++: deep hierarchical feature learning on point sets in a metric space [EB/OL]. (2017-06-07)[2022-08-06]. https://arxiv.org/abs/1706.02413.

[6] Xu Y F, Fan T Q, Xu M Y, et al. Spidercnn: deep learning on point sets with parameterized convolutional filters[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11212: 87-102.

[7] Zhang Z Y, Hua B S, Chen W, et al. Global context-aware convolutions for 3D point cloud understanding[C]// 2020 International Conference on 3D Vision (3DV), November 25-28, 2020, Fukuoka, Japan. New York: IEEE Press, 2020: 210-219.

[8] Maturana D, Scherer S. VoxNet: a 3D Convolutional Neural Network for real-time object recognition[C]// 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), September 28-October 2, 2015, Hamburg, Germany. New York: IEEE Press, 2015: 922-928.

[9] Kumar K S C, Al-Stouhi S. Multi-scale voxel class-balanced ASPP for LIDAR pointcloud semantic segmentation[C]//2021 IEEE Winter Conference on Applications of Computer Vision Workshops, January 5-9, 2021, Waikola, HI, USA. New York: IEEE Press, 2021: 117-124.

[10] Park J, Kim C, Kim S, et al. PCSCNet: Fast 3D semantic segmentation of LiDAR point cloud for autonomous car using point convolution and sparse convolution network[J]. Expert Systems With Applications, 2023, 212: 118815.

[11] Graham B. Spatially-sparse convolutional neural networks [EB/OL]. (2014-09-22)[2022-08-06]. https://arxiv.org/abs/1409.6070.

[12] Riegler G, Ulusoy A O, Geiger A. OctNet: learning deep 3D representations at high resolutions[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6620-6629.

[13] Engelcke M, Rao D, Wang D Z, et al. Vote3Deep: Fast object detection in 3D point clouds using efficient convolutional neural networks[C]//2017 IEEE International Conference on Robotics and Automation, May 29-June 3, 2017, Singapore. New York: IEEE Press, 2017: 1355-1361.

[14] Graham B, Engelcke M, Maaten L V D. 3D semantic segmentation with submanifolds sparse convolutional networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 9224-9232.

[15] Wang X L, Girshick R, Gupta A, et al. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7794-7803.

[16] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-22, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-7141.

[17] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11211: 3-19.

[18] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation [M]//Navab N, Hornegger J, Wells W M, et al. Medical image computing and computer-assisted intervention-MICCAI 2015. Lecture notes in computer science. Cham: Springer, 2015, 9351: 234-241.

[19] He T, Shen C H, van den Hengel A. DyCo3D: robust instance segmentation of 3D point clouds through dynamic convolution[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 19-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 354-363.

[20] Zhu Z, Xu M D, Bai S, et al. Asymmetric non-local neural networks for semantic segmentation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2019: 593-602.

[21] Dai A, Chang A X, Savva M, et al. ScanNet: richly-annotated 3D reconstructions of indoor scenes[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 2432-2443.

[22] Armeni I, Sax S, Zamir A R, et al. Joint 2D-3D-semantic data for indoor scene understanding[EB/OL]. (2017-02-03)[2022-08-06]. https://arxiv.org/abs/1702.01105.

[23] Xu M T, Ding R Y, Zhao H S, et al. PAConv: position adaptive convolution with dynamic kernel assembling on point clouds[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June

20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2018: 3172-3181.

[24] Choy C, Gwak J, Savarese S. 4D spatio-temporal ConvNets: minkowski convolutional neural networks [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 3070-3079.

[25] 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), June 15-20, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2019: 6410-6419.