DOI: 10.19650/j.cnki.cjsi.J2108289

双边特征聚合与注意力机制点云语义分割*

王溪波[1],曹士彭[1,2],赵怀慈[2],刘鹏飞[2],邰炳昌[3]
(1.沈阳工业大学信息科学与工程学院 沈阳 110870; 2.中国科学院沈阳自动化研究所光电信息处理重点实验室 沈阳 110016; 3.卓越新时代认证有限公司 沈阳 110000)

摘 要:机器视觉是环境感知的重要手段之一,是自动驾驶、机器人、工业检测等领域的研究热点,而点云数据的精细分析是其中的一项关键技术。针对大尺度真实场景点云数据分割精度低的问题,提出了一种适用于点云数据语义分割的网络结构。首先,构建了一个双边特征聚合结构,通过分别处理点云的几何信息和语义信息,达到充分利用点云特征信息的目的。其次,使用近邻特征的高维空间相关性计算点与点之间的相互作用,进行局部邻域的上下文信息增强。提出了一种混合池化结构代替最大值池化,减少信息损失,使用横向跨层池化连接来增强特征多样性。最后,引入注意力机制提取全局特征,滤除尺度噪声,增强特征在空间上的表现力。实验结果表明,该方法在大尺度真实场景点云数据集 S3DIS 上的平均交并比为 68.2%,平均准确率为 80.7%,比 PointNet 提高了 20.6% 和 14.5%,客观指标优于已有的代表性方法。

关键词: 机器视觉;点云语义分割;双边特征聚合;跨层混合池化;注意力机制
中图分类号: TP391.41 TH74 **文献标识码**: A **国家标准学科分类代码**: 520.20

Semantic segmentation of point cloud via bilateral feature aggregation and attention mechanism

Wang Xibo[1], Cao Shipeng[1,2], Zhao Huaici[2], Liu Pengfei[2], Tai Bingchang[3]
(*1. School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China; 2. Key Laboratory of Optical-Electronic Information Processing, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China; 3. Excellence New Era Certification Limited Company, Shenyang 110000, China*)

Abstract: Machine vision is one of the important measure manners for environmental perception. It is a research hotspot in the fields of automatic driving, robot, industrial detection and so on. The fine analysis of point cloud data is one of the key technologies. To solve the problem of low segmentation accuracy of large-scale point cloud data of real scene, a bilateral feature aggregation network architecture for semantic segmentation of the point cloud is proposed. Firstly, a bilateral feature aggregation module is formulated to aggregate local features by processing the geometric information and semantic information of the point cloud. The aim is to make full use of the feature information of the point cloud. Secondly, the high-dimensional spatial correlation of nearest neighbor features is used to calculate the impact between points. The context information of local neighborhood is enhanced. A hybrid-pooling architecture is proposed to replace the max-pooling to reduce the information loss of max-pooling, and the horizontal skip connection pooling is used to enhance feature diversity. Finally, an attention module is introduced to extract global features, which can filter scale noise and enhance the spatial expressiveness of features. Experimental results show that the mean intersection over union of the proposed method is 68.2%, and the mean accuracy is 80.7%. These two values are 20.6% and 14.5% higher than those of the PointNet. The objective indicator is better than the existing representative methods.

Keywords: machine vision; point cloud semantic segmentation; bilateral feature aggregation; skip connection hybrid-pooling; attention mechanism

收稿日期:2021-07-23 Received Date: 2021-07-23
*基金项目:国家自然科学基金(U2013210)项目资助

0 引 言

机器视觉是机器感知的重要方式之一。近年来随着自动驾驶、增强现实等应用领域的快速发展,三维视觉研究取得了极大进展。相比于二维图像,三维数据所包含的深度信息可以使机器更好地识别真实世界。点云是一种基本的三维数据表示形式。大型复杂场景点云数据的细粒度分析[1-3]在自动驾驶、增强现实、机器人等应用中有着巨大的潜力。比如,在工业检测中,对利用三维扫描仪得到的点云数据进行分析处理;为移动机器人提供语义地图,提高其目标感知与场景理解能力等任务目标。针对上述任务,本文重点研究面向真实点云场景的语义分割。

点云数据与普通图像数据有较大差异,其分布有稀疏性、不规则性、无序性等特点。因此,点云数据的语义分割任务比二维图像的语义分割更具挑战性。特别是对于从现实世界中收集的数百万甚至数十亿个点组成的大尺度场景点云数据,对算法效率、内存占用、网络捕捉复杂结构的能力等方面要求更高。

在对点云数据的处理方式上,一些工作[4-6]基于传统算法改进,如随机采样一致性算法(random sample consensus, RANSAC),基于密度的聚类算法(density based spatial clustering of applications with noise, DBSCAN),区域增长算法(region growing)等。但传统算法依赖于点云质量,稳定性差。近年来发展较快的是利用深度学习方法进行点云处理分析,主要包括以下 3 种:1)基于投影的网络[7-9]:点云数据的不均匀性导致无法直接对点云使用卷积神经网络。为了利用目前已有的二维卷积神经网络,一些工作选择将三维点云投影到二维图像上,再把二维图像作为训练数据输入网络。但在投影过程中可能会导致几何信息丢失,且这种方法缺乏捕捉非局部几何特征的能力;2)基于体素化的网络[10-12]:针对点云的无序、不规则性,将无序点云体素化为有序体素块,再使用三维卷积神经网络处理有序体素块。该方法的主要局限是计算成本过高,尤其是在处理大规模点云时,无法满足实际应用。而且体素块的体积设置会影响最终的分割效果,体积设置过小时,会由于点云的稀疏性导致有空体素块,浪费计算成本;体积设置过大会导致计算成本过高。没有有效的自动设置体素块体积的方法,过分依赖经验。此外,体素化会导致空间分辨率和细粒度几何形状信息的损失;3)基于点云的网络[1,13-15]:该方法一般用多层感知器提取点云的逐点特征。文献[1]是使用神经网络直接处理点云数据的开创性工作,网络的每一层输入为上一层输出的几何嵌入特征,但其使用的最大值池化会导致信息损失过大。文献[13]进一步用 k 近邻或球半径方式分组划分局部区域,但其提取的几何信息不够丰富,且忽略了点云的语义信息,导致网络不能充分捕捉物体局部细节,最终分割结果不理想。文献[14]利用动态图卷积提取特征并寻找近邻关系。文献[15]将近邻点的相对位置投影到卷积权重上,直接使用邻域和局部中心点之间的关系来学习卷积的动态权重。文献[14]、[15]都缺少对全局特征的利用,导致网络的空间感知能力随网络加深而逐渐混乱。

投影或体素化的方法对于实际应用来说有以下局限性:1)需要耗时、耗计算资源的处理步骤;2)其生成的中间表示可能会丢失部分物体本身的几何信息及周围环境的背景信息。

为避免上述局限,本文设计了一种端到端网络直接处理用于细粒度分析的点云数据。而且,针对上述真实大尺度点云数据语义分割面临的问题,本文提出如下解决方案:1)针对网络捕捉局部细节能力有限的问题:通过构建双边结构同时处理点云的几何特征和语义特征,将两种信息融合,再利用特征的高维空间相关性进行特征信息增强,以达到强化网络捕捉局部细节能力的目的。其中几何特征可以丰富网络的基本几何感知,增强局部特征在高维特征空间的泛化能力。语义特征是高层次知识表示,可以赋予网络高度泛化的类别特征感知能力;2)针对最大值池化导致信息损失过大的问题:在最大值池化的基础上引入注意力池化,构建混合池化方式弥补最大值池化信息损失,再用跨层池化拼接方式增加特征信息多样性;3)针对网络缺乏全局特征的问题:目前大部分工作都将关注的重点放在如何更有效地聚合局部特征上,但在语义分割过程中,全局特征可以表征不同位置点之间的上下文关系。因此,本文方法引入一个注意力模块分别从空间及通道维度为点云中的每个点聚合全局特征信息。

1 理论分析

本文方法所设计的网络结构如图 1 所示。网络由编码层、解码层以及注意力模块组成。编码层中,由双边信息增强单元、近邻特征高维空间相关性及跨层混合池化单元组成双边池化模块。级联 5 个双边池化模块组成双边特征聚合模块,构成整个编码层,以编码点云的局部特征。在解码层中,用基于距离的最近邻点进行插值上采样,逐层增大解码层中传播的特征向量,并用跨层连接方式将编码过程中产生的同尺寸特征向量与上采样产生的特征向量进行拼接,以弥补上采样过程中可能的信息损失。最后,在进行语义预测前引入注意力机制,从空间维度及通道维度两种维度提取全局特征信息。

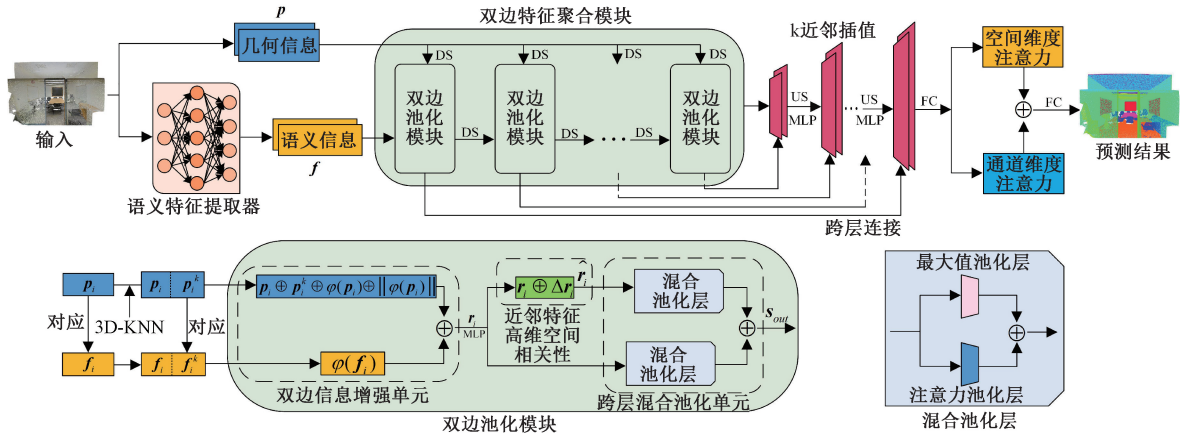

图 1　网络整体结构

Fig. 1　Overall network architecture

1.1　最远点采样与 k 近邻算法融合

为了提取局部特征,首先要构建局部邻域。对于一组大规模点云 $N=\{x_1,x_2,\cdots,x_N\}$,其中 x_N 表示点云集合中的每个点,首先使用基于欧式距离的最远点采样算法选择 N 中的一组点作为中心点,记为 $\{p_1,p_2,\cdots,p_n\}$,使得 $p_i(1\leqslant i\leqslant n)$ 在度量空间中是距离点集 $\{p_1,p_2,\cdots,p_{i-1}\}$ 的最远点。与采用随机采样算法找中心点相比,在给定相同数量中心点的情况下,最远点采样能够更好的覆盖整个点云,克服点云稀疏性的影响。对于中心点 $p_i(1\leqslant i\leqslant n)$,使用基于欧氏距离度量的 k 近邻算法找到它的 k 个近邻点 $\forall p_i^k\in N$。由于点云的稀疏性,若用球半径方法构建局部邻域,可能会发生球半径内点云数量过少的情况,而 k 近邻方法可以有效处理该问题。

1.2　双边特征聚合模块

点云的特征信息主要包括两方面:1)几何信息:包含点云的形状信息和空间位置信息,可以增强网络的空间感知能力;2)语义信息:是泛化的高级特征表示,可以增强网络的类别特征理解能力。为了充分利用这两种信息,本文构建了双边特征聚合模块。

1)语义特征提取器

一些数据集除点云的三维空间坐标外,还会包含诸如点云的 RGB 颜色、光照强度等其他信息。为了创建整个场景的整体印象,使用多层感知器作为语义特征提取器,将数据集提供的所有信息输入多层感知器中融合,并将输出特征提升为特定维度,获取初步的语义知识信息。多层感知器包含输入层、隐藏层、输出层,每层有一定数量的神经元,其优点在于它可以灵活的在高维嵌入空间中表示特征。图 1 展示了语义特征提取器的结构及其传入双边结构的语义特征信息 $\boldsymbol{f}$。多层感知器 M 的操作细节如式(1)所示。

$$M=\zeta_\theta(\mathrm{BN}(\tau_{1\times1}^m(\boldsymbol{x}))) \tag{1}$$

其中,ζ 为参数化的非线性激活函数,本文方法选择 Leaky ReLU 激活函数,θ 为卷积核中可学习的参数集合,BN 为批归一化处理,τ 为卷积操作,其上标 m 表示多层感知器的输出维度,下标 1×1 表示卷积核尺寸,$\boldsymbol{x}$ 表示多层感知器的输入。

2)双边信息增强单元

本文通过构建一组双边结构分别处理点云的几何、语义特征信息,并使其互相补充、增强。在提供的三维空间几何信息不完备的情况下,会弱化局部特征在高维特征空间的泛化能力。针对该问题,将中心点及近邻点在三维欧氏空间中的绝对坐标、近邻点相对于中心点的局部相对坐标与两者间的欧氏距离融合在一起,作为双边信息增强单元的几何约束,以提供给网络更丰富的几何信息。点云的绝对坐标可以获取全局信息,增强网络的空间感知定位能力。而相对坐标所提取的特征信息可以增强网络应对点云无序性的能力,两者间的欧氏距离作为补充信息。最终将点的局部几何特征表示为 $\hat{\boldsymbol{p}}_i$,公式如下:

$$\varphi(p_i)=p_i^k-p_i,\ \varphi(p_i)\in R^3 \tag{2}$$

$$\hat{\boldsymbol{p}}_i=M(p_i\oplus p_i^k\oplus\varphi(p_i)\oplus\|\varphi(p_i)\|),\ \hat{p}_i\in R^{\frac{d_{out}}{2}} \tag{3}$$

其中,p_i、p_i^k 分别表示中心点、近邻点在三维空间中的 $x-y-z$ 绝对坐标,$-$ 表示空间坐标相减,以构造近邻点关于中心点的局部相对坐标,$\oplus$ 表示特征拼接操作,$\|\cdot\|$表示欧氏距离,M 表示多层感知器,d_{out} 是 5 个级联的双边池化模块各自的输出的特征向量维度,分别设定为[16, 64, 128, 256, 512]。

在低维特征空间中,将局部邻域的语义特征记为 $\hat{\boldsymbol{f}}_i$,

其构成方式如下式所示：

$$\varphi(\boldsymbol{f}_i)=\boldsymbol{f}_i^k-\boldsymbol{f}_i,\varphi(\boldsymbol{f}_i)\in R^{d_f} \tag{4}$$

$$\hat{\boldsymbol{f}}_i=M(\boldsymbol{f}_i\oplus\varphi(\boldsymbol{f}_i)),\hat{\boldsymbol{f}}_i\in R^{\frac{d_{out}}{2}} \tag{5}$$

其中，$\boldsymbol{f}_i$、$\boldsymbol{f}_i^k$ 分别为中心点 p_i、近邻点 p_i^k 在语义特征空间中所对应的语义信息，－表示特征维度相减，以构造局部语义上下文信息，d_f 是语义特征 $\boldsymbol{f}_i$、$\boldsymbol{f}_i^k$ 的维度，d_{out} 是整个编码模块最终输出的特征向量维度。

最后，为了充分利用所有的几何及语义信息，对于每个局部邻域，将几何特征 $\hat{\boldsymbol{p}}_i$ 与低维语义特征 $\hat{\boldsymbol{f}}_i$ 拼接，获得局部邻域上下文信息 $\boldsymbol{r}_i$，公式如下：

$$\boldsymbol{r}_i=\hat{\boldsymbol{p}}_i\oplus\hat{\boldsymbol{f}}_i,\ \boldsymbol{r}_i\in R^{d_{out}} \tag{6}$$

3）近邻特征的高维空间相关性

在同一局部区域内的各个点之间会产生扰动，而在传统算法中没有考虑高维特征空间内中心点与近邻点特征之间的关系，造成了空间或语义的间隙。本文通过在高维特征空间中密集连接局部邻域内的中心点与近邻点，以自适应特征调整的方式，基于局部区域特征寻找点与点之间的相互作用关系，调整局部特征，增强局部邻域的上下文信息，更好地表示局部区域。网络结构如图 1 所示，公式如下：

$$\hat{\boldsymbol{r}}_i=\boldsymbol{r}_i\oplus\Delta\boldsymbol{r}_i,\ \hat{\boldsymbol{r}}_i\in R^{d_{out}} \tag{7}$$

其中，$\Delta\boldsymbol{r}_i$ f_{imp} 的计算公式如下：

$$\Delta\boldsymbol{r}_i=\sum_{j=1}^{k}f_{imp}(\boldsymbol{r}_i,\boldsymbol{r}_i^j) \tag{8}$$

$$f_{imp}(\boldsymbol{r}_i,\boldsymbol{r}_i^j)=\mathrm{M}(\boldsymbol{r}_i^j-\boldsymbol{r}_i) \tag{9}$$

式中：$\boldsymbol{r}_i$、$\boldsymbol{r}_i^j$ 分别表示中心点、近邻点在高维特征空间中对应的特征向量，$\hat{\boldsymbol{r}}_i$ 表示增强后的局部特征，$\Delta\boldsymbol{r}_i$ 表示局部邻域内点之间互相影响的偏移量，k 表示所选取的近邻点的个数，函数 f_{imp} 表示计算中心点与近邻点之间的影响因子。

4）跨层混合池化单元

为处理点云的无序性，文献[1]开创性地使用最大值池化来聚合邻域点集合的特征信息，但最大值池化对信息损失较大。受文献[2]启发，本文引入注意力机制，构造注意力池化层，自动学习重要的局部特征。与最大值池化层混合使用，以弥补其信息损失，获得更精确的细粒度局部邻域信息。混合池化层结构如图 1 所示。

（1）注意力池化层

计算注意力权重：给定一组局部特征向量 $\boldsymbol{r}_i=\{\boldsymbol{r}_i^1\cdots\boldsymbol{r}_i^j\cdots\boldsymbol{r}_i^k\}$，构造函数 α 为其中的每个特征向量 $\boldsymbol{r}_i^j$ 学习出一个注意力权重值，令网络能够自适应学习重要特征。该函数由共享权重的多层感知器与一个 softmax 函数构成，公式如下：

$$\boldsymbol{w}_i^k=\alpha(\boldsymbol{r}_i^j,\boldsymbol{W}) \tag{10}$$

其中，$\boldsymbol{w}_i^k$ 是 k 个近邻特征的一组可学习的注意力权重，$\boldsymbol{W}$ 是共享权重的多层感知器所包含的权重。

聚合近邻特征：α 函数为 k 个近邻特征学习出的注意力权重可以视为网络自动选择重要近邻特征的软掩膜。这些带有权重值的特征最后以加权求和的方式聚合起来，公式如下：

$$\mathrm{attentive}(\boldsymbol{r}_i)=\sum_{j=1}^{k}(\boldsymbol{r}_i^j\cdot\boldsymbol{w}_i^k) \tag{11}$$

（2）混合池化

如图 1 所示，对于一组局部特征向量，一方面使用最大值池化聚合出整个局部邻域的最显著特征，用于表述该区域；另一方面，使用注意力池化方式学习局部邻域的高维质心来获得更多细节信息，对整组特征向量进行细粒度处理。然后将两种池化层的输出拼接在一起，作为整个单元的输出。将混合池化单元记为 H，公式如下：

$$\mathrm{H}(\boldsymbol{r}_i)=\max_k(\boldsymbol{r}_i)\underset{\boldsymbol{w}_i^k}{\oplus}\mathrm{attentive}(\boldsymbol{r}_i) \tag{12}$$

式中：max( )表示最大值池化。

（3）跨层双池化

浅层局部特征含有局部结构信息，但特征泛化能力较弱。深层局部特征含有高级语义信息，特征泛化能力强，但缺乏局部结构信息。受残差网络[16]启发，将不同深度的网络信息拼接，以增加特征信息多样性，并避免梯度消失或爆炸。但若直接将两层网络的输出跨层连接，会导致冗余信息过多，浪费计算成本。所以，本文选择将不同深度网络的输出分别通过两个混合池化层聚合出代表性信息后，再进行跨层连接，跨层双池化结构细节如图 1 所示。

对于浅层局部邻域特征 $\boldsymbol{r}_i$ 与利用近邻特征高维空间相关性增强后所得到的深层局部邻域特征 $\hat{\boldsymbol{r}}_i$，分别用两个混合池化层聚合出代表性信息后进行信息融合。产生的新局部邻域特征记为 $\boldsymbol{s}_{out}$，同时它也是双边池化模块的最终输出，公式如下：

$$\boldsymbol{s}_{out}=M(H(\boldsymbol{r}_i)\oplus H(\hat{\boldsymbol{r}}_i)),\ \boldsymbol{s}_{out}\in R^{d_{out}} \tag{13}$$

$\boldsymbol{s}_{out}$ 会作为高级语义信息传入级联的下一个双边池化模块，直到级联的 5 个双边池化模块都编码完成，最后会作为整个编码层的输出传入解码层。

1.3 损失函数

为了保持密集邻域的几何完整性，本文将近邻区域作为一个整体来考虑，而不是考虑单个近邻点。将几何信息视为对近邻区域的约束，与近邻点位置拼接在一起，公式如下：

$$\widetilde{\boldsymbol{p}}_i^k=p_i^k\oplus M(p_i\oplus p_i^k\oplus(p_i-p_i^k)\oplus\|p_i-p_i^k\|) \tag{14}$$

其中，$\widetilde{\boldsymbol{p}}_i^k$ 表示带有几何约束的近邻区域，其他参数含义如式(3) 所示。再通过最小化 l_2 距离来鼓励带约束

近邻区域的几何中心点靠近局部区域的中心点位置,公式如下:

$$\mathcal{L}(p_i) = \left\| \frac{1}{k}\sum_{i=1}^{k} \widetilde{\boldsymbol{p}}_i^k - p_i \right\|_2 \tag{15}$$

1.4　k 近邻插值方法

语义分割任务的目标是获得原始点云集合中每一个点的分类标签。所以,将编码层输出的特征映射恢复为输入点集的原始尺寸在该任务中是很重要的。如图 1 所示,本文在解码层中使用基于距离的最近邻插值上采样和跨层特征连接。在上采样过程中,首先使用基于欧氏距离的 k 近邻算法为每个查询点找到 k 个最近邻点,然后通过对近邻点逆距离加权求平均值方式对点的特征映射进行上采样,计算公式如下:

$$\boldsymbol{f}^d(p_i) = \frac{\sum_{i=1}^{k} (d(p_i)\boldsymbol{f}_{p_i}^m)}{\sum_{i=1}^{k} d(p_i)}, \tag{16}$$

其中,

$$d(p_i) = \frac{1}{dis(p_i, p_i^k)^2}, \tag{17}$$

$\boldsymbol{f}^d$ 表示新生成的更大尺寸的特征图,$\boldsymbol{f}_{p_i}^d$ 是近邻点的特征向量,p_i 是中心点,p_i^k 是近邻点,m 表示特征维度,$dis(\,)^2$ 函数计算欧氏距离,本文中近邻点数 k 取 3。最后,通过跨层连接将上采样的特征映射与编码层产生的对应尺寸的中间特征拼接起来,再用多层感知器进行信息融合。

1.5　注意力模块

本文之前的所有工作都只挖掘了点的局部关系。而对于语义分割任务,全局信息在确定每个单独点的类别标签时也很重要,因为两个在空间上偏离很大的点也可能属于同一个语义类别。此外,对于高维特征,特征通道之间也存在相互依赖关系。因此,为了捕捉每个点的全局上下文信息,抑制无意义的局部结构特征,本文引入注意力模块,从空间维度和通道维度[3]两种维度获取点与点之间的全局关系。

1)空间维度注意力

为了在点与点之间建立丰富的全局上下文关系,使用空间维度注意力模块自适应地聚合局部特征的空间上下文信息,将不同空间位置的点对相互之间的影响因子作为权重系数,加权求和。给定来自解码器的特征向量 $\boldsymbol{F} \in R^{N\times d}$,首先将其送入两个全连接层内,分别获得两个新的特征向量 $\{\boldsymbol{F}', \boldsymbol{F}''\} \in R^{N\times d}$,其中 N 是点云数量,d 是特征维度。再利用 $\boldsymbol{F}'$、$\boldsymbol{F}''$ 求得归一化后的空间维度注意力权重 v_{ij},其表示点云中点 j 对点 i 的影响因子,公式如下:

$$v_{ij} = \mathrm{softmax}(\boldsymbol{F}_i' \cdot \boldsymbol{F}_j''), \tag{18}$$

之后将 $\boldsymbol{F}$ 送入另一个全连接层,生成一个新的特征向量 $\boldsymbol{D} \in R^{N\times d}$。最终,获取空间维度注意力后的输出特征记为 $\hat{\boldsymbol{F}} \in R^{N\times d}$,公式如下:

$$\hat{\boldsymbol{F}}_i = \sum_{j=1}^{N} (v_{ij}\boldsymbol{D}_j) \oplus \boldsymbol{F}_i, \tag{19}$$

此时,全局空间结构信息就通过注意力方式与每个点所包含的特征向量聚合在一起。

2)通道维度注意力

通道维度注意力建模特征通道之间的依赖关系,从而提高特征的可辨别性。获取方式与空间维度注意力类似,但针对的是特征通道结构信息,而不是点的坐标位置信息。最终输出的特征是通过将全局通道结构信息与每个单通道信息聚合在一起而获得的,记为 $\widetilde{\boldsymbol{F}} \in R^{N\times d}$。

最后将 $\hat{\boldsymbol{F}}$ 与 $\widetilde{\boldsymbol{F}}$ 进行信息融合,便获得结合了局部与全局特征信息的每个点的语义标签。通过从全局角度进行的注意力操作后,每个点的特征信息都会被赋予全局信息,这样就可以最大程度的利用点之间的复杂关系,生成更准确的分割结果。

2　实验验证

2.1　数据集

在该项工作中,本文的目标是真实点云场景的语义分割。为了验证本文提出的方法,本文在真实场景的三维基准数据集 S3DIS[17] 上进行了实验。斯坦福大尺度室内三维场景(stanford 3D large-scale indoor spaces, S3DIS)数据集是从室内工作环境中采集的。该数据集包含 6 个子区域,每个区域包含 50 个不同的房间。根据房间大小,构成每个房间的点云数据从 50 万到 250 万不等。所有的点云都提供了三维坐标与颜色信息,并被标记为 13 种语义类别之一。

2.2　实验设置

对于 S3DIS 数据集,训练时的批处理大小为 2,每次输入网络的点云数量约为 5×2^{13} 个点。本文方法在单个 GeForce RTX 2080Ti GPU 上训练 100 轮。使用 Adam 优化算法来最小化损失函数。学习率从 0.01 开始,每 10 轮后以 0.5 的速率衰减。使用 Linux 操作系统,基于 Python 和 Tensorflow 平台实现该工作。

2.3　消融实验

表 1 展示了在 S3DIS 数据集上单独对某些结构进行测试的消融实验结果,以说明本文方法所设计模块的有效性。评估指标采用平均交并比(mIoU)。计算公式如下:

$$\mathrm{mIoU} = \frac{1}{k}\sum_{i=0}^{k-1} \frac{p_{ii}}{\sum_{j=0}^{k-1} p_{ij} + \sum_{j=0}^{k-1} p_{ji} - p_{ii}} \tag{20}$$

其中,方法 1 是基础模型,未使用双边增强单元,仅为网络输入几何信息,未使用语义信息,且未使用高维空间相关性、跨层池化结构,池化方式仅为最大值池化。方法 2 使用了双边信息增强单元,它只是最大程度地聚合了基本的局部几何信息和语义信息,相比方法 1 有明显提升。方法 3、4 是与 2 进行不同池化方式的效果对比,证明混合池化的方式效果最好。方法 5 在方法 4 的基础上添加了跨层池化方式。方法 6 又添加了高维空间相关性进行信息增强。方法 7 添加了注意力模块,是本文所提出的完整网络架构,效果最好。在该项消融实验中,可以看出本文所提出的各个单元是如何相互补充以达到最佳效果的。

表 1 消融实验结果

Table 1 Results of ablation studies

方法	双边信息增强单元	跨层池化	高维空间相关性	池化方式	注意力模块	mIoU/%
1	×	×	×	最大值池化	×	53.40
2	√	×	×	最大值池化	×	65.71
3	√	×	×	注意力池化	×	65.91
4	√	×	×	混合池化	×	66.51
5	√	√	×	混合池化	×	67.48
6	√	√	√	混合池化	×	67.52
7	√	√	√	混合池化	√	68.22

2.4 实验结果对比

1)评估指标对比

表 2 所示为本文与其他网络对 S3DIS 数据集进行语义分割的评估指标对比。评估指标采用平均交并比(mIoU)、类别平均准确率(mAcc)、总体精度(OA)。假设共有 k 个类别,i 表示真实类别,j 表示预测类别。定义 p_{ii} 表示类别预测正确的点的数量,p_{ij}、p_{ji} 分别表示假负、假正的点的数量。则 OA 的计算公式表示为:

$$\mathrm{OA}=\frac{\sum_{i=0}^{k-1}p_{ii}}{\sum_{i=0}^{k-1}\sum_{j=0}^{k-1}p_{ij}} \tag{21}$$

mAcc 的计算公式表示为:

$$\mathrm{mAcc}=\frac{1}{k}\sum_{i=0}^{k-1}\frac{p_{ii}}{\sum_{j=0}^{k-1}p_{ij}} \tag{22}$$

由表 2 可知,本文在 mAcc、mIoU 方面取得了更优的性能表现。相比于对比网络,本文最大的不同之处在于构建双边结构分别处理点云的几何、语义信息,两种信息互补,相互促进,提升分割效果。对比网络中,文献[1]、[13]和[20]都只利用了点云的绝对坐标与相对坐标,未提供充分的几何约束。文献[18]用学习局部相对坐标的方式处理点云的无序性,但冗余信息较多。文献[14]、[19]用图的方式处理点云间的关系,缺少对语义信息的利用。文献[24]令语义、实例分割的特征信息互补。由于语义、实例特征信息的重复、冗余信息较多,虽然在 OA 指标上要高于本文方法 0.96%,但在 mAcc、mIoU 指标上分别低于本文方法 6.43%、4.12%,整体看来效果不如本文方法。

表 2 S3DIS 数据集的语义分割结果

Table 2 Semantic segmentation results on S3DIS dataset

%

方法	mAcc	OA	mIoU
PointNet[1]	66.2	78.6	47.6
PointNet++[13]	67.1	81.0	54.5
PointCNN[18]	75.6	88.1	65.4
SPG[19]	73.0	85.5	62.1
DGCNN[14]	-	84.1	56.1
ShellNet[20]	-	87.1	66.8
PointWeb[21]	76.2	87.3	66.7
A-SCN[22]	-	81.6	52.7
MPNet[23]	-	86.8	61.3
InsSem-SP[24]	74.3	88.5	64.1
Ours+	80.7	87.1	68.2

2)语义分割结果可视化

图 2 为 S3DIS 数据集中 3 个房间的分割结果可视化对比图。其中图 2(a)为人工标注的参考标准,图 2(b)为本文方法的分割结果,图 2(c)为 PointNet 方法的分割结果,图 2(d)为 DGCNN 方法的分割结果,图 2(e)为 SPG 方法的分割结果。每一行是同一个房间。由图示可以看出,相比于对比方法,本文方法在细节处取得了更好的语义分割效果。PointNet 方法对于房间 2 中位置相近的椅子、电脑、墙壁等都没能很好的进行区分。DGCNN、SPG 方法主要是在纹理细节不丰富的墙壁处分割效果不好,例如房间 3、房间 1 所示。SPG 方法在房间 3 中的

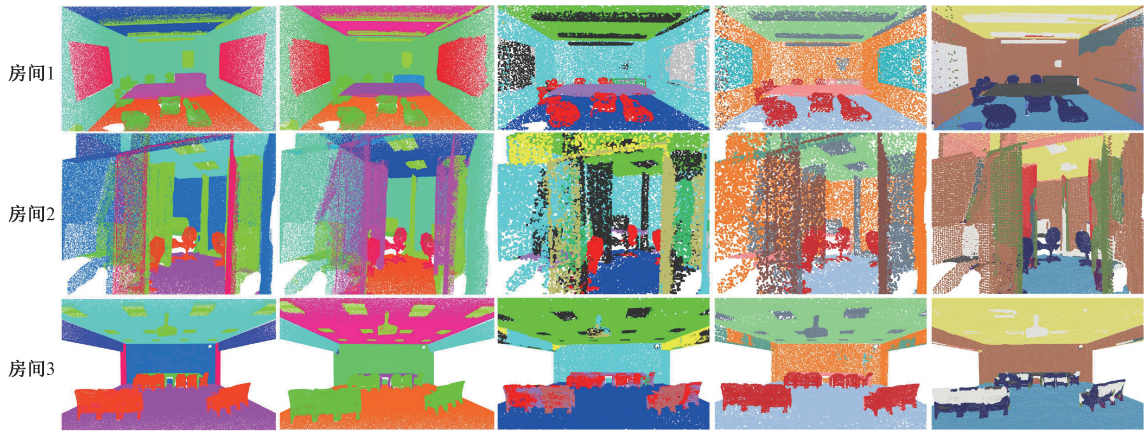

图 2　S3DIS 数据集上语义分割结果可视化

Fig. 2　Visualization of semantic segmentation results on S3DIS dataset

座椅靠背处也出现了分割错误。而本文方法在以上区域都取得了相对更好的分割效果,说明其更好地挖掘了点云局部区域的细粒度信息。

3)网络复杂度分析

表 3 列出了不同方法在 S3DIS 数据集上训练的模型参数量与网络每训练一轮所需的时间。除 PointWeb 方法实验时使用了 4 张 2080Ti 显卡外,其余方法均使用单张 2080Ti 显卡进行测试。PointNet 方法由于其整体网络结构较为简单,所以训练一轮耗时最短。SPG 方法由于其特有的超图构造,参数量最少,但训练一轮所需时间也较长。PointWeb 方法由于其提出的权重适应调整模块中对特征向量进行了大量的元素级相乘,所以计算时间最长,且极其耗费显卡内存,实验过程中使用了 4 张 GPU 显卡进行训练。本文方法由于网络结构较其它对比网络更为复杂,所以参数量较多,虽然时间相比于 PointNet 方法略长,但是精度提高了 20.6%。

表 3　不同网络在 S3DIS 数据集上的复杂度分析

Table 3　Complexity analysis of different networks on S3DIS

方法	参数量/百万	时间/s
PointNet[1]	0.8	293
PointNet++[13]	0.97	392
SPG[19]	**0.25**	1 738
DGCNN[14]	1.97	632
PointWeb[21]	1.22	1 412(4 GPU)
Ours+	1.8	413

4)室外场景测试

为了进一步验证本文的可行性,针对自动驾驶应用方向,在 Semantic3D 数据集上进行了分割测试。该数据集是由超过 40 亿个点组成的大型室外场景三维点云数据集,涵盖了一系列不同的城市场景:教堂、街道、铁路轨道、广场、村庄、足球场、城堡等,在真实世界中的覆盖范围达到 160×240×30 m^3。原始点云被分为 8 个类别:人造地形、自然地形、高植被、低植被、建筑物、硬景观、扫描伪影、汽车,其中每个点包含有三维坐标、RGB 颜色、强度值信息。测试结果可视化如图 3 所示,图 3(a)为输

图 3　Semantic3D 数据集测试结果

Fig. 3　Test results on Semantic3D dataset

入的点云,图 3(b)为语义分割结果。由图可知,本文方法很好的提取出了室外场景地图中的地面、建筑、植物等地图要素,对自动驾驶汽车的定位、路线决策都能提供丰富的辅助信息依据。

3 结 论

针对真实场景点云数据的语义分割,本文提出了一种新的网络结构,构建了双边特征聚合模块,分别处理点云的几何信息、语义信息后进行融合并增强,提取局部上下文信息。利用高维空间中点与点之间的相互作用增强局部邻域的上下文信息。引入注意力池化结合最大值池化,解决其信息损失过大问题,并利用横向跨层双池化结构增加特征信息多样性,避免梯度消失或爆炸。最后利用注意力模块为点云的逐点特征结合全局特征。

实验结果表明,相比于对比网络,本文在多项客观指标取得了更优性能。在应用方面,本文所训练的模型可以嵌入三维扫描仪内,实现工业零部件自动分类;也可应用于自动驾驶领域,用该模型构建环境语义地图,为车辆提供用于理解周围环境的高层次语义信息,实现车辆的高质量视觉定位,提高自主导航模块的环境感知与场景理解能力。

参考文献

[1] QI C R, SU H, MO K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, Hawaii, USA. New York: IEEE, 2017: 652-660.

[2] HU Q, YANG B, XIE L, et al. Randla-net: Efficient semantic segmentation of large-scale point clouds[C]. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, USA. New York: IEEE, 2020: 11108-11117.

[3] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation[C]. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15-21, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 3146-3154.

[4] 单吉超,李秀智,张祥银,等. 室内场景下实时地三维语义地图构建[J]. 仪器仪表学报, 2019, 40(5): 240-248.
SHAN J CH, LI X ZH, ZHANG X Y, et al. Real-time 3D semantic map building in indoor scene [J]. Chinese Journal of Scientific Instrument, 2019, 40(5): 240-248.

[5] 邱佳月,赖际舟,李志敏,等. 面向复杂场景的激光雷达地面分割方法[J]. 仪器仪表学报, 2020, 41(11): 244-251.
QIU J Y, LAI J ZH, LI ZH M, et al. A lidar ground segmentation algorithm for complex scenes[J]. Chinese Journal of Scientific Instrument, 2020, 41(11): 244-251.

[6] 钱昱来,盖绍彦,郑东亮,等. 基于局部和全局信息的快速三维人耳识别[J]. 仪器仪表学报, 2019, 40(11): 99-106.
QIAN Y L, GAI SH Y, ZHENG D L, et al. Fast 3D ear recognition based on local and global information[J]. Chinese Journal of Scientific Instrument, 2019, 40(11): 99-106.

[7] SU H, MAJI S, KALOGERAKIS E, et al. Multi-view convolutional neural networks for 3d shape recognition[C]. Proceedings of 2015 IEEE International Conference on Computer Vision, December 13-16, 2015, Santiago, Chile. New York: IEEE, 2015: 945-953.

[8] CHEN X, MA H, WAN J, et al. Multi-view 3d object detection network for autonomous driving [C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, Hawaii, USA. New York: IEEE, 2017: 1907-1915.

[9] LANG A H, VORA S, CAESAR H, et al. Pointpillars: Fast encoders for object detection from point clouds[C]. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15-21, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 12697-12705.

[10] LE T, DUAN Y. Pointgrid: A deep network for 3d shape understanding [C]. Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition, June 18-21, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 9204-9214.

[11] MENG H Y, GAO L, LAI Y K, et al. Vv-net: Voxel vae net with group convolutions for point cloud segmentation [C]. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision, Oct. 27-Nov. 3, 2019, Seoul, Korea (South). New York: IEEE, 2019: 8500-8508.

[12] CHEN Y, LIU S, SHEN X, et al. Fast point r-cnn[C]. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision, Oct. 27-Nov. 3, 2019, Seoul, Korea (South). New York: IEEE, 2019: 9775-9784.

[13] 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. December 3-9, Long Beach, CA, USA. New York: IEEE, 2017: 5105-5114.

[14] WANG Y, SUN Y, LIU Z, et al. Dynamic graph cnn for learning on point clouds [J]. Acm Transactions On Graphics (TOG), 2019, 38(5): 1-12.
[15] WU W, QI Z, FUXIN L. Pointconv: Deep convolutional networks on 3d point clouds[C]. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15-21, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 9621-9630.
[16] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.
[17] ARMENI I, SENER O, ZAMIR A R, et al. 3d semantic parsing of large-scale indoor spaces[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, June 26-July 1, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 1534-1543.
[18] LI Y, BU R, SUN M, et al. Pointcnn: Convolution on x-transformed points[J]. Advances in Neural Information Processing Systems, 2018, 31: 820-830.
[19] LANDRIEU L, SIMONOVSKY M. Large-scale point cloud semantic segmentation with superpoint graphs[C]. Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition, June 18-21, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 4558-4567.
[20] ZHANG Z, HUA B S, YEUNG S K. Shellnet: Efficient point cloud convolutional neural networks using concentric shells statistics [C]. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision, Oct. 27-Nov. 3, 2019, Seoul, Korea (South), New York: IEEE, 2019: 1607-1616.
[21] ZHAO H, JIANG L, FU C W, et al. Pointweb: Enhancing local neighborhood features for point cloud processing [C]. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15-21, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 5565-5573.
[22] XIE S, LIU S, CHEN Z, et al. Attentional shapecontextnet for point cloud recognition [C]. Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition, June 18-21, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 4606-4615.
[23] HE T, GONG D, TIAN Z, et al. Learning and memorizing representative prototypes for 3d point cloud semantic and instance segmentation [C]. Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVIII 16. Springer International Publishing, 2020: 564-580.
[24] LIU J, YU M, NI B, et al. Self-prediction for joint instance and semantic segmentation of point clouds[C]. European Conference on Computer Vision, August 23-28, 2020, Edinburgh, England, United Kingdom. 2020: 187-204.

作者简介

王溪波,1985 年于辽宁大学获得学士学位,1991 年于沈阳工业大学获得硕士学位,2012 年于东北大学获得博士学位。现为沈阳工业大学教授,主要研究方向为实时系统及嵌入式软件、智能信息系统。
E-mail: wangxb@ sut. edu. cn

Wang Xibo received his B. Sc. degree from Liaoning University in 1985, received his M. Sc. degree from Shenyang University of Technology in 1991, and received his Ph. D. degree from Northeast University in 2012. He is currently a professor at Shenyang University of Technology. His main research interests include real-time system, embedded software, and intelligent information system.

曹士彭,现为沈阳工业大学硕士研究生,主要研究方向为计算机视觉、点云分析、深度学习。
E-mail:caoshipeng@ sia. cn

Cao Shipeng is currently a master student at Shenyang University of Technology. His main research interests include computer vision, point cloud analysis, and deep-learning.

赵怀慈(通信作者),2003 年于中科院沈阳自动化研究所获得博士学位,现为中科院沈阳自动化研究所研究员,主要研究方向为图像处理、复杂系统建模与仿真技术,指挥、控制、通信与信息处理技术。
E-mail: hczhao@ sia. cn

Zhao Huaici (Corresponding author) received his Ph. D. degree from Shenyang Institute of Automation, Chinese Academy of Sciences in 2003. He is currently a researcher at Shenyang Institute of Automation, Chinese Academy of Sciences. His main research interests include image processing, complex system modeling and simulation technology, command, control, communication and information processing technology.

邰炳昌,1986 年于空军工程学院获得学士学位,现为卓越新时代认证有限公司高级工程师、注册审核员,主要研究方向为人工智能,图像分析处理。
E-mail:tbc9981@ 163. com

Tai Bingchang received his B. Sc. degree from the Air Force Engineering College in 1986. He is currently a senior engineer and a registered auditor at Excellence New Era Certification Limited Company. His main research interests include artificial intelligence, image analysis and processing.