[文章编号] 1003－4684(2026)02-0044-06

基于双通道自注意力的点云语义分割

舒 军，王 帅，吴 启

（湖北工业大学电气与电子工程学院，太阳能高效利用及储能运行控制湖北省重点实验室，湖北 武汉 430068）

[摘 要] 针对点云语义分割存在局部特征提取细粒度特征描述能力不足，以及忽略局部与全局特征融合的问题。本文在主干网络中加入双通道自注意力机制，其中空间自注意模块学习邻域点之间的相互关联性，捕捉上下文信息，提高局部特征的分辨能力，通道自注意模块聚合通道结构信息，减少噪声影响。同时设计了多密度特征融合模块，将不同密度的点云作为基于双通道自注意力机制的骨干网络的输入，聚合不同密度的点云特征，融合局部与全局的特征，提高语义分割的性能。在 S3DIS 和 Semantic3D 数据集上的大量实验表明，双通道自注意力机制的有效性。

[关键词] 双通道自注意力；多密度特征融合；点云；语义分割

[中图分类号] TP391 [文献标识码] A

随着三维传感器的快速发展，传感器获取的点云包含了丰富的三维坐标、颜色、强度值、入射方向等属性信息，越来越受到人们的关注。近年来，很多的学者对 3D 点云数据进行了大量研究。然而，点云数据是无序和非结构化的，不同于图像固定在一个规则的像素网格。为了应对这一挑战，针对点云的数据分析出现了许多不同的方法。

在三维点云的研究中，语义分割是自动驾驶和机器人定位领域的一个热点研究课题。采用点云形式的输入的分割算法可以大致分为三类：基于投影的算法、基于体素的算法和基于原始点云的算法。基于投影的点云语义分割首先将点云投影到二维平面上，利用图像分割算法对投影后的图像进行分割[1-2]，得到每个像素点的语义标签。最后，将语义标签映射回点云中，即可得到点云的语义分割结果。但是由于点云的三维结构信息丢失了一部分，因此可能会导致部分信息的损失。基于体素的点云语义分割[3-4]将点云分割成一系列的体素。对每个体素进行聚合，得到体素的特征表示。接着，利用卷积神经网络(CNN)得到每个体素的语义标签。最后，将语义标签映射回点云中，即可得到点云的语义分割结果。但是需要消耗大量的计算资源，尤其是需要处理高密度的点云数据时，计算量会更大。此外，体素的大小和数量也会对分割的精度和计算效率产生影响。与基于投影或基于体素的点云分割方法相比，基于点的点云分割可以更加准确地捕捉点云数据中的细节信息。基于点的方法无需额外的操作，直接处理点云信息，充分利用了点云数据的三维结构信息，同时利用 CNN 在图像分割领域的成功经验，可以有效地处理大规模点云数据。

Qi 等[5]提出了具有里程碑意义的 PointNet，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。然而，该网络未考虑局部特征提取问题。在这之后，一系列的方法被提出[6-8]，不简单的依靠每个点来提取特征，通过利用局部结构特征聚合点云的局部几何信息。然而，上述的方法没有考虑不同局部特征之间的关系，不能获得丰富的上下文信息。为解决上述问题，将在图像处理领域获得巨大进展的卷积神经网络引入[9-10]，通过卷积神经网络获得不同局部结构特征之间的关系。然而，这些方法仍存在不足，导致计算成本成指数上升。

随着研究的深入，在自然语言处理和二维图像处理任务中取得巨大成功的自注意力也被用于三维点云的处理。自注意力可以考虑输入矢量之间的关系，不受输入顺序的影响。换句话说，它具有置换不变性。因此，自注意力可以很容易地处理点云。许多研究人员在设计神经网络架构时，引入和修改了自注意力机制模块[11]。自注意力机制不仅能够学习丰富的上下文，而且所有的操作都是可以并行的，可以代替卷积神经网络中的卷积运算。尽管自注意

[收稿日期] 2023－07－07
[基金项目] 国家青年科学基金项目(61603127)
[第一作者] 舒 军(1973－2024)，男，湖北武汉人，工学博士，湖北工业大学副教授，研究方向为计算机视觉。
[通信作者] 王 帅(1998－)，男，湖北襄阳人，湖北工业大学硕士研究生，研究方向为计算机视觉。

力机制能够关注主要信息,具有很好的通用性,但有些问题仍然需要在未来的研究中予以解决。本文同样也提出了一种改进点云语义分割的方法,设计了双通道自注意力模块,考虑了点与点之间的空间几何关系和特征关联性。通过双通道自注意力机制学习点的通道特征和邻域点特征之间的相关性,并进一步获取关键信息,从而丰富了上下文特征描述。

在图像领域,多分辨率特征提取和融合可以显著增强对象分类和分割的能力。基于这一现象[12,13],本文提出了一种基于多密度特征融合的点云语义分割网络,该网络能够聚合不同密度的特征,提高了语义分割的性能。首先,通过改变采样率,得到不同密度的点云。低密度点云在所提出的网络中适合于提取目标的全局形状特征,而高密度点云适合于从局部特征中提取细节。然后,从不同尺度的点云中提取特征。最后,通过应用特征融合操作从提取的特征中获得新的特征集。

1　本文算法

整体的网络结构如图 1 所示。首先,应用降采样以获得不同密度的点云。然后,利用提出的基于双通道自注意力机制的网络架构提取每个密度的点云特征。最后对提取的特征进行融合和优化。

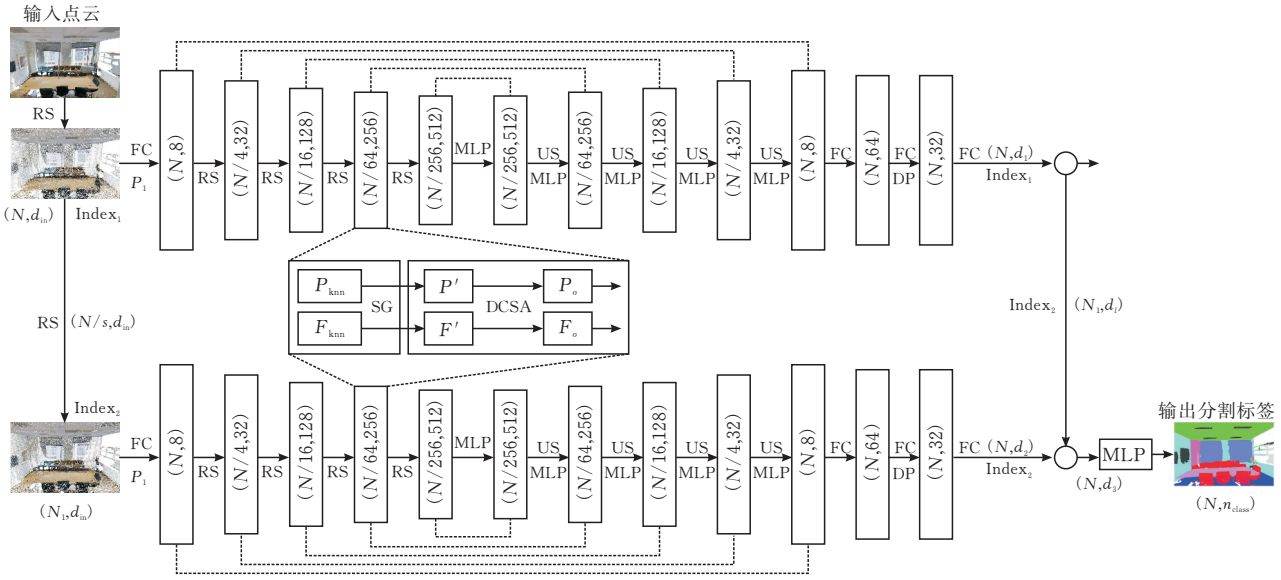

图 1　网络整体架构

1.1　基于双通道自注意力机制的主干网络

基于双通道自注意力机制的主干网络的具体描述为:给定一个点云集合,每个点具有 3 个坐标和 d 维特征(例如,颜色、法线等),分别表示为 P 和 F。目标是合理地使用 P 和 F 从点云集合中提取特征。如图 1 所示,主干网络采用编解码器结构,有利于全特征融合。编码器部分由两个模块组成:SG(抽样和分组)和 DCSA(双通道自注意力)。

SG:首先采用 RS(随机采样)算法对原始点云进行下采样以确定 N 个中心点,并采用基于 KNN(K-Nearest Neighbor)算法寻找中心点,最终得到输入点云集合。在 SG 过程中,将一些几何信息添加到 P 和 F 以获得 P' 和 F',并且作为 DCSA 模块输入。它们可以描述为:

$$P' = \mathrm{cat}(P_{\mathrm{knn}}, P_k - P_c, \| P_k - P_c \|, P_c) \tag{1}$$

$$F' = \mathrm{cat}(F_{\mathrm{knn}}, P_k - P_c, \| P_k - P_c \|, P_c) \tag{2}$$

其中 P_{knn} 是由 KNN 从投影坐标找到的 K 个邻居点的坐标;F_{knn} 是由 KNN 从投影特征找到的 K 个邻居点的特征;P_c 为最近邻采样的中心点;P_k 是 KNN 从原始坐标找到的 K 个邻居点的特征;$P_k - P_c$ 为中心点与邻居点的相对坐标;$\| P_k - P_c \|$ 为中心点与邻居点的欧式距离,在通道维度上将它们连接起来,然后将其送入 DCSA 模块。

DCSA:主要使用自注意机制增强坐标和特征之间的关系。DCSA 模块的具体结构如图 2 所示。它接收两个输入,即 SG 之后的坐标 P' 和特征 F'。对 P' 和 F' 进行线性变换得到 $P_{\mathrm{q}}, P_{\mathrm{k}}, P_{\mathrm{v}}, P_{\mathrm{r}}$ 和 $F_{\mathrm{q}}, F_{\mathrm{k}}, F_{\mathrm{v}}, F_{\mathrm{r}}$。这里 q、k、v 和 r 分别表示查询、键、值和残差。具体定义如下:

$$P_{\mathrm{q}}, P_{\mathrm{k}}, P_{\mathrm{v}}, P_{\mathrm{r}} = L(P') \tag{3}$$

$$F_{\mathrm{q}}, F_{\mathrm{k}}, F_{\mathrm{v}}, F_{\mathrm{r}} = L(F') \tag{4}$$

式中 L 表示线性变换。然后通过 DCSA 运算得到 P_o 和 F_o,该过程定义如下:

$$P_o = P_v \otimes P_a + P' \tag{5}$$

$$F_o = F_v \otimes F_a + F' \tag{6}$$

其中 $\otimes$ 表示矩阵乘法,由上式可以看出,坐标和特征的信息都得到了增强,上式中的 P_a 和 F_a 由 q 和 v 加权得到,具体过程定义如下:

$$P_a = \mathrm{softmax}(sum(P_q{}^T \otimes P_v)) \quad (7)$$

$$F_a = \mathrm{softmax}(sum(F_q{}^T \otimes F_v)) \quad (8)$$

其中 $\otimes$ 同样也是表示矩阵乘法；sum 表示矩阵加法；最后通过 softmax 分配权重。

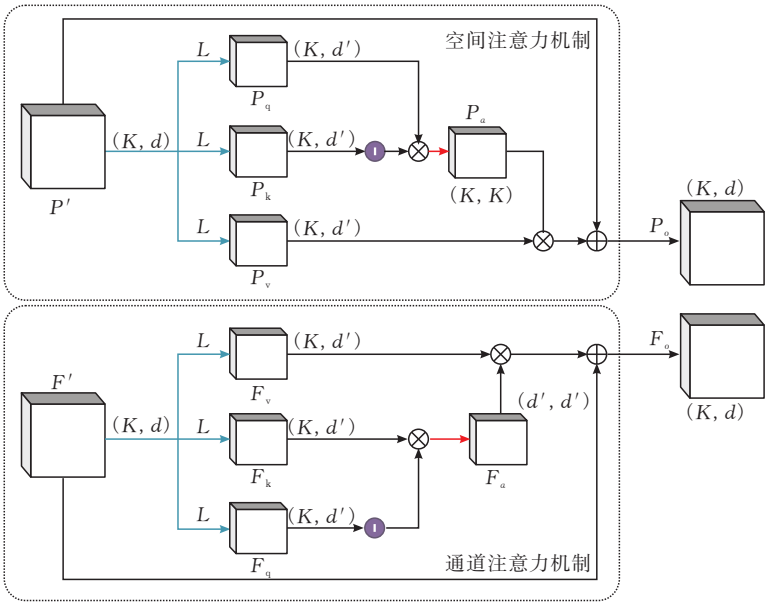

图 2 双通道自注意力机制

对于解码部分，采用 U-Net 结构，应用 4 个解码层，采用最近邻插值的方法进行上采样，同时将上采样之后的特征连接对应编码层编码后产生的特征。最后在经过上述的 4 个解码层之后，获得最后的分割结果。

1.2 多密度特征融合

为了使本文的方法不受尺度变化的影响，提出了一种基于不同密度的多尺度特征提取方法。低密度点云的特征适合于表示全局形状特征，而高密度点云适合于描述详细的局部特征。根据点云的特征，不同密度的特征是互补的。为了构建多尺度特征融合网络，对输入数据进行随机下采样操作。通过两个尺度下采样来构建点云。因此，获得两个不同密度的点云 P_1 和 P_2，分别表示在第一和第二下采样过程之后的点云。记录下两个采样点云中每个采样点的位置，可用于之后的特征融合。

具体的结构如图 1 所示，以一定的采样比例对原始点云执行下采样。所选择的点分别存储在 Index_1 和 Index_2 中用于第一和第二下采样过程。然后，使用基于双通道自注意力机制的主干网络执行特征提取，其中 d_1 和 d_2 是每个下采样点云的特征维度。最后，执行特征融合以获得相关的特征集合，其中“—”指示根据中不存在的描述符，并且“+”指示特征融合。基于所提取的特征，多层感知器(MLP)被用来获得每一个点在每个类别上的得分。具体的结构如图 1 所示。详细的步骤可表示如下。

步骤 1：对输入数据执行下采样。Index_1 表示随机生成且没有重复 N 个点的索引，在范围[0，N]中随机生成 N/S 个索引，其由 Index_2 表示；

步骤 2：使用 Index_1 作为点的索引，生成新的点云 P_1。使用 Index_2 作为索引以生成新的点云 P_2；

步骤 3：由基于双通道自注意力机制的骨干网络提取这些不同密度的点云的特征；

步骤 4：使用 Index_2 获得 P_1 中的点，然后获得这些点的新特征集 (N_1, d_1)，并将 P_2 的特征集 (N_1, d_2) 与新特征集 (N_1, d_1) 组合以获得特征集 (N_1, d_3)；

步骤 5：对 (N_1, d_3) 执行 MLP 操作以分类 P_2 中的每个点，实现点云的语义分割。

2 实验分析

2.1 实验环境和参数设置

在 Ubuntu 系统上基于 TensorFlow 框架进行计算，使用 NVIDIA Quadro P6000 GPU 进行加速。采用 Adam 优化器，并将两个数据集的 Batchsize 分别设置为 6 和 3。初始学习速率设置为 0.01，最大迭代次数为 100。

2.2 评价指标

用整体准确率(OA)，平均准确率(mAcc)和平均交并比(mIoU)这 3 个指标来定量评估语义分割性能。其中 OA 是所有分类正确的点所占的比例，mAcc 是每个标签的准确率的平均值，mIoU 是每个类别预测标签与真实标签的交集与并集之比的平均数[14]，公式如下：

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

$$mAcc=\frac{1}{c}\cdot\sum_{i=1}^{c}\frac{TP}{(TP+FN)} \tag{10}$$

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

式中：TP 是实际为真、预测为真的样本数；FN 是实际为真、预测为假的样本数；FP 是实际为假、预测为真的样本数；c 是语义标签数量。

2.3　S3DIS 实验和结果分析

S3DIS 是大场景室内 3D 点云数据集，该数据集包含 6 个教学和办公 Area，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。通过进行 6 倍交叉验证，评估所提出算法的性能。如表 1 所示，在 6 个区域中，mIoU 为 70.9%，高于大多数现有网络。在天花板、地板、chair、Board、clutter 类别上取得了最佳性能，相对于之前最佳结果，分别提升了 0.8%、0.9%、1.8%、0.8%、0.7%。此外，在 window、door 等类别上，分割精度也优于大多数方法。

选取了 4 个常见的环境，对分割结果进行可视化图像对比分析，如图 3 所示，结果表明，在会议室场景对于柱子和写字板的分割更为准确，对于写字板边缘区分更清晰，分类准确率更高；在餐厅场景中，对于 clutter 和椅子腿的分割更加准确，对物体边缘特征的识别更好，轮廓更清晰，误差更小；在办公室场景中，对于靠在墙面上的杂物以及墙面附近的柱子，都有明显的效果；在走廊场景中，对几何形状的边缘表现出更好的性能，误差更小，相比于 PointNet++，对边缘的区分更为精确。从上述的表现看来，本文提出的算法有效地学习到了点云的空间信息，在物体边界处具有更好的区分性。

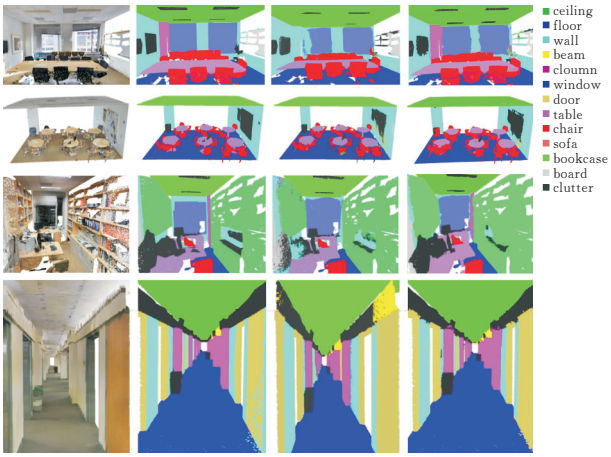

图 3　S3DIS 数据集分割结果

表 1　S3DIS 数据集 6 个区域交叉验证结果

Model	mIoU	OA	mAcc	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	book	Board	clutter
Pointnet	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)	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)	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	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	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	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	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
Ours	**70.9**	**87.5**	**82.0**	**94.2**	**97.0**	80.5	57.8	48.3	64.8	72.4	68.5	**78.2**	68.1	64.1	**66.7**	**60.8**

2.4　Semantic3D 实验和结果分析

采用 Semantic3D 数据集中的 reduce-8 作为数据集，其中训练集和测试集分别包含 15 个完整的点云数据集和 4 个地区的点云数据。实验结果表明，相较于其他算法，本文的模型在 Semantic3D 数据集上的分割精度较高，mIoU 为 77.8%，OA 为 95.1%。在 building 和 hardscape 类别上表现最佳，相对于之前最佳结果分别提升了 0.1% 和 1.1%。此外，在 road 和 natural 等类别上也取得了较好的效果，而在 bush 类别上分割效果相对较差，精度为 51.8%。该结果可能是由于植被数量较少，且 natural、tree 和 bush 等类别的特征相似，难以进行分割，对模型训练和性能产生了一定影响。值得注意的是，hardscape 类别的分布不均匀，其形状和结构变化较大，内部的几何形状、颜色和纹理等特征也随环境变化而变化，在这种情况下，该算法仍然取得了最佳性能。

测试集分割结果的可视化图像如图 4 所示。从整体看来分割效果比较好，建筑道路等物体边界比较清晰。从上述的数据分析和可视化结果中可以看出，算法能够识别点云结构中的细节和复杂部分，有效地区分不同结构的特征和细节。展现出网络优秀

的特征提取、空间聚合和边界分割能力，证明了该算法特征提取模块的有效性。

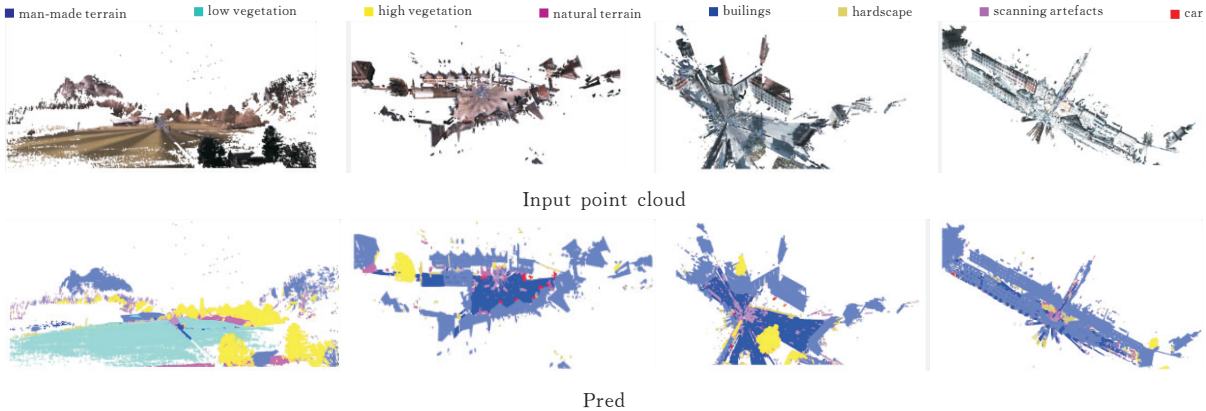

图 4 Semantic3D 数据集语义分割结果

表 2 Semantic3D 数据集语义分割结果

Model	mIoU	OA	rode	natural	tree	bush	buildings	hardscope	artefact	car
SnapNet	59.1	88.6	82.0	77.3	79.7	22.9	91.1	18.4	37.3	64.4
ShellNet	69.3	93.2	96.3	90.4	83.9	41.0	94.2	34.7	43.9	70.2
GACNet	70.8	91.9	86.4	77.7	**88.5**	**60.6**	94.2	37.3	43.5	77.8
SPG	73.2	94.0	**97.4**	**92.6**	87.9	44.0	83.2	31.0	63.5	76.2
RandLA-Net	77.4	94.8	95.6	91.4	86.6	51.5	95.7	51.5	69.8	76.8
KPCnov	74.6	92.9	90.9	82.2	84.2	47.9	94.9	40.0	**77.3**	**79.7**
Ours	77.8	**94.9**	95.4	90.7	87.7	51.8	**95.9**	**52.6**	70.8	77.7

2.5 网络结构分析

为了验证提出的 DCSA 机制和 MDFF 模块的有效性，在相同的网络框架下进行了逐个模块调整的测试，并在 S3DIS 数据集上进行评估。在没有添加 DCSA 和 MDFF 两个模块时，mIoU 为 68.1%。当分别加入 DCSA 和 MDFF 模块后，mIoU 分别提高了 1.1%和 2.0%，分别达到了 69.2%和 70.1%。当两个模块共同作用时，mIoU 提高了 2.8%，达到了 70.9%。通过该消融实验的结果，证明了所提出模块的有效性。

表 3 网络结构分析

DCSA	MDFF	mIoU(S3DIS)
		68.1
√		69.2
	√	70.1
√	√	70.9

2.6 自注意力机制方式选择

在 S3DIS 数据集上进行了自注意力机制的消融实验，通过在构建的网络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)机制以及空间和通道并行作用的 DCSA 机制，评估了这些不同自注意力机制对点云分割性能的影响。如表 4 所示，DCSA机制比 CSA 机制和 SSA 机制分别高1.0%和0.6%，这得益于 DCSA 增强了坐标与特征之间的依赖关系。从而证明了双通道自注意力机制的有效性。

表 4 自注意力机制选择

CSA	SSA	DCSA	mIoU(S3DIS)
√			69.9
	√		70.3
		√	70.9

3 总结

针对现有算法存在的一些问题，提出了双通道自注意力机制和多密度特征融合模块对算法进行优化。实验结果表明，包含空间自注意力机制和通道自注意力机制的双注意力机制能够有效学习点的特征，更大程度的挖掘点的上下文信息。采用的多密度特征融合模块，可以聚合具有不同密度的点云特征，提高语义分割的性能。在 S3DIS 和 Semantic3D 数据集上进行了实验，在这些数据集上分别达到了 70.9%和 77.8%的准确率。从可视化图像中可以看出，对物体的边缘具有很好的区分，在部分类别上有明显的提升。本文提出的方法表现出比 PointNet++、SPG 和 KPCnov 拥有更好的性能。

[参 考 文 献]

[1] WU B, WAN A, YUE X, 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 2018, May 21, 2018-May 25, 2018, 2018: 1887-1893.

[2] ZHANG P, LI G, LIU C, et al. End-to-end BEV Perception Via Homography Matrix[C]. 6th IEEE Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2023, February 24, 2023 - February 26, 2023, 2023: 1352-1356.

[3] MENG H Y, GAO L, LAI Y K, et al. VV-net: Voxel VAE net with group convolutions for point cloud segmentation[C]. 17th IEEE/CVF International Conference on Computer Vision, ICCV 2019, October 27, 2019 - November 2, 2019, 2019: 8499-8507.

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

[5] QI C R, SU H, MO K, et al. PointNet: Deep learning on point sets for 3D classification and segmentation [C]. 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, July 21, 2017 - July 26, 2017, 2017: 77-85.

[6] QI C R, YI L, SU H, et al. PointNet++: Deep hierarchical feature learning on point sets in a metric space [C]. 31st Annual Conference on Neural Information Processing Systems, NIPS 2017, December 4, 2017 - December 9, 2017, 2017: 5100-5109.

[7] HUANG Q, WANG W, NEUMANN U. Recurrent Slice Networks for 3D Segmentation of Point Clouds [C]. 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, June 18, 2018 - June 22, 2018, 2018: 2626-2635.

[8] ZHAO H, JIANG L, FU C W, et al. Pointweb: Enhancing local neighborhood features for point cloud processing[C]. 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, June 16, 2019 - June 20, 2019, 2019: 5560-5568.

[9] HUA B S, TRAN M K, YEUNG S K. Pointwise Convolutional Neural Networks[C]. 31st Meeting of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018, June 18, 2018 - June 22, 2018, 2018: 984-993.

[10] WU W, QI Z, FUXIN L. PointCONV: Deep convolutional networks on 3D point clouds[C]. 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2019, June 16, 2019 - June 20, 2019, 2019: 9613-9622.

[11] GUO M H, CAI J X, LIU Z N, et al. PCT: Point cloud transformer[J]. Computational Visual Media, 2021, 7(02): 187-199.

[12] 王艺娴, 胡雨凡, 孔庆群, 等.三维点云语义分割:现状与挑战[J]. 工程科学学报, 2023, 45(10): 1653-1665.

[13] 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.

[14] 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(08): 12-18.

Point Cloud Semantic Segmentation Based on Dual-channel Self-attention

SHU Jun, WANG Shuai, WU Qi

(*Hubei Key Laboratory for High-efficiency Utilization of Solar Energy and Operation Control of Energy Storage System, Hubei Univ. of Tech., Wuhan* 430068, *China*)

Abstract: Although the deep learning model based on point cloud has made remarkable progress in the past years, there are still some problems, such as the lack of fine-grained feature description ability of local feature extraction and the neglect of local and global feature fusion. Aiming at the above problems, a dual-channel self-attention mechanism is added to the backbone network. The spatial self-attention module learns the correlation between neighborhood points, captures context information, and improves the resolution of local features. The channel self-attention module aggregates channel structure information to reduce noise impact. At the same time, a multi-density feature fusion module is designed. The point clouds with different densities are used as the input of the backbone network based on the dual-channel self-attention mechanism. The point cloud features with different densities are aggregated, and the local and global features are fused to improve the performance of semantic segmentation. A large number of experiments on S3DIS and Semantic3D datasets show that this method can obtain competitive results.

Keywords: Dual-channel self-attention; multi-density feature fusion; point cloud; semantic segmentation

[责任编校: 赵　瞳]