工程科学学报
Chinese Journal of Engineering

三维点云语义分割：现状与挑战

王艺娴 胡雨凡 孔庆群 曾慧 张利欣 樊彬

3D point cloud semantic segmentation: state of the art and challenges

WANG Yixian, HU Yufan, KONG Qingqun, ZENG Hui, ZHANG Lixin, FAN Bin

引用本文:
王艺娴, 胡雨凡, 孔庆群, 曾慧, 张利欣, 樊彬. 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. doi: 10.13374/j.issn2095–9389.2022.12.17.004
WANG Yixian, HU Yufan, KONG Qingqun, ZENG Hui, ZHANG Lixin, FAN Bin. 3D point cloud semantic segmentation: state of the art and challenges[J]. *Chinese Journal of Engineering*, 2023, 45(10): 1653–1665. doi: 10.13374/j.issn2095–9389.2022.12.17.004

在线阅读 View online: https://doi.org/10.13374/j.issn2095–9389.2022.12.17.004

工程科学学报, 第 45 卷, 第 10 期: 1653−1665, 2023 年 10 月
Chinese Journal of Engineering, Vol. 45, No. 10: 1653−1665, October 2023
https://doi.org/10.13374/j.issn2095-9389.2022.12.17.004; http://cje.ustb.edu.cn

三维点云语义分割: 现状与挑战

王艺娴[1)], 胡雨凡[1)], 孔庆群[2,3)], 曾　慧[1)], 张利欣[1)], 樊　彬[1)✉]

1) 北京科技大学智能科学与技术学院, 北京 100083　2) 中国科学院自动化研究所, 北京 100190　3) 中国科学院大学, 北京 100049
✉通信作者, E-mail: bin.fan@ieee.org

摘　要　随着获取点云数据成本下降以及 GPU 算力的提高, 众多三维视觉场景如自动驾驶、工业控制、MR/XR 对三维语义分割的需求日益旺盛, 这进一步推动了深度学习模型在三维点云语义分割任务中的发展. 近期, 深度学习模型在网络架构上持续创新, 如 RandLA-Net 和 Point Transformer, 并突破性地以更低的计算成本提高了分割准确率, 但已有的三维点云语义分割综述介绍的研究工作包含大量早期以及被舍弃的方法, 没有系统地整理这些新型高效的方法, 不能很好地体现研究现状. 此外, 这部分综述以输入网络的不同数据类型分类各点云语义分割方法, 不能有效地体现各方法的演进关系, 也不利于对比不同方法的分割性能. 针对以上问题, 本文面向近 3 年的研究成果和最新的研究进展, 重点归纳了三维点云语义分割中基于不同网络架构的方法、面临的挑战及潜在研究方向, 并从 3 个层面对三维点云语义分割进行了系统地综述. 通过本文, 读者可以较系统地了解三维点云语义分割的数据获取方式、常见数据集及模型的评价指标, 对比基于不同网络架构的三维点云语义分割方法的发展过程、分割性能和优缺点, 并进一步认识三维点云语义分割现存的挑战和潜在的研究方向.

关键词　三维视觉; 点云; 语义分割; 深度学习; 网络架构
分类号　TP391;TP183

3D point cloud semantic segmentation: state of the art and challenges

WANG Yixian[1)], *HU Yufan*[1)], *KONG Qingqun*[2,3)], *ZENG Hui*[1)], *ZHANG Lixin*[1)], *FAN Bin*[1)✉]

1) School of Intelligence Science and Technology, University of Science and Technology Beijing, Beijing 100083, China
2) Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
3) University of Chinese Academy of Sciences, Beijing 100049, China
✉Corresponding author, E-mail: bin.fan@ieee.org

ABSTRACT　Decrease in the cost of acquiring 3D point cloud data coupled with the rapid advancements in GPU computing power have resulted in an increased demand for 3D point cloud semantic segmentation in numerous 3D visual applications, including but not limited to autonomous driving, industrial control, and MR/XR, which further advances the development of deep learning methods in 3D point cloud semantic segmentation. Recently, many novel deep learning network architectures, such as RandLA-Net and Point Transformer, have been proposed and have achieved notable improvements in semantic segmentation accuracy while decreasing the computational load. However, previous research on 3D point cloud semantic segmentation methods has focused primarily on relatively early works, whose approaches have been gradually abandoned over the years and cannot accurately reflect the current research status. Moreover, the existing methods have been categorized based on their input data types, making it difficult to compare the segmentation performance of different techniques and not providing a comprehensive view of the relationship between methods using different network architectures. Therefore, this paper reviews the mainstream 3D semantic segmentation methods developed in the last three years using different deep learning network architectures and is organized into three levels. First, the two principal 3D point cloud data acquisition methods, including their customary datasets and metrics to evaluate model performance, are introduced. Second, a systematic

收稿日期: 2022−12−17
基金项目: 北京市自然科学基金资助项目(4202073); 国家自然科学基金资助项目(62076026, 61973029)

review of 3D semantic segmentation methods based on different network architectures is organized, followed by a statistical analysis of the evaluation of performance between different models on two 3D segmentation datasets—S3DIS and ScanNet. The analysis of model performance on these two commonly used datasets includes model structure relevance, strengths, and limitations. Finally, an insightful discussion of the remaining methodological and application challenges and potential research directions is provided. This paper offers an extensive overview of the recent three-year research progress in 3D point cloud semantic segmentation and summarizes various network architecture pipelines, elucidates their fundamental operations, compares the model performance across multiple architectures, discusses their notable strengths and limitations, most importantly, concludes the current challenges and promising research directions for future investigations. Furthermore, this paper enables researchers to effortlessly identify the relevant research and research hotspots among different 3D point cloud semantic segmentation methods based on the analyses presented and aims to update the reviews on 3D point cloud semantic segmentation methods with a better viewpoint and highlight key properties and contributions of proposed methods, providing promising research directions for the main challenges.

KEY WORDS 3D vision; point cloud; semantic segmentation; deep learning; network framework

三维点云语义分割是计算机视觉中一个基本问题, 其主要任务是针对给定的描述三维场景的数据, 如三维点云、颜色-深度 (RGB-D) 图, 通过三维点云语义分割算法, 输出三维场景中每个点的语义标签值. 三维点云语义分割是自动驾驶导航规划、工业自动控制抓取等高级人工智能任务的基础任务, 也是目前三维计算机视觉、深度学习中的研究热点.

早期, 由于直接获取大量三维点云的成本较高, 点云一般需要由图像转化得到, 因此许多传统点云语义分割方法, 如条件随机场, 是让模型先学习图像特征, 再将图像特征转化为深度信息, 如点的三维坐标和语义标签, 并将这些包含深度信息的二维像素投影为带有语义标签的三维点, 实现三维点云语义分割[1], 我们称这些方法为传统方法. 随着大场景点云数据集[2] 的出现及 GPU 算力的提升, 一些深度学习方法, 如 PointConv[3]、DGCNN[4]、Point Transformer[5], 已逐渐代替传统方法成为主流. 这些方法主要利用深度神经网络[6] 学习更丰富的点云特征, 从而得到更准确的语义信息来预测标签. 但在应用这些方法的过程中也出现了如模型缺乏训练数据、模型的复杂度较高导致推理速度较慢、模型的占用计算机内存过高等挑战.

现有综述如 Guo 等[7], Xie 等[8] 都是面向所有三维视觉任务, 并非针对三维点云语义分割, 且其中介绍的点云语义分割工作包含大量早期以及被舍弃的方法, 不能体现点云语义分割领域现阶段的主要关注点. 针对三维分割的综述[9] 则根据模型输入的数据类型来分类三维点云语义分割方法, 不能很好地对比不同方法的分割准确率. 此外, 针对 Transformer 在三维视觉应用方面的综述[10−12], 没有对比其他深度学习网络在三维点云语义分割上的应用, 而这些方法在现阶段仍然属于该领域的研究热点. 因此, 本文从以下 3 个层面对三维点云语义分割进行综述: (1) 如何获取点云数据并评价不同点云语义分割方法; (2) 不同点云语义分割网络架构的出现原因及性能对比; (3) 现有点云语义分割方法在实际应用中存在的挑战, 以及潜在的研究方向. 其中, 本文在第 2 部分详细介绍了不同类型点云语义分割网络要解决的问题并列举了代表性工作, 结合图示阐明了不同网络的基本计算过程及演化关系, 同时在两个常用的用于评估模型语义分割性能的点云数据集上做了详细的性能对比, 分析了每类网络的优缺点; 在第 3 部分针对不同网络架构的缺点, 进一步总结了点云语义分割面临的 3 个挑战及潜在研究方向.

1 三维点云语义分割的常用数据集与评价指标

三维点云是三维点云语义分割问题的数据样本, 是对一个三维空间中所有物体进行曲面采样而得到的一个点集, 用于描述特定的三维场景. 若用一个矩阵 $\boldsymbol{P}$ 表示一个三维点云, 用一个 n 维特征空间中的向量 $\boldsymbol{p}=[p_1,p_2,\cdots,p_n]^{\mathrm{T}}\in\mathbf{R}^n$ 表示点云中的一个点, 则一个由 m 个点组成的三维点云可表示为 $\boldsymbol{P}=\{\boldsymbol{p}_i=[p_{i1},p_{i2},\cdots,p_{in}]^{\mathrm{T}}|i=1,\cdots,m\}$, 其中 p_{ij} 可为点的三维坐标或 RGB 等特征值. 不同于二维图像中像素的有序紧密排列, 三维点云中的点是无序稀疏分布在三维空间中的, 如图 1 为图像像素与三维点的对比. 三维点云中的每个点根据不同的数据获取方式, 对应激光雷达扫描空间的一个测量点, 或对应 RGB-D 图的一个像素. 由于点的坐标等信息与点间的排列顺序无关, 因此要求点云语义分割方法具有置换不变性. 同时点的语义

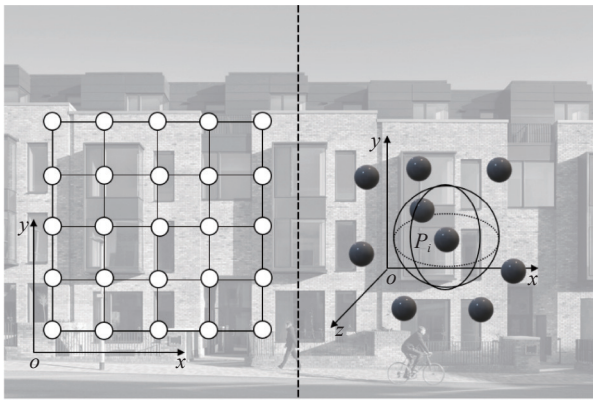

图 1 二维图像中的像素 (左) 与三维点云中的点 (右)

Fig.1 Pixels in a 2D image (left) and points in a 3D point cloud (right)

也不受三维坐标旋转、平移的影响，因此还要求方法具有刚体不变性.

1.1 点云数据获取方式及常用数据集

获取点云数据的方法有二维图像投影、激光雷达扫描等. 早期，由于使用激光雷达的成本较高，点云数据基本由图像投影得到. 随着雷达技术的发展，使用雷达扫描空间直接获取点云数据是目前常见的方法. 随着深度学习技术的不断发展，基于深度神经网络的模型对大场景点云数据的需求日益急迫. 因此，有研究人员陆续发布了公开的点云数据集，方便模型的性能评估和不同模型的对比. 按照点云数据的获取方式，常见的点云数据集可分为激光雷达点云数据集和 RGB-D 数据集.

1.1.1 激光雷达点云数据集

激光雷达点云数据是通过雷达发射的激光能量来测量传感器和被测物体之间的距离等信息而直接生成的点云数据集. 该类数据集主要由不同的激光扫描仪捕获帧或帧序列得到，数据的分辨率较高、连续范围广、噪音较小[13]. 以下列举 3 个在点云语义分割研究中使用较多的雷达点云数据集，其中 S3DIS[2] 和 Semantic3D[14] 数据集中所有物体均为静态，SemanticKITTI[15] 数据集中有静态和动态物体.

(1) S3DIS[2].

S3DIS 数据集由美国斯坦大学、普林斯顿大学和芝加哥丰田技术大学的研究人员共同开发，并于 2016 年公开. 它是一个室内雷达点云数据集，由固定的地面扫描仪扫描了总面积超过 6000 m^2，6 个大型建筑内的 272 个房间的室内场景得到. 完整的 2D-3D-S3DIS 数据集包含超过 70000 个 RGB 图像，以及每个 RGB 图像对应的表面法线、语义注释、相机信息等. 一般的 S3DIS 数据集指仅包含三维点云的空间坐标、RGB 和语义标签，其中点被划分为 13 类语义类别.

(2) Semantic3D[14].

该数据集由瑞士苏黎世联邦理工学院的研究人员开发，并于 2017 年公开，是一个室外雷达点云数据集. 它由固定的地面激光扫描仪在总面积超 115200 m^2 的中欧区域内扫描得到，其中 15 个场景作为训练集，15 个场景作为测试集. 数据集中的物体均为静态，包含三维点云的空间坐标、强度、颜色和标签信息，其中点被分为 8 类语义类别标签.

(3) SemanticKITTI[15].

该数据集由德国波恩大学的研究人员开发，并于 2019 年公开，是一个室外雷达点云数据集. 它基于 2012 年公开的 KITTI[16] 数据集，注释了 KITTI 所有 22 个序列中超过 43000 个德国城市的街区场景，其中序列 00 到 10 作为训练集，序列 11 到 21 作为测试集；点被分为 28 个类别，有 6 个类别附加了移动或不移动的标签，1 个类别作为由于错误测量而无法识别的类. 训练集包含点云的空间坐标和标签信息，测试集实施在线测评.

理想的点云数据集中，不同类别样本差异大、数量分布均匀且样本类别无限多，但实际采集的数据集往往存在不同类别数据数量不平衡、数据总体类别少等问题. 目前针对此类问题，研究者尽可能使用数据量大、数据种类丰富的数据集来训练网络. 而针对三维点云语义分割模型应用的不同场景，室内场景主要使用 S3DIS 训练模型，室外场景主要使用 Semantic3D 和 SemanticKITTI.

1.1.2 RGB-D 数据集

RGB-D 数据集是通过 RGB-D 相机拍摄的，具有像素级颜色和深度信息的图像，计算每个像素的三维空间后，间接生成的点云数据集. 由于 RGB-D 数据不如雷达点云数据准确，雷达点云比 RGB-D 数据更常用于三维点云语义分割，因此本文只介绍一个常用的 RGB-D 数据集——ScanNet[17]. 该数据集由斯坦福大学、普林斯顿大小和慕尼黑工业大学的研究人员共同开发，并于 2018 年公开. 它是一个实例级的室内 RGB-D 视频数据集，收集了 1513 个场景，其中 1201 个场景用于训练，312 个场景用于测试. 数据集含 1513 张像素级语义标注的 RGB-D 图像，以及由图像处理得到的三维数据 (点云的空间坐标、颜色和体素级类别标签)，其中体素被分为 21 个类别. ScanNetv2 是 ScanNet 的最新版本.

表 1 汇总了常用的点云语义分割数据集信息, 包含数据类型、传感器信息、场景信息等.

1.2　点云语义分割的评估指标

总体准确率 (OAcc), 平均准确率 (mOAcc) 和平均交并比 (mIoU) 是评估三维点云语义分割精确度的常见指标. 点云中的一个点作为一个训练样本, 假设所有点分为 M 个语义类别, i 表示第 i 个语义类别, i 的值域为 $\{0, \cdots, M\}$. $\boldsymbol{c}$ 是一个 $M \times M$ 的混淆矩阵, c_{ij} 的第一个下标表示样本的真实标签类别, 第二个下标表示样本的预测标签类别, 因此 c_{ij} 表示真实标签为第 i 类, 而预测标签为第 j 类的样本数量. 每个类别的交并比称为 IoU_i, 各指标具体的计算方法如下式所示:

$$\mathrm{OAcc} = \sum_{i=0}^{M} \frac{c_{ii}}{\sum_{j=0}^{M} c_{ij}} \tag{1}$$

$$\mathrm{mAcc} = \frac{1}{M+1}\mathrm{OAcc} = \frac{1}{M+1}\sum_{i=0}^{M} \frac{c_{ii}}{\sum_{j=0}^{M} c_{ij}} \tag{2}$$

$$\mathrm{mIoU} = \frac{1}{M+1}\sum_{i=0}^{M} \mathrm{IoU}_i = \frac{1}{M+1}\sum_{i=0}^{M} \frac{c_{ii}}{\sum_{j=0}^{M}(c_{ij}) + \sum_{j=0}^{M}(c_{ji}) - c_{ii}} \tag{3}$$

2　方法现状

由于目前已有较多相关的综述介绍了早期点云语义分割方法, 如 Guo 等[7]、Xie 等[8] 的工作, 因此本文主要介绍 2019 年至今较新的研究工作. 本文根据深度学习方法使用的网络架构, 将三维点云语义分割方法分为基于卷积神经网络 (Convolutional neural networks, CNN) 、基于图神经网络 (Graph neural networks, GNN)、基于注意力 (Attention) 网络、基于 Transformer 和基于其他网络的 5 类方法. 图 2 为近 3 年按工作发表时间排序的点云语义分割领域的主要研究成果.

2.1　基于 CNN 的方法

早期, 由于获取激光雷达点云数据的成本较高, 基于 CNN 的三维点云语义分割方法主要是让网络先完成对图像数据的分割, 再将图像的分割结果投影为带有语义标签的点云数据[57]. 但这样的方法容易引入噪声且内存开销大. 近年来, 由于使用激光雷达的成本不断降低, 出现了如 PointNet[58] 等直接将点云作为网络输入的研究成果. 目前, 基于 CNN 的方法主要可分为以下 2 类: (1) 将点云转化为图像或体素后作为二维卷积或三维卷积的输入; (2) 重新定义一种点卷积, 直接将点云作为点卷积的输入.

在将点云转化为体素输入三维卷积的方法中, Choy 等[20] 为了解决三维卷积性能差的问题的, 提

表 1　点云语义分割常用数据集

Table 1　Popular point cloud semantic segmentation datasets

Dataset name	Dataset type	Sensors	Scene type	# scenes	# classes	Year
S3DIS[2]	LiDAR point clouds	Matterport camera	indoor	272	13	2016
Semantic3D[14]	LiDAR point clouds	Terrestrial laser scanners	outdoor	30	8	2017
SemanticKITTI[15]	LiDAR point clouds	Mobile laser scanners	outdoor	43552	28	2019
ScanNet[17]	RGB-D images	RGB-D camera	indoor	1513	21	2018

Note: "#" represents "the number of".

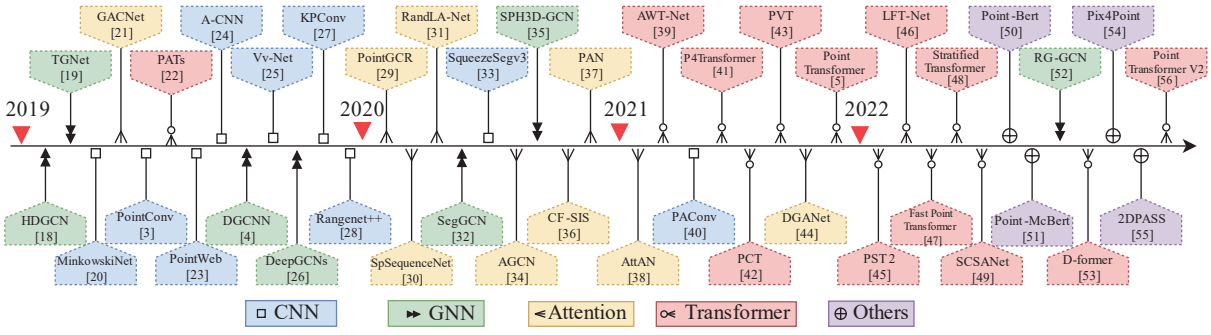

图 2　三维点云语义分割方法发展里程

Fig.2　Significant milestones in 3D point cloud semantic segmentation methods

出一种作用在稀疏张量上的广义三维稀疏卷积. 其比二维卷积和二维−三维混合卷积更能提高模型的鲁棒性，同时不会过多增加模型的计算和内存开销. Meng 等[25]设计了一种 Vv-net 将点云体素化. 网络首先用一种基于核函数的插值变分自动编码器来编码每个体素内的局部几何形状，并进一步用径向基函数来计算每个体素内的局部连续表示，最后用三维群等变 CNN 学习体素特征.

由于将点云转化为图像或体素不能有效利用点云的空间特征，且会不同程度地增加数据处理成本和结构化噪声，因此有不少研究人员在如何设计高效的点卷积上开展工作. PointConv[3]中引入了一种重加权采样密度的点卷积和一种提高模型内存效率的顺序求和技术. PointWeb[23]在具有子流形稀疏卷积和稀疏卷积的 U 形网络上，应用基于原始坐标和移位坐标的点聚类方法进行语义分割. Komarichev 等[24]提出一种直接作用在点云上的环形卷积神经网络 (Annularly convolutional neural network, A-CNN)，其通过改变扩张型 K-最近邻图查询中的环形结构和方向来对近邻点排序，再对这些有序的近邻点应用标准点卷积. Thomas 等[27]通过定义一种新的核点卷积 (KPConv) 来学习点云的局部特征，并提出一种学习核心点的局部偏移的可变形卷积，使模型可在点云的不同位置进行不同的位移. Xu 等[40]用一个动态的卷积权重矩阵来构造卷积核，并提出一种位置自适应卷积 (Position adaptive convolution, PAConv)，其中权重矩阵的系数可由分数网络自适应地学习点的相对位置关系得到. SqueezeSegv2[59]用上下文聚合模块改进 SqueezeSeg[60]，以提高其对脱落噪声的鲁棒性. SqueezeSegv3[33]在 SqueezeSegv2 的基础上增加了空间自适应卷积，从而针对不同空间位置采用不同的卷积核进行卷积.

2.2 基于 GNN 的方法

针对点云语义分割中使用二维卷积或三维卷积时需要先将点云转化为图像或体素的问题，除了设计以点云作为输入的点卷积，还可以使用 GNN 建立关于点云的特殊图结构[4,18−19,26,32,35,52,61]，再使用图卷积来探索每个点的邻居信息，从而更好地利用点云的空间特征，提高分割精度.

Liang 等[18]提出采用多层动态图卷积 (Dynamic graph convolution, DGConv) 构造的分层动态图卷积网络 (Hierarchical depth-wise graph convolutional neural network, HDGCN). DGConv 将深度图卷积和点卷积结合，深度图卷积用来降低内存消耗，同时学习跨通道的特征，点卷积用来学习每个通道的独立特征. Wang 等[4]设计的动态图卷积网络 (Dynamic graph convolutional neural network, DGCNN) 以输入的 N 个点为中心，逐层计算出各自的 K 近邻点以动态构建局部邻域图，然后用边缘卷积计算中心点与近邻点间的边缘特征. 但边缘特征的固定尺寸使模型在不同尺度和输入点数时不能有较好的性能. 深度图卷积网络 (DeepGCNs)[26]是将残差连接、密集连接和扩张卷积应用到图卷积网络 (GCN) 中，构建了比以往 GCN 都深的 56 层网络，解决了 GCN 中叠加网络的梯度消失问题.

球形核函数[62]具有平移不变性 (不改变经过平移变换的输入结果) 和非对称性 (不同顺序顶点对的输出不同) 的特点. Lei 等[35]用具有球形核函数的可分离图卷积网络 (Spherical kernel with graph neural networks for point cloud, SPH3D-GCN)，学习局部点云在空间的几何关系，并设计了适用于 SPH3D-GCN 的池化和非池化操作，使模型更适用于大规模点云的语义分割. Lei 等[32]提出一种基于模糊机制的球形核函数，并将其应用到深度可分离的图卷积网络中形成分割图卷积网络 (SegGCN)，解决了点云空间边界不连续导致的分割不准问题. TGNet[19]是在不同尺度的邻域中，用一系列泰勒加权的高斯核函数来学习由粗到细的局部图语义特征，使模型对变尺度的输入具有鲁棒性.

2.3 基于注意力的方法

CNN 对不同特征进行各向同性的卷积操作在一定程度上限制了语义分割的准确性，而基于注意力的网络可以选择关注与中心点最相关的点，学习其与语义信息最相关的特征，从而降低计算成本，并快速获取最有效的信息.

Wang 等[21]设计了一个图注意力卷积网络 (Graph attention convolutional network, GACNet)，它先给中心点的不同邻居点分配合适的注意力权重，在学习特征的同时学习注意力权重分布，并根据学习到的分布确定卷积核的形状，从而使模型学习到最相关的邻居特征，避免对象之间的特征污染. 注意力图卷积网络 (AGCN)[34]作为一种基于注意力的 GCN，通过在 GCN 中叠加多层的点注意力层，来学习局部结构之间的关系，并附加一个全局点图，辅助点注意力层学习单点间的相关性，从而更好地聚合局部信息. Liang 等[63]在图卷积网络上定义了基于注意力的 K 近邻点，通过为每个近邻点分配不同的权重，自动选择并聚合最重要的近邻点特征.

Ma 等[29] 提出一种点全局上下文推理方法 (Point-GCR), 模型使用通道注意力来学习一个通道图, 其中图节点为输入点在某个通道上的特征, 图边为任意两个通道特征间的相关性, 同时学习点云的空间相关性和不同通道特征的相关性. Shi 等[30] 为了解决四维点云 (三维点云视频帧) 语义分割中时间和空间信息丢失的问题, 在三维稀疏卷积中添加跨帧全局注意模块和跨帧局部插值模块, 设计出 SpSequenceNet. Zhang 等[38] 提出注意力对抗学习网络, 从而让网络更加关注不同的邻域信息. DGANet[44] 通过集成由一种偏移注意力机制实现的扩张图注意力模块, 进一步差异化构建的局部图的每条边, 从而更好地学习边缘特征.

为了设计一种高效轻量的网络, Hu 等[31] 提出 RandLA-Net, 其使用类似 PointNet++[64] 的分层结构, 在每层的特征提取中使用具有注意力池化的局部特征聚合模块来学习复杂的局部特征. PAN[37] 基于一种新型局部注意力边缘卷积层和逐点空间注意力模块. 其中, 局部注意力边缘卷积层用来在沿多方向搜索的邻域中构建相邻点的局部图, 逐点空间注意力用来生成所有点的互相关矩阵, 学习更精确的长距离空间的上下文特征.

2.4　基于 Transformer 的方法

尽管注意力机制可以让模型筛选学习最重要的信息[65], 但研究人员往往需要耗费很大精力为不同的任务设计特别的注意力模块, 如通道注意力、空间注意力等[66], 且不同的注意力模块间的计算复杂度不同, 不支持并行计算. 为了提高注意力的鲁棒性和计算效率, Transformer 使用多头自注意力 (Multi-head self-attention, MSA) 来建立模型的关注点, 其对输入的无序序列具有天然的输出不变性, 当输入点云发生置换和刚体变换时, Transformer 具有较稳定的输出. 同时, Transformer 兼有并行计算和不同输入单元间最大路径短的特点, 因此能让模型一次性为每个点建立起其与剩余所有点之间的最相关关系.

相较于使用缩放点积注意力的模型 (图 3(a)), 基于 Transformer 的模型 (图 3(b)) 融合了多头自注意力、残差连接 (Add) 和正则化 (Norm)、前馈网络 (Feed-forward network, FFN) 等模块, 其中自注意力用于建立相关关系, 多头用于进一步聚合在不同特征空间建立的点的相关关系, 残差连接用于补充可能丢失的点的空间信息, 正则化用于稳定模型的输入和输出, 前馈网络用于转化点间的相关关系为每个点的特征. 根据网络中是否混合使用除 Transformer 之外的其他网络, 基于 Transformer 的方法可进一步分为纯 Transformer 网络和混合 Transformer 网络[10].

2.4.1　纯 Transformer 网络

纯 Transformer 采用编码器-解码器结构, 不使用任何 CNN 或 GNN, 其中编码器由多个 Transformer 层叠加, 每个 Transformer 层的结构如图 3(b) 所示. 解码器也由多个 Transformer 层叠加, 其输入为编码器最后一层的输出.

PATs[22] 用低参高效的组混杂注意力取代高参低效的 MSA. PCT[42] 在编码器加入邻域点的向量映射结构和 4 个连续的偏移注意力结构, 增强模型学习局部上下文特征的能力. PVT[43] 通过 Transfor-

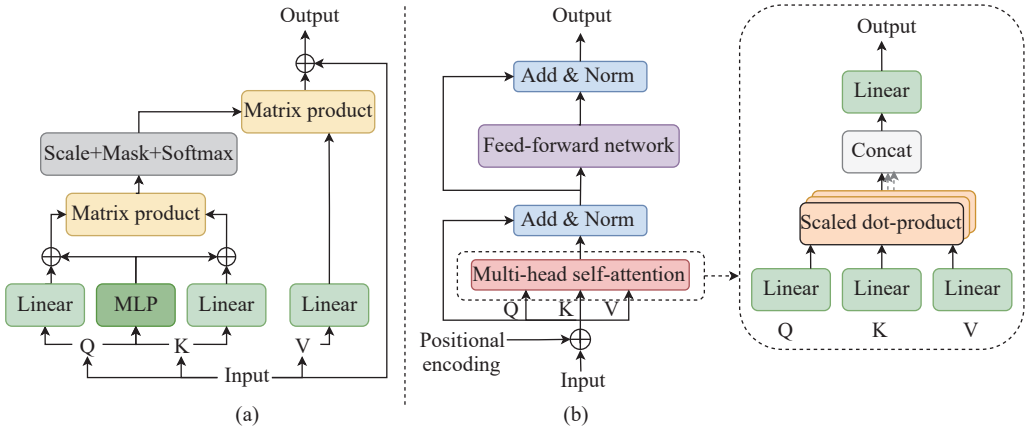

图 3　缩放点积注意力和单层 Transformer 的结构 (其中注意力模块的输入分为查询 Q, 键 K 和值 V, 并得到带权重的输出. 最右侧虚线框内为多头自注意力的结构). (a) 缩放点积注意力; (b) 单层 Transformer

Fig.3　Structures of scaled dot-product attention and single-layer Transformer encoder (the attention module has three inputs: a query vector Q, key vector K, value vector V, and weighted output. The structure of multi-head self-attention is shown inside the right-most dashed box): (a) scaled dot-product attention; (b) single-layer Transformer encoder

mer 分别增强作用于点和作用于体素的特征提取，并提出一种线性复杂度的局部注意力算法，同时对相对位置编码来计算相对注意力. Point Transformer[5] 在每个点的邻域中引入自注意力机制，并利用多层感知机 (Multi-layer perceptron, MLP) 对中心点与近邻点的相对位置编码，学习自注意力特征. Point Transformer 类似 PointNet++[64]，在编码器中通过最远点采样下采样出最远点，减少深层网络需要计算的点云数量，同时用最远点最大程度地保留点云的特征信息；在解码器中通过线性插值将点云数量上采样到输入规模，从而预测逐点的类别标签. Point Transformer V2[56] 在 Point Transformer 的基础上提出具有权重编码层的分组向量注意力，让注意力之间可交互信息. 同时在关系向量上增加位置编码，提高了模型的推理速度.

为了降低计算成本，LFT-Net[46] 仅在编码阶段使用 Transformer 来学习点云高纬度的局部特征，并用自注意力加权的转换池化模块取代一般池化，避免过多局部特征的丢失. Fast Point Transformer[47] 提出一种低空间计算复杂度的局部自注意力模块，同时基于体素散列构建模型，使网络的推理速度比 Point Transformer 快 129 倍.

2.4.2 混合 Transformer 网络

由于纯 Transformer 在训练时需生成关于所有输入点的注意力图，其计算复杂度高. 为降低 Transformer 的使用成本，同时利用 CNN 等网络的优势，混合 Transformer[48,49,53,56] 结合了这些网络.

Stratified Transformer[48] 仅在编码器中使用 Transformer，并将编码器的第一层作为点的向量映射层，使用高效的 KPConv[27] 学习点云的局部特征. 其在 Transformer 中使用一种具有更好上下文学习能力的分层键采样策略，学习点云的多尺度特征. Segment-Fusion [67] 是一种新的片段特征融合方法，先通过图割算法将点分组成片段，并将点特征融合成片段特征；然后利用基于 Transformer 的片段融合网络，融合不同片段的上下文信息，同时将注意力矩阵与邻接矩阵相乘，限制不同分段之间的信息交互. D-former[53] 以扩张的方式在局部和全局范围内交替进行自我注意，在不增加所涉及的补丁的情况下扩大感受野，从而降低计算成本.

图 4 总结了上述 4 种点云语义分割方法的异同，其中上半部分展示了应用于三维点云语义分割模型的常见编码器–解码器结构，图 4 (a) ~ (d) 为不同架构网络在编码器中每一层的基本操作. 其中卷积网络需先从邻域搜索近邻点，再通过对近邻点卷积操作提取中心点特征；图卷积网络则在卷积之前构建点之间的图关系，再由卷积去学习中心点和近邻点间的相关性；注意力图网络是在图关系的基础上通过注意力机制如缩放点积注意力，提取到中心点与附近不同的近邻点更细微的相关信息；Transformer 用每个点的位置编码取代邻域搜索，在全域使用多头自注意力，再用前馈网络进行特征融合，建立所有点之间的复杂关系，进而学习到更丰富的语义信息.

2.5 基于其他网络的方法

2.5.1 无/自监督网络

无/自监督网络一般先在大规模未标注的预训练数据集上预训练，学习三维场景表征等代理任务，然后在特定点云数据集上正式训练，学习语义信息，又称“微调”；最后在带标签的测试集上，计算模型预测点级语义标签的准确性. 因为这种方法使用的“监督信号”是点云自身的属性，模型的训练不需要人工标注的数据，因此被称为“无监督方法”.

近年，Transformer 在点云语义分割上取得巨大进展. 为了更好地将标准 Transformer 用跨模态数据训练，Point-BERT[50] 在不改变 Transformer 的架构的前提下，设计了一种自监督训练方法. 在预训练阶段，输入的点云先被划分为多个局部点云补丁，并使用简化的 PointNet 将补丁转变为点的向量映射，然后用预训练的标记生成器将点的向量映射转化为标记并添加随机掩码，然后输入 Transformer. 如此，Transformer 通过预测被添加掩码的标记而非语义标签，实现无监督预训练. 在测试阶段，模型在同一测试集上预测点的语义标签. 最近，为了解决 Point-BERT 可能为局部补丁生成错误的标记，Point-McBert[51] 在 Point-BERT 的基础上为每个补丁提供了多选的标记作为监督信号，进一步提高了模型性能.

2.5.2 跨模态监督网络

图像在描述同一物体分辨率比点云高，因此含有更多纹理信息，但缺乏空间信息；点云有准确的空间坐标，但分辨率低. 跨模态监督网络结合图像和点云，实现信息互补，进一步提高数据利用率和模型性能. 相关工作如 Pix4Point[54]，利用了大量低成本的图像来辅助点云语义分割，通过先在图像数据集上预训练 Transformer 编码器，并将其作为点云网络的编码器，在点云数据集上微调，实现了分割性能的叠加. 此外，为解决车载三维雷达点云和相机图像实时视场重合小，纯融合信息有限

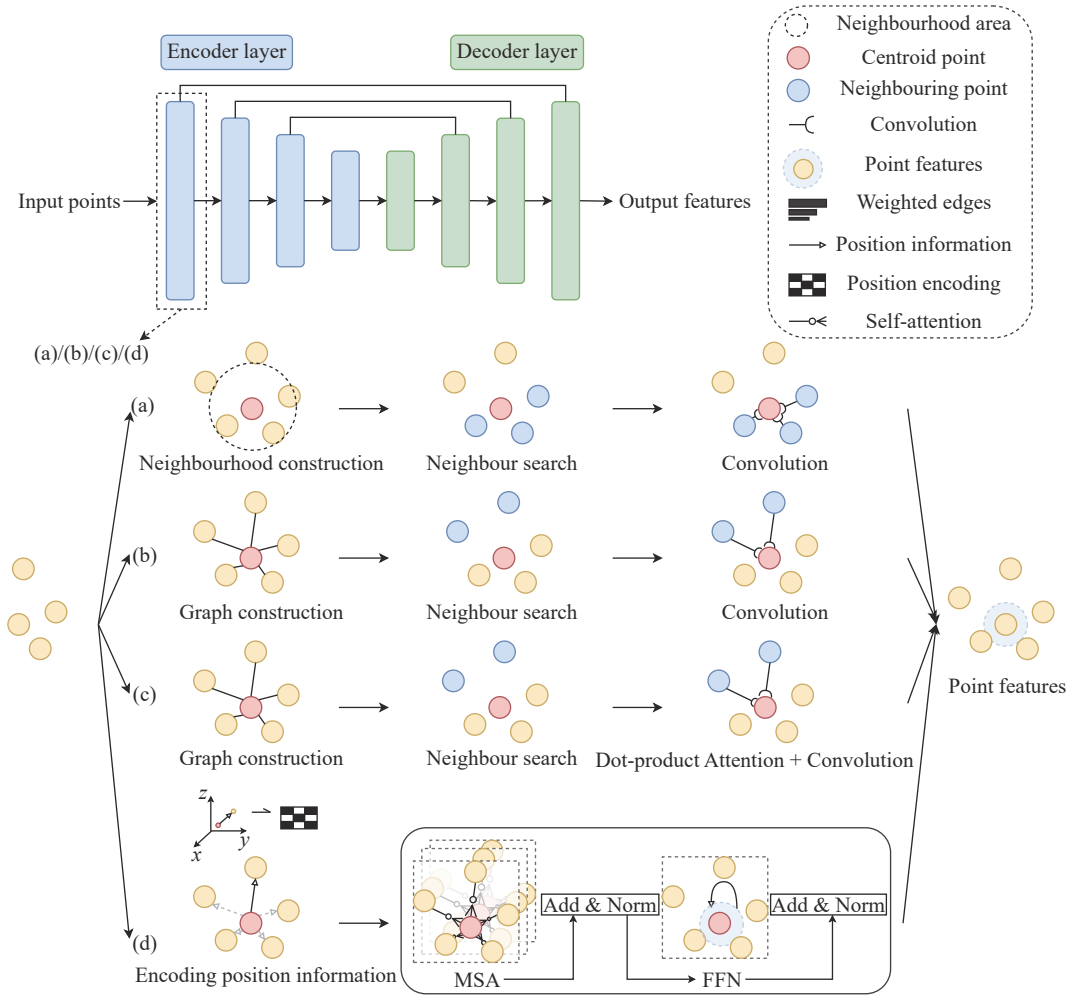

图 4　不同网络架构的基本过程示例(其中 MSA 为多头自注意力, Add 为残差连接, Norm 为正则化, FFN 为前馈网络). (a) 卷积网络; (b) 图卷积网络; (c) 注意力图网络; (d) Transformer

Fig.4　Illustrations of fundamental operations for different network structures (Note: MSA, Add, Norm, and FFN denote multi-head self-attention, residual connection, normalization, and feed-forward networks, respectively): (a) convolution network; (b) graph convolution network; (c) attention graph network; (d) transformer

的问题, 2DPASS[55] 通过在重叠区使用融合网络, 在非重叠区使用纯点云网络, 然后将融合网络学习到的语义信息再蒸馏到纯点云网络, 实现二维先验最大化辅助三维语义分割.

S3DIS 和 ScanNet 数据集是评估点云语义分割模型性能的常见数据集. 由于模型超参数如批尺寸和每个批的点数影响评估结果, 为了更直观且公平地对比基于不同网络架构的方法, 表 2 汇总了不同网络架构的典型模型在 S3DIS 上的部分超参数设置和评估结果. 由于公开的各模型在 ScanNet 数据集上的超参数设置评估较少, 表 2 仅展示模型在 S3DIS 上的部分超参数设置. 同时, 为更全面地对比各网络架构的模型性能, 表 2 还列举了基于 MLP 的模型. 由表 2 可以看出各种方法在语义分割任务上有一定的准确率, 同时仍存在一些显著的缺点:

(1) CNN 是最常见的网络, 但传统卷积不适于计算非结构化点云数据, 因此有大量工作通过重构卷积核, 实现了创新的网络架构. 但 CNN 的性能受到卷积类型、激活函数、网络深度等多因素影响, 给模型优化带来很大负担. 而 GNN 通过构建图来使用图卷积捕获点云特征, 避免了点云结构化带来的问题, 且构建的图节点是无序的, 因此模型具有置换不变性. 但 GNN 中非图部分的参数梯度和关系图中的参数梯度是一起反向传播的, 因此图传播模型难以描述; 且随着网络层数增加, 图计算量剧增, 因此难以构建深层 GNN 提取粒度更细的点云特征.

(2) 基于注意力的方法一般使用 CNN 或 GNN 作为骨干网络, 通过自适应加权学习关键信息, 但不同注意力模块的计算成本和优化方法差异较大, 模型鲁棒性欠佳. 基于 Transformer 的方法是

表 2 不同网络架构的点云语义分割方法在 S3DIS 和 ScanNet 上的评估性能对比

Table 2 Performance evaluation of different semantic segmentation architecture methods on the S3DIS and ScanNet datasets

Method	Input	Architecture	S3DIS batch size	S3DIS number of batch points	S3DIS 6-fold mIoU/%	S3DIS tested on Area5 mIoU/%	ScanNet test set overall accuracy/%	ScanNet test set mIoU/%
Pointnet[58]	points	MLP			47.60	41.1	73.9	14.69
Pointnet++[64]	points				54.50	51.5	84.5	38.28
MinkowskiNet[20]	voxels	CNN				65.35		72.1
PointConv[3]	points					50.34		55.6
PointWeb[23]	points		16		66.70	60.28	85.9	
A-CNN[24]	points						85.4	
KPConv[27]	points				70.60	67.1		68.4
PAConv[40]	points			4096		66.58		
HDGCN[18]	points	GNN			66.85	59.33		
TGNet[19]	points		16			58.70	66.2	
DGCNN[4]	points		12	4096	56.10			
ResGCN-28[26]	points					60.00		
SegGCN[32]	points		8	8192		63.60		58.9
SPH3D-GCN[35]	points		16	8192	68.90	59.5		61
GACNet[21]	points		16			62.85		
RandLA-Net[31]	points	Attention			70.00			
AGCN[34]	points			4096	56.63			
PAN[37]	points		32		66.30		86.7	42.1
PATs[22]	points	Pure Transformer				60.1		
PCT[42]	points					61.33		
PVT[43]	voxels+ points			4096		68.21		
Point Transformer[5]	points		8		73.50	70.4		
LFT-Net[46]	points					65.2		
Fast Point Transformer[47]	voxels					70.1		72.1
Point Transformer V2[56]	points					71.6		75.2
Stratified Transformer[48]	points	Hybrid Transformer	16			72		
MinkNet18+Segment-Fusion[67]	points					65.3		

基于注意力的方法的发展，具有对输入的点云具有天然的置换不变和刚体不变性，且由于网络中大量使用自注意力，从而能很好地学习语义分割需要的远距离上下文信息. 但该方法的参数量大，且因基于点积注意力的二次计算复杂度和多头机制，需要较高的计算成本.

(3) 基于其他网络架构的方法是为了解决全监督网络训练数据成本高而新兴的方法，如无监督网络能有效减少对标注数据的依赖，从而降低标注成本；跨模态网络能利用大量图像数据训练，从而减少了点云数据的成本. 但这类方法的模型性能受其他因素影响大，如代理任务影响无监督网络的性能，点云与图像的映射关系影响跨模态网络的性能.

3 挑战和潜在的研究方向

3.1 全监督学习缺乏训练数据

近年来，全监督深度学习方法逐渐在点云语义分割中占主导地位. 相比传统方法，这些全监督网络虽免去了人工特征设计，但在训练阶段需要使用大量带语义标签的数据，且这些数据的数量、密度、质量 (是否有噪音、遮挡) 和多样性对网络

的性能有很大影响. 虽然现有的公共数据集提供了几个室内和室外场景, 但它们不能充分满足实际应用的需求, 因此有以下 3 个有价值的研究方向.

3.1.1 开发高效的数据增强方法

数据增强是提高数据效率和节约标注成本的关键技术. 与二维图像数据增强不同, 三维数据增强方法需要针对不同类型的数据集, 如雷达点云数据集和 RGB-D 数据集, 设计特殊的增强策略. 常见的数据增强操作有旋转、缩放、裁剪、颜色抖动等, 需要根据不同的模型进行大量超参数的调优, 从而确定具体策略, 成本很高. 目前有研究人员通过设计三维点云数据增强网络, 来降低三维点云的数据增强成本. 如 Point Augment[68] 作为一个综合点云数据增强和分类的网络, 采用对抗学习策略来联合优化数据增强器和分类器, 从而让增强器能学习生成最适合分类器的增强样本.

3.1.2 设计合适的代理任务

代理任务 (Pre-text task)[69] 是在预训练阶段, 在训练数据集上学习数据的低级通用特征, 从而辅助模型在正式训练中学习高级语义特征的预训练任务. 合适的代理任务可以极大地降低模型对语义标签的学习依赖, 减少点云数据集需要的标注量, 使模型可在低标注成本的大规模数据集上训练. 目前有针对特定模型研究的代理任务, 如 LESS[70] 先在预训练阶段, 利用数据集中保留的每个部件的每个类别的唯一标注点, 通过弱监督学习生成点云的伪语义标签, 然后再在正式训练阶段, 结合生成的伪标注学习语义信息, 但这种代理任务的通用性还有待进一步研究.

3.1.3 融合多源数据

图像数据的采集成本低、数据分辨率高、数据处理较方便、计算复杂度较低, 但其数据质量受环境因素影响大, 如光照不足或拍摄角度不理想会导致图像模糊, 同时其缺乏目标的空间位置和几何纹理等信息. 而三维点云数据的可靠性高, 对环境变化如光照变化不敏感, 且可提供目标较精确的空间信息, 但其数据稀疏, 使用的采集设备成本高. 结合图像和雷达点云描述三维场景, 则可以用低成本带来丰富的数据信息, 但如何快速、准确地建立图像与点云的映射关系仍是一个亟需解决的问题. 目前通过融合图像与点云数据进行三维点云语义分割的方法, 除了如 Pix4Point 这样的跨模态监督网络, 主要有基于点云投影图像的方法[71], 如 RangeNet++[28] 先将激光点云投影为距离图像, 并对距离图像进行二维语义分割, 再将图像语义分割结果重投影回点云, 进行三维语义分割.

3.2 高效轻量的网络架构

结合表 2 和相关工作, 基于 Transformer 的方法一般比基于 CNN 或 GNN 的方法有更高的分割准确率, 但其参数量和计算复杂度较高. 实际应用中, 系统对模型的内存占用和计算消耗有着严格的限制, 这给 Transformer 在语义分割上的应用带来很大挑战. 最近, 在图像语义分割领域, SegNeXt[72] 使用深度可分离卷积来实现 Transformer 的注意力网络, 从而加快了 Transformer 的推理速度. 但在点云语义分割上, 由于二维卷积的效果不佳且三维卷积的计算复杂度普遍较高, 仍无有效降低 Transformer 计算复杂度的方法. 目前有研究如何用低计算复杂度的网络实现 Transformer 的性能, 如 PointNeXt[73] 通过优化网络训练策略和有效的模型优化, 用计算复杂度低的 CNN 取代 Transformer, 实现了比纯 Transformer, 如 Point Transformer[5] 更高的分割精度.

3.3 大规模实时三维语义分割

随着获取实时点云数据成本的进一步降低, 自动驾驶、交通运输、AR/XR 等领域对大规模实时点云语义分割的需求逐渐增加. 这方面面临的挑战主要有 3 个: (1) 采集场景往往比较混乱, 如街道、车流, 这些场景中有许多形状或表面反射率相似的物体, 且物体间遮挡现象严重, 使采集到的物体点云有较多重合和空洞, 给分割带来难度. (2) 场景的每一帧一般包含数万个点, 如果同时处理这些点会让模型的计算复杂度急剧增加, 从而对计算元件和存储器有很高的要求. (3) 现有模型的内存占用过高, 无法用于低存储空间的嵌入式处理器; 且模型推理速度过慢, 几乎不能匹配雷达毫秒级的实时帧获取速度. 最近有研究提出了一种轻量化、高推理速度的实时三维点云语义分割网络[74], 其在 NVIDIA RTX 3090TI 上的推断时间约为 38.5 ms, 在 ZCU104 MPSOC FPGA 平台上的实时处理速度可达每秒 56 帧, 这为自动驾驶车辆控制器的其他感知任务留出了充足的时间.

4 结语

本文开篇介绍了三维点云语义分割的具体任务, 并引出要综述的三维点云语义分割的三个层面. 第一层面是在第 1 节介绍了点云数据的获取方式、常用数据集及评价指标; 第二层面是在第 2 节按不同网络架构, 将点云语义分割方法分为基于 CNN、GNN、注意力、Transformer 和其他网络

架构的 5 类，并依模型出现的时间顺序介绍了不同类别从 2019 年到目前最新的研究进展，然后对比了各方法在 S3DIS 和 ScanNet 上的性能，并归纳总结了各类方法的优缺点、适用范围和应用场景；第三层面是针对第 2 节中不同方法存在的问题，在第 3 节提出了点云语义分割方法面临的挑战和潜在的研究方向. 通过本文，可以初步了解点云语义分割的具体任务，特别是不同网络架构模型的最新研究进展和性能对比.

参 考 文 献

[1] Riemenschneider H, Bódis-Szomorú A, Weissenberg J, et al. Learning where to classify in multi-view semantic segmentation // *European Conference on Computer Vision*. Zurich, 2014: 516

[2] Armeni I, Sener O, Zamir A R, et al. 3D semantic parsing of large-scale indoor spaces // 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Las Vegas, 2016: 1534

[3] Wu W X, Qi Z A, Li F X. PointConv: deep convolutional networks on 3D point clouds // 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Long Beach, 2020: 9613

[4] Wang Y, Sun Y B, Liu Z W, et al. Dynamic graph CNN for learning on point clouds. *ACM Trans Graph*, 2019, 38(5): 1

[5] Zhao H S, Jiang L, Jia J Y, et al. Point transformer // 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). Montreal, 2021: 16259

[6] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, 313(5786): 504

[7] Guo Y L, Wang H Y, Hu Q Y, et al. Deep learning for 3D point clouds: A survey. *IEEE Trans Pattern Anal Mach Intell*, 2020, 43(12): 4338

[8] Xie Y X, Tian J J, Zhu X X. Linking points with labels in 3D: A review of point cloud semantic segmentation. *IEEE Geosci Remote Sens Mag*, 2020, 8(4): 38

[9] He Y, Yu H S, Liu X Y, et al. Deep learning based 3D segmentation: A survey [J/OL]. *arXiv Preprint* (2021-3-10) [2022-12-17]. https://arxiv.org/abs/2103.05423&ampshy;

[10] Lahoud J, Cao J L, Khan F S, H, et al. 3D Vision with Transformers: A Survey [J/OL]. *arXiv preprint* (2022-8-8) [2022-12-17]. https://arxiv.org/abs/2208.04309

[11] Lu D N, Xie Q, Wei M Q, et al. Transformers in 3D point clouds: A survey [J/OL]. *arXiv preprint* (2017-5-24) [2022-12-17]. https://arxiv.org/abs/2205.07417

[12] Zeng J H, Wang D C, Chen P. A survey on transformers for point cloud processing: An updated overview. *IEEE Access*, 2022, 10: 86510

[13] Gao B, Pan Y C, Li C K, et al. Are we hungry for 3D LiDAR data for semantic segmentation? A survey of datasets and methods. *IEEE Trans Intell Transp Syst*, 2021, 23(7): 6063

[14] Hackel T, Savinov N, Ladicky L, et al. Semantic3D. net: A new large-scale point cloud classification benchmark [J/OL]. *arXiv preprint* (2017-5-24) [2022-12-17]. https://arxiv.org/abs/1704.03847

[15] Behley J, Garbade M, Milioto A, et al. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences // 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). Seoul, 2019: 9296

[16] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite // 2012 *IEEE Conference on Computer Vision and Pattern Recognition*. Providence, 2012: 3354

[17] Dai A, Chang A X, Savva M, et al. ScanNet: richly-annotated 3D reconstructions of indoor scenes // 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Honolulu, 2017: 5828

[18] Liang Z D, Yang M, Deng L Y, et al. Hierarchical depthwise graph convolutional neural network for 3D semantic segmentation of point clouds // 2019 *International Conference on Robotics and Automation* (*ICRA*). Montreal, 2019: 8152

[19] Li Y, Ma L F, Zhong Z L, et al. TGNet: Geometric graph CNN on 3-D point cloud segmentation. *IEEE Trans Geosci Remote Sens*, 2019, 58(5): 3588

[20] Choy C, Gwak J Y, Savarese S. 4d spatio-temporal convnets: Minkowski convolutional neural networks // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach, 2019: 3075

[21] Wang L, Huang Y C, Hou Y L, et al. Graph attention convolution for point cloud semantic segmentation // 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Long Beach, 2019: 10296

[22] Yang J C, Zhang Q, Ni B B, et al. Modeling point clouds with self-attention and gumbel subset sampling // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: 2019: 3323

[23] Zhao H S, Jiang L, Fu C W, et al. PointWeb: enhancing local neighborhood features for point cloud processing // 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Long Beach, 2019: 5565

[24] Komarichev A, Zhong Z C, Hua J. A-CNN: Annularly convolutional neural networks on point clouds // 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Long Beach, 2019: 7421

[25] Meng H Y, Gao L, Lai Y K, et al. VV-net: Voxel VAE net with group convolutions for point cloud segmentation // 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). Seoul, 2019: 8500

[26] Li G H, Müller M, Thabet A, et al. DeepGCNs: can GCNs go as deep as CNNs? // 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). Seoul, 2019: 9267

[27] Thomas H, Qi C R, Deschaud J E, et al. KPConv: flexible and deformable convolution for point clouds // 2019 *IEEE/CVF*

International Conference on Computer Vision (*ICCV*). Seoul, 2019: 6411

[28] Milioto A, Vizzo I, Behley J, et al. RangeNet++: fast and accurate LiDAR semantic segmentation // 2019 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*). Macau, 2019: 4213

[29] Ma Y N, Guo Y L, Liu H, et al. Global context reasoning for semantic segmentation of 3D point clouds // 2020 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*). Snowmass, 2020: 2931

[30] Shi H Y, Lin G S, Wang H, et al. SpSequenceNet: semantic segmentation network on 4D point clouds // 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Seattle, 2020: 4574

[31] Hu Q Y, Yang B, Xie L H, et al. Randla-net: Efficient semantic segmentation of large-scale point clouds // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle, 2020: 11108

[32] Lei H, Akhtar N, Mian A. SegGCN: efficient 3D point cloud segmentation with fuzzy spherical kernel // 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Seattle, 2020: 11611

[33] Xu C F, Wu B C, Wang Z N, et al. SqueezeSegV3: Spatially-adaptive convolution for efficient point-cloud segmentation // *European Conference on Computer Vision*. Glasgow, 2020: 1

[34] Xie Z Y, Chen J Z, Peng B. Point clouds learning with attention-based graph convolution networks. *Neurocomputing*, 2020, 402: 245

[35] Lei H, Akhtar N, Mian A. Spherical kernel for efficient graph convolution on 3D point clouds. *IEEE Trans Pattern Anal Mach Intell*, 2020, 43(10): 3664

[36] Wen X, Han Z Z, Youk G, et al. CF-SIS: Semantic-instance segmentation of 3D point clouds by context fusion with self-attention // *Proceedings of the* 28*th ACM International Conference on Multimedia*. Seattle, 2020: 1661

[37] Feng M T, Zhang L, Lin X F, et al. Point attention network for semantic segmentation of 3D point clouds. *Pattern Recognit*, 2020, 107: 107446

[38] Zhang G G, Ma Q H, Jiao L C, et al. AttAN: Attention adversarial networks for 3D point cloud semantic segmentation // *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. Yokohama, 2020: 789

[39] Huang H, Fang Y. Adaptive wavelet transformer network for 3D shape representation learning // *International Conference on Learning Representations*. Hefei, 2022: 1

[40] Xu M T, Ding R Y, Zhao H S, et al. PAConv: position adaptive convolution with dynamic kernel assembling on point clouds // 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Nashville, 2021: 3173

[41] Fan H H, Yang Y, Kankanhalli M. Point 4D transformer networks for spatio-temporal modeling in point cloud videos // 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Nashville, 2021: 14204

[42] Guo M H, Cai J X, Liu Z N, et al. PCT: Point cloud transformer. *Comp Visual Media*, 2021, 7(2): 187

[43] Zhang C, Wan H C, Shen X Y, et al. PVT: Point-voxel transformer for point cloud learning [J/OL]. *arXiv preprint* (2022-5-25) [2022-12-17].https://arxiv.org/abs/2108.06076

[44] Wan J, Xie Z, Xu Y Y, et al. DGANet: A dilated graph attention-based network for local feature extraction on 3D point clouds. *Remote Sens*, 2021, 13(17): 3484

[45] Wei Y M, Liu H, Xie T T, et al. Spatial-temporal transformer for 3D point cloud sequences // 2022 *IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*). Waikoloa, 2022: 1171

[46] Gao Y B, Liu X B, Li J, et al. LFT-net: Local feature transformer network for point clouds analysis. *IEEE Trans Intell Transp Syst*, 2023, 24(2): 2158

[47] Park C, Jeong Y, Cho M, et al. Fast point transformer // 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New Orleans 2022: 16949

[48] Lai X, Liu J H, Jiang L, et al. Stratified transformer for 3D point cloud segmentation // 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New Orleans, 2022: 8500

[49] Xu S J, Wan R, Ye M S, et al. Sparse cross-scale attention network for efficient LiDAR panoptic segmentation // *Proceedings of the AAAI Conference on Artificial Intelligence*. Online, 2022: 2920

[50] Yu X M, Tang L L, Rao Y M, et al. Point-BERT: Pre-training 3D point cloud transformers with masked point modeling // 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New Orleans, 2022: 19313

[51] Fu K X, YuanM Z, Wang M N. Point-McBert: A Multi-choice self-supervised framework for point cloud pre-training [J/OL]. *arXiv preprint* (2022-8-15) [2022-12-17]. https://arxiv.org/abs/2207.13226

[52] Zeng Z Y, Xu Y Y, Xie Z, et al. RG-GCN: A random graph based on graph convolution network for point cloud semantic segmentation. *Remote Sens*, 2022, 14: 4055

[53] Wu Y X, Liao K L, Chen J T, et al. D-former: A u-shaped dilated transformer for 3d medical image segmentation. *Neural Comput Appl*, 2022, 35: 1931

[54] Qian G C, Zhang X D, Hamdi A, et al. Improving standard transformer models for 3D point cloud understanding with image pretraining [J/OL]. *arXiv preprint* (2022-11-22) [2022-12-17]. https://arxiv.org/abs/2208.12259

[55] Yan X, Gao J T, Zheng C D, et al. 2DPASS: 2D priors assisted semantic segmentation on LiDAR point clouds // *European Conference on Computer Vision*. Tel Aviv, 2022: 677

[56] Wu X Y, Lao Y X, Jiang L, et al. Point transformer V2: Grouped vector attention and partition-based pooling [J/OL]. *arXiv preprint* (2022-10-11) [2022-12-17]. https://arxiv.org/abs/2210.05666

[57] Mousavian A, Pirsiavash H, Košecká J. Joint semantic

segmentation and depth estimation with deep convolutional networks // 2016 *Fourth International Conference on 3D Vision* (*3DV*). Stanford, 2016: 611

[58] Charles R Q, Hao S, Mo K C, et al. PointNet: deep learning on point sets for 3D classification and segmentation // 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Honolulu, 2017: 652

[59] Wu B C, Zhou X Y, Zhao S C, et al. SqueezeSegV2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a LiDAR point cloud // 2019 *International Conference on Robotics and Automation* (*ICRA*). New York, 2019: 4376

[60] Wu B C, Wan A, Yue X Y, et al. SqueezeSeg: convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LiDAR point cloud // 2018 *IEEE International Conference on Robotics and Automation* (*ICRA*). Brisbane, 2018: 1887

[61] Xu Q G, Sun X D, Wu C Y, et al. Grid-GCN for fast and scalable point cloud learning // 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Seattle, 2020: 5661

[62] Lei H, Akhtar N, Mian A. Octree guided CNN with spherical kernels for 3D point clouds // 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Long Beach, 2019: 9631

[63] Liang Z D, Yang M, Li H, et al. 3D instance embedding learning with a structure-aware loss function for point cloud segmentation. *IEEE Robotics Autom Lett*, 2020, 5(3): 4915

[64] Qi C R, Yi L, Su H, et al. PointNet++: Deep hierarchical feature learning on point sets in a metric space // *Advances in Neural Information Processing Systems*. Long Beach, 2017: 1

[65] Liu J W, Liu J W, Luo X L. Research progress in attention mechanism in deep learning. *Chin J Eng*, 2021, 43(11): 1499
(刘建伟, 刘俊文, 罗雄麟. 深度学习中注意力机制研究进展. 工程科学学报, 2021, 43(11): 1499)

[66] Guo M H, Xu T X, Liu J J, et al. Attention mechanisms in computer vision: A survey. *Comput Vis Media*, 2022, 8(3): 331

[67] Thyagharajan A, Ummenhofer B, Laddha P, et al. Segment-fusion: Hierarchical context fusion for robust 3D semantic segmentation // 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). New Orleans, 2022: 1236

[68] Li R H, Li X Z, Heng P A, et al. PointAugment: an auto-augmentation framework for point cloud classification // 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). Seattle, 2020: 6378

[69] Xiao A R, Huang J X, Guan D Y, et al. Unsupervised representation learning for point clouds: A survey [J/OL]. *arXiv preprint* (2022-6-5) [2022-12-17]. https://arxiv.org/abs/2202.13589

[70] Liu M H, Zhou Y, Qi C R, et al. LESS: Label-efficient semantic segmentation for LiDAR point clouds // *European Conference on Computer Vision*. Tel Aviv, 2022: 70

[71] Jhaldiyal A, Chaudhary N. Semantic segmentation of 3D LiDAR data using deep learning: A review of projection-based methods. *Appl Intell*, 2023, 53(6): 6844

[72] Guo M H, Lu C Z, Hou Q B, et al. SegNeXt: Rethinking convolutional attention design for semantic segmentation [J/OL]. *arXiv preprint* (2022-9-18) [2023-12-17]. https://arxiv.org/abs/2209.08575

[73] Qian G C, Li Y C, Peng H W, et al. PointNeXt: Revisiting PointNet++ with improved training and scaling strategies [J/OL]. *arXiv preprint* (2022-10-12) [2022-12-17]. https://arxiv.org/abs/2206.04670

[74] Xie X, Bai L, Huang X M. Real-time LiDAR point cloud semantic segmentation for autonomous driving. *Electronics*, 2021, 11(1): 11