基于点查询的弱监督 3D 点云语义分割算法研究

摘要

点云是常用的三维数据格式之一，因其极大程度保留了三维空间中物体的轮廓信息，在机器人、自动驾驶、工业检测等领域有广泛的应用。近年来，随着深度传感器的发展和普及，点云数据的获取变得越来越容易，点云数据集变得越来越庞大，对点云数据进行完全标注的成本也越来越高。针对这一问题，最直接的方法是利用弱监督学习，即给定一小部分标签进行训练，模型仍能保持较高的性能。本文围绕三维点云弱监督语义分割开展相关研究，复现基于点查询的弱监督 3D 点云语义分割算法，并在其基础上进行优化和改进，最后使用 0.1% 的标注数据在 SemanticKITTI 数据集上进行模型性能的验证。本文的主要工作如下：

1. 复现基于点查询的弱监督点云语义分割算法，并在其基础上进行改进。算法主要由点特征提取网络、点查询网络、多尺度信息融合网络三部分构成。点特征提取网络使用 RandLA-Net 的局部特征聚合（Local Feature Aggregation, LFA）模块进行特征聚合，该模块使用局部空间编码将邻域内所有点的位置信息与特征信息融合，通过注意力池化启发式地选出较为重要的特征，最后将局部空间编码和注意力池化堆叠多次以实现更好的效果。随机采样被用于减少点云的规模，增大网络的感受野。点查询网络通过 KNN 算法获取标记点邻域内的特征并将其作为标记点特征，为网络的推理提供更多依据，同时将稀疏的监督信号传递到邻域及更广阔的空间中。多尺度信息融合网络将提取出的四个尺度的邻域特征进行融合，并通过一系列简单的多层感知机直接推断语义标签，有利于网络结合浅层细节信息和高层抽象信息进行学习。

2. 报告本文改进的算法在上述数据集上的性能，并与其他先进的全监督、弱监督算法对比。数据显示，在 SemanticKITTI 数据集中，本文改进的算法性能上较复现的算法提高了 2.5%，且参数量少于该网络。本文还报告了算法在数据集上的可视化结果，用于判断本文方法与真实标签之间的差异。

实验表明本文的改进方法仅使用 0.1% 的标注数据即取得较为理想的效果，且简单轻量的网络更容易在真实场景中应用，本文为解决点云数据集标注成本过高的问题提供了解决方案和实现参考。

关键词：弱监督；3D 点云；点查询；监督信号

1 引言

近年来，随着消费级深度传感器的快速发展和普及，三维数据在自动驾驶、增强现实、机器人技术等领域有更广泛的应用。三维数据直接对真实场景进行建模，极大地保留了场景中的形状轮廓、空间尺度和几何结构等信息，有助于智能体实现场景的理解与感知。三维数据有多种表现形式，包括深度图、点云、网格和体素等 [5]，点云数据是欧式空间中带有坐标的一组点的集合。相比于其他三维数据表现形式，点云数据直接从物体的表面进行离散采样，最大程度地保留物体的原始轮廓信息。此外，点云数据的表现形式简单，适用于端到端的智能网络的学习。所以点云数据是场景感知、模式识别相关任务的首选表现形式。

随着深度神经网络的发展，计算机视觉成为人工智能热门的研究内容之一。物体分类、目标检测、语义分割是计算机视觉中的三大任务。语义分割旨在为图像或点云数据按照相应的语义类别进行细粒度划分，以不同颜色表征不同的语义信息（如图 1所示），给予计算机场景级的感知和理解能力，为自动驾驶、智能拾取、导航规划等任务提供决策依据。

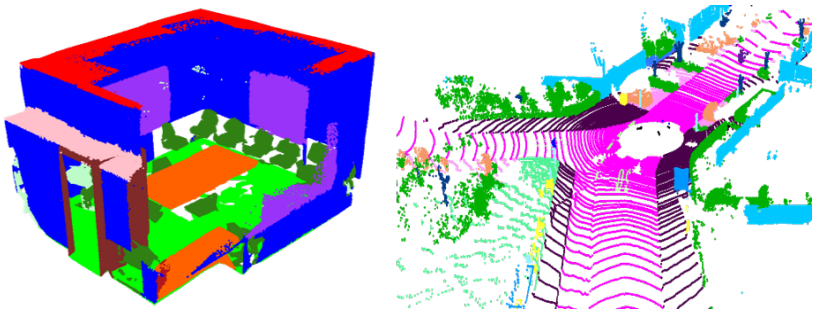

图 1. 点云语义分割效果图

在二维图像语义分割工作 [17,22] 的鼓舞下，越来越多的学者尝试对三维点云进行语义分割。相较于二维图像，三维点云不受光照条件影响，对环境的鲁棒性强；能提供深度信息，易于区别前景和背景。但同时三维点云也缺少二维图像所特有的归纳偏置，其无序、非结构化的特点决定其并不能使用简单的卷积神经网络，而需要通过精心设计的特征聚合算子 [2,33]，完成局部特征的提取。

同时，点云语义分割所需的海量标记数据也成为了其发展的阻碍。由于点云的数据量庞大，结构复杂，对点云进行完全标注的成本高昂，极大制约了点云数据集的发展和点云语义分割在实际场景中的应用。以室内 RGB-D 视频数据集 ScanNet [4] 为例，其包含 250 万张 RGB-D 图像和来自 707 个室内场景的 1513 次 3D 扫描，点云数据由 RGB-D 图像重建而成。由于使用了结构化传感器，研究者们可以使用 iPads 或手机等便携式设备进行 RGB-D 数据的获取。因此，如此海量的数据扫描只需要 20 人即可完成。相比之下，对这些数据进行标注却是一件费时费力的事情。据 [33] 统计，超过 500 人参与了数据的标注工作，为了确保标注的准确性，每个场景需要由 2 至 3 人分别标注，每个 3D 扫描的平均标注时间为 22.3 分钟。对于室外点云数据集 SemanticKITTI [1]，标注时间同样需要超过 1700 个小时。

在数据集越来越庞大，标注成本越来越高昂的现状下，弱监督学习成为缓解这一矛盾的有效方法。弱监督学习是一种仅需要弱标签或少量标签进行模型学习的方法，可以有效解决数据集中标签不足的问题。在给定少部分标记点的条件下，模型的性能基本接近全监督模型的性能，在成本和性能之间取得平衡。

随着人工智能的发展，数据集越来越庞大的同时，模型的复杂度也越来越高。深层的网络分辨率低，可以很好地抽象出图像的语义信息，但丢弃了图像的细节特征；浅层网络则相反，对于细节的表达能力强，但语义性弱。多尺度信息融合旨在将网络中不同层的特征通过一定的方法进行融合，以充分利用网络学习到的细节和全局语义信息。

2 相关工作

2.1 全监督点云语义分割

随着点云数据获取成本的下降，基于深度学习的全监督点云语义分割近年来取得了飞速的发展。

基于点的方法是点云语义分割常用的方法之一，该方法直接将离散的点作为网络的输入，不需要额外的前处理。开创性的工作是 Qi 等学者于 2017 年提出的 PointNet [20]，针对点云数据的无序性，作者首次提出使用对称性函数处理点云，即直接使用多层感知机 (MLP) 提取每个点的特征，使用最大池化汇聚全局特征后将全局特征和单点特征进行融合。但是，该方法并没有考虑点的局部上下文关系。团队于同年提出改进方案 PointNet++ [21]，通过采样 (sampling) 和分组 (grouping) 的方式在各个尺度上构建邻域并聚合邻域内的特征，最后使用 PointNet 网络提取该特征至下一层级。PointCNN [13] 提出 X 变换，通过 MLP 学习 X 变换矩阵并对输入的点云特征进行加权和置换，变换后的点云特征可以直接用于经典卷积运算；Thomas 等在 KPConv [29] 中提出核点可变形卷积算子，卷积的权值由点到核点的距离决定，为满足置换不变性，规定每个点都要与核点进行运算，卷积核还可以根据空间中点的分布进行变形以更好地提取点的特征。Hu 等人提出轻量级网络 RandLA-Net [8]，聚焦于局部几何轮廓信息的汇聚，从而可以应用随机采样提高运行效率，在处理大场景点云时有较好的效果。

基于投影的方法得益于成熟的二维图像处理技术。学者通常采用透视投影、球面投影或垂直投影的方法，将三维点云投影成二维图像，再利用成熟的卷积神经网络对二维图像进行特征提取。华南理工大学团队提出多传感器感知融合网络 PMF [40]，将三维点云投影为透视图的同时与真实的二维图像进行特征融合，实现跨模态学习。Cortinhal 等人提出的 Salsanext [3] 则是将点云球面投影为 RV 视图，应用空洞卷积提高模型的感受野，从而捕获更多的上下文信息。PolarNet [36] 则是应用垂直投影的方法，将三维点云投影为鸟瞰图，为了适应雷达点云近密远疏的特点而采用环形划分的方式，充分考虑鸟瞰图点云空间分布不均的问题。

基于体素的方法往往将不规则分布的点云体素化为规则的立方体，相较于投影的方法，体素直接在三维空间中进行划分，包含更多的空间几何信息。但由于点云的非均匀分布，造成体素的高度冗余和额外的内存消耗。Zhu 等 [39] 提出圆柱体划分和圆柱体卷积运算，一定程度上解决激光雷达点云密度分布不均的问题。

尽管这些全监督的方法在现有的数据集上取得了显著的效果，但是它们均需要大量的标记数据作为监督信号进行训练，这在现实的应用场景中是昂贵而奢侈的。

2.2 无监督点云语义分割

当数据集没有标签时，如何挖掘数据内在的规律作为监督信号成为无监督点云语义分割的关键。较早的工作是 Saudar 等 [24] 通过重新组合部分被随机移位的点云，从而在原始点

云数据中学习高级的语义表示。Song 等 [25] 利用点云序列中的运动信息进行学习，将运动物体的几何一致性作为约束条件提供有效的监督信号。尽管这些方法已经取得令人鼓舞的进步，但性能相比于全监督和弱监督学习仍存在不少差距。

2.3 弱监督点云语义分割

弱监督点云语义分割按标签的类别可以分为两类：不准确监督和不完全监督。

不准确监督具体表现在点云仅含有云级（cloud-level）或段级（seg-level）等标签，这一类标签也称为弱标签，因为其并不直接标注每个点的类别。代表性工作是 Wei 等 [31] 提出的点云类激活图，通过子云级的弱标签训练一个分类网络，利用卷积神经网络的定位能力 [37] 和类激活映射图技术 [38] 生成点级伪标签。清华大学 Tao 等人 [27] 利用过分割预处理点云场景，为实例的中心段进行标注，并通过结构分组和语义分组网络进行聚类。由于云级和段级标签往往缺少细节信息，导致物体边缘轮廓的分割效果较差，这在自动驾驶等对边界判断要求较高领域是难以接受的。

不完全监督表示训练过程中，只能通过少部分有标签的点完成对模型的约束。较早的工作是 Xu 等人 [35] 通过设计梯度相似、旋转不变、颜色平滑等多个分支约束，为所有的点提供充足的监督信号。Li 等人 [12] 提出局部对比正则化和全局对比正则化，充分利用生成的伪标签进行学习。本文的基线网络 SQN [7] 提出语义查询网络，在每一特征层中查询标记点临近的点，将监督信号尽可能地传递到更广的区域。Unal 等人 [30] 提出第一个涂鸦标注的激光雷达语义分割数据集，并提出训练、伪标签、蒸馏三个可与任意模型结合的独立的模块。LESS [16] 对原始点云采用预分割，减少标注工作量的同时获得了大量传播标签，设置教师网络学习点云序列，对学习单帧点云的学生网络进行蒸馏。Kong [10] 等人利用激光雷达点云场景中的先验结构进行一致性约束，提出 LaserMix 对点云进行划分混合。这些方法在仅有少量标记点的条件下充分挖掘有标记数据和无标记数据之间的关系，为更多的点提供点级的监督信号，取得了接近全监督语义分割的性能。

3 本文方法

本章将详细阐述本文改进算法的实现。改进的算法以 RandLA-Net [8] 的局部特征聚合模块为主干（backbone），在此基础上应用点查询网络和多尺度信息融合网络以充分利用稀疏的监督信号。

3.1 问题陈述

对于一个不完全监督的点云语义分割任务，本文设计了一个深度学习框架，它直接使用无序的点作为输入。点云可以表示为一个有 N 个点的集合 $\{P_n|n = 1, 2, ..., N\}$，其中 P_n 表示点云中的第 n 个点且 $P_n \in \mathbb{R}^3$，即每个点包含世界坐标 (x, y, z) 信息。对于点的特征可以用集合 $F \in \mathbb{R}^{N\times d}$ 表示，即每个点有 d 维度的特征（包括颜色、强度等信息）。

由于是不完全监督，只有少部分的点拥有标签信息，本文称这些点为标记点或查询点，设有 M 个标记点，记为 $\{PL_n|n = 1, 2, ..., M\}$，$PL \in \mathbb{R}^{M\times 3}$；同理，这些点的特征记为 $FL \in \mathbb{R}^{M\times d}$，它们的标签记为 $L \in \mathbb{R}^M$。

设本文的改进算法为 $\mathcal{F}$，对于训练阶段，所有点和标记点作为算法的输入，标记点的预测标签 $\hat{Y} \in \mathbb{R}^M$ 作为算法的输出，记为 $\hat{Y} = \mathcal{F}(P, F, PL)$。使用损失函数 $\mathcal{L}$ 计算预测的偏差 $\mathcal{L}(\hat{Y}, L)$ 进行训练。对于推理阶段，将所有点作为算法的输入，预测所有点的标签 $\hat{Y} = \mathcal{F}(P, F) \in \mathbb{R}^N$。

3.2 网络整体框架

本文的网络总体架构如图2所示，网络由点特征提取网络、点查询网络和多尺度信息融合网络三部分组成。为了简洁起见，图2仅展示单一查询点时模型的结构，在实际训练中，查询点占所有点比例的 0.1%。

在点特征提取网络中，理论上可以使用任何点云特征聚合算子进行特征提取，本文选择轻量级网络 RandLA-Net [8] 中的局部特征聚合（Local Feature Aggregation, LFA）模块，同时使用随机采样（Random Sample, RS）降低点云的规模以更好地聚合上下文信息。

经过四层 LFA 和四次随机采样，输入的点云被提取为四个尺度的特征信息，此时输入标记点，通过点查询网络，使用 K 近邻算法获得标记点周围的 K 个点，并将这 K 个点的特征进行拼接。对每一个尺度拼接后的信息通过逐元素求和的方式进行融合，由于监督信号较弱，直接使用一系列简单的多层感知机进行特征维度变换，最后输出查询点被分类到各个类别的置信度。

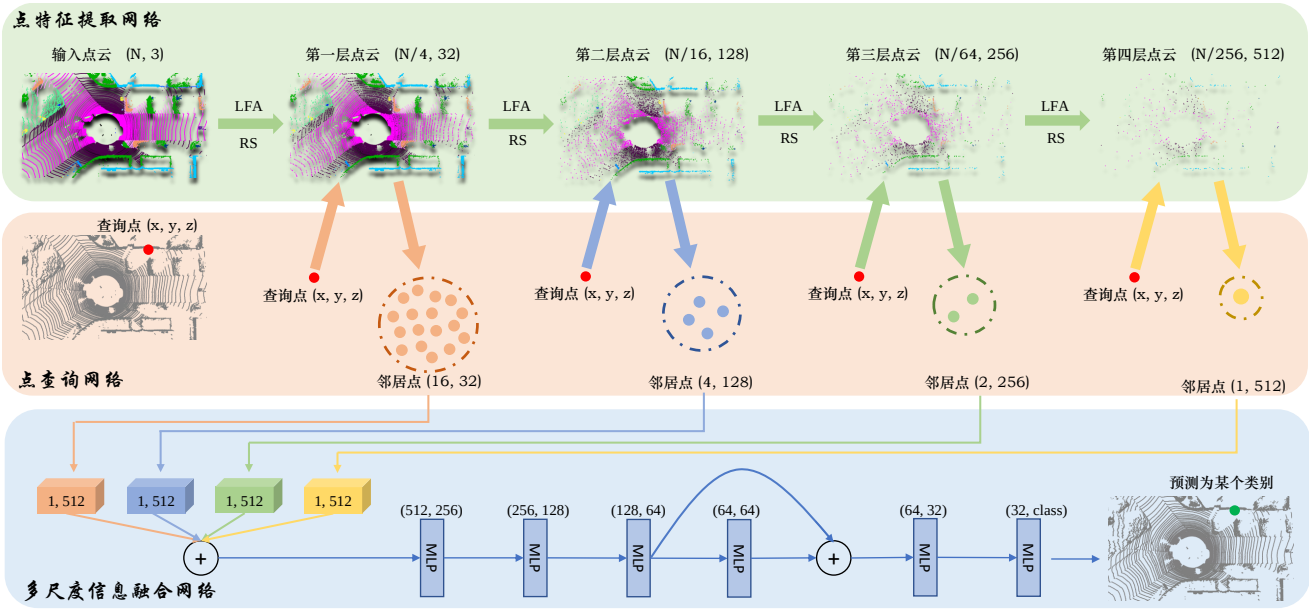

图 2. 网络整体架构图（以一个查询点为例）

3.3 点特征提取网络：局部特征聚合模块

点特征提取网络使用局部特征聚合模块 [8]（LFA）进行特征提取；使用随机采样（RS）降低点云的规模。每经过一层 LFA 和 RS，点的数量减少为原来的 1/4，特征的数量增加为原来的 2 倍（第一层除外）。

LFA [8] 的整体架构如图3所示，其将局部空间编码（Local Spatial Encoding）和注意力池化（Attentive Pooling）模块连续堆叠两次，并采用 ResNet [6] 残差连接的方式将输入的信息与输出的信息结合，以增大网络的感受野。

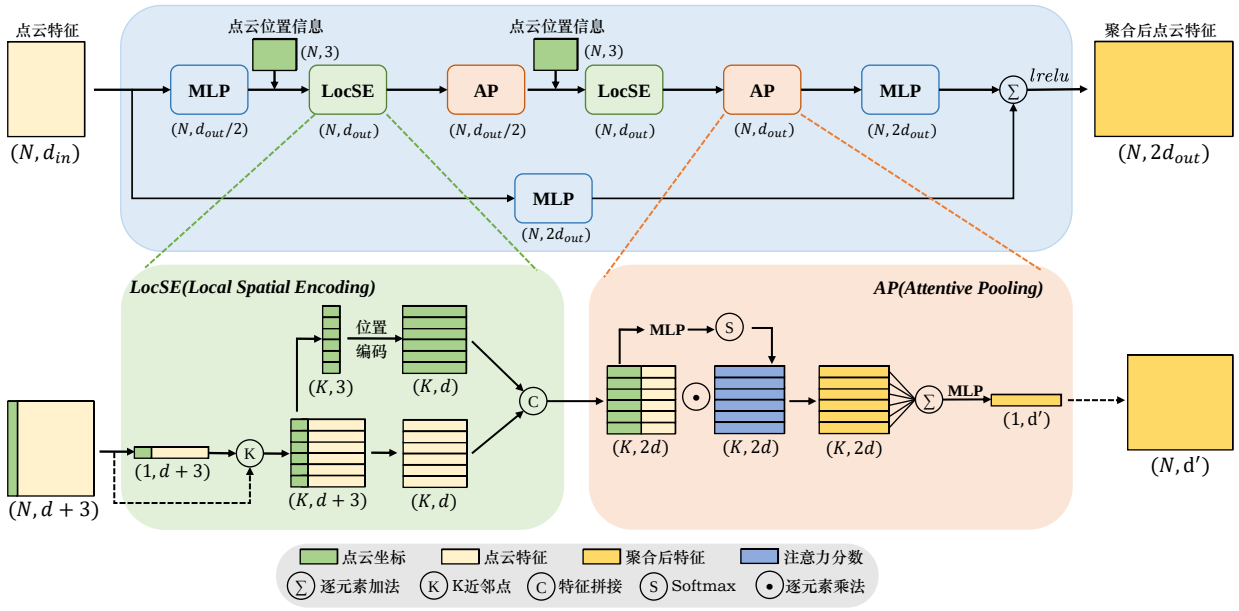

图 3. 局部特征聚合模块整体架构

3.3.1 局部空间编码（Local Spatial Encoding）

给定所有点的坐标 $P \in \mathbb{R}^{N\times 3}$ 和所有点的特征 $F \in \mathbb{R}^{N\times d}$，局部空间编码将点邻域的空间位置信息嵌入到点的特征中，从而使邻域内点的特征更加注重其局部空间结构。在网络每一层的特征提取过程中，局部空间编码都将原始的坐标信息对特征进行嵌入，不断地重复强调点的邻域空间信息，从而使整个网络有效地学习到复杂的几何轮廓。

具体包含以下步骤：

1. 寻找 K 近邻点。对于点云中的某个点，采用简单的 K-近邻算法根据空间欧式距离选择离该点最近的 K 个点。

2. 位置编码。对于 K 个近邻点 $\{P_n | n = 1, 2, ..., K\}$，设其中心点为 P_i，将中心点的坐标、K-近邻点的坐标、相对位置和欧式距离这些位置信息进行拼接并通过多层感知机进行编码，最终得到 $K \times d$ 维的空间信息特征：

$$MLP(P_i \oplus P_k \oplus (P_i - P_k) \oplus ||P_i - P_k||) \tag{1}$$

3. 空间信息嵌入。将经过位置编码后获得的空间信息与 K 近邻点的特征进行拼接 $P \oplus F \in \mathbb{R}^{K\times 2d}$。

3.3.2 注意力池化（Attentive Pooling）

由于平均池化和最大池化会造成大部分信息丢失，Hu [8] 等人在此提出一个精确到每个特征的注意力池化模块。

具体步骤为，对拼接后的特征经过一个可学习的线性 MLP 和 $Sigmoid$ 激活函数，从而计算每个特征的注意力分数；再将注意力分数与拼接的特征相乘，表示网络对各个特征的重要性进行打分，较为重要的特征有较高的数值，对下游网络的影响较大；最后将近邻点的特征进行求和作为中心点的特征。

3.4 点查询网络

经过点特征提取网络对输入点云进行四层的特征提取后，对于每一个尺度，点查询网络致力于为查询点（有标记点）收集尽可能多的邻域特征，并充分利用这些特征进行推理，同时在训练时将稀疏的监督信号传递到邻域中的其他点，这是实现弱监督学习的关键。

由于点云数据存在局部语义相似性，可以假设查询点的临近点与查询点共享相似的语义信息和类别标签；且经过点特征提取网络后，每一个点均聚合了其邻域内其它点的特征，即使在物体边缘处邻域内点的类别不同，这些点所携带的特征仍可以作为查询点的特征。因此，查询点与其临近点即可共享相似的监督信号，以实现将稀疏的监督信号传递到更宽更广的上下文语义空间中，使网络进行充分学习。

点查询网络主要包括两个部分：邻域点搜索和点特征拼接。

3.4.1 邻域点搜索

邻域点搜索旨在查询并取出查询点邻域内点的特征，如图4所示。邻域采用 K-近邻算法定义，即给定某个查询点 PL_i 及其 (x, y, z) 坐标，邻域内点的集合 $\{P_i^1, P_i^2, ..., P_i^K\}$ 为欧式空间中距离查询点 PL_i 最近的 K 个点，同时通过索引获取这些点的特征 $\{F_i^1, F_i^2, ..., F_i^K\}$。需要强调的是，虽然这些点都是没有标记的，但他们将会共享查询点的标记。

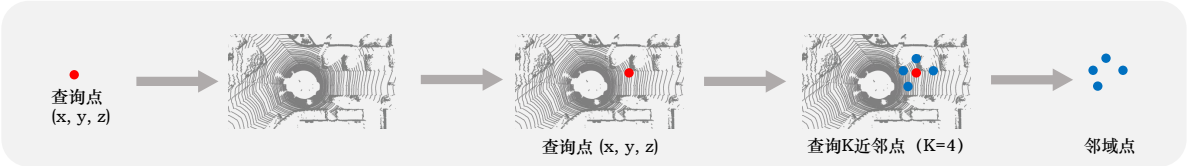

图 4. 邻域点搜索方法

在点查询网络中，需要在每一个尺度的点特征中进行邻域点搜索。由于不同尺度点的规模不同，网络的感受野也不同：浅层的网络点数量多，特征维数低，点所聚合的邻近点特征少，网络的感受野较小；深层的网络则相反，仅有较少的点，且每个点聚合的特征多，所涉及的邻域广，网络的感受野较大。基于此，邻域点搜索在不同的尺度中查询不同数量的点，以更好地适应网络的感受野变化。具体而言，上述 K 值的在每一个尺度的选取为 $[16, 4, 2, 1]$。

3.4.2 点特征拼接

对于查询到的邻域点特征 $\{F_i^1, F_i^2, ..., F_i^K\}$，采用特征图拼接的方式进行融合：

$$\vec{R} = F_i^1 \oplus F_i^2 \oplus ... \oplus F_i^K \tag{2}$$

每一个尺度中查询到的 K 个向量都需要通过拼接的方式转化为一个紧凑的表示，其代表某个尺度下查询点的邻域信息。总体而言，点查询网络使用标记点邻域的特征代替标记点的特征，在更大的接受域中推断点的类别标签，并同时将监督信号传递到更广阔的空间中。需要注意的是，即使标记点本身并不存在（经过多次随机采样后，标记点有较大的概率被丢弃），依旧可以使用邻域的特征表示标记点的特征。

3.5 多尺度信息融合网络

多尺度信息融合网络将来自点查询网络的四个尺度的特征进行融合，并通过一系列简单的多层感知机直接推断标记点的类别，使网络同时关注高维语义抽象信息和低维局部细节信息。

具体而言，经过点特征拼接后，四个尺度的邻域特征分别表示为四个紧凑、有相同维度的特征向量 $\{\vec{R_1}, \vec{R_2}, \vec{R_3}, \vec{R_4}\}$，一个简单的多尺度信息融合网络将这些不同尺度的特征向量进行逐元素加法融合，并通过多层 MLP 得到估计类别：

$$\hat{Y} = MLP(\vec{R_1} + \vec{R_2} + \vec{R_3} + \vec{R_4}) \tag{3}$$

网络能够进行充分训练得益于多尺度信息融合网络的简单设计。由于网络的监督信号过于稀疏（0.1% 标记数据），设计复杂的多尺度信息融合网络将会引起灾难性的过拟合现象。选择逐元素加法而不是拼接的融合方式同样出于这样的考虑，由于四个尺度的特征向量维度相同，进行逐元素加法可以极大限度地节省网络参数，避免过拟合现象，使用比 baseline 更少的参数实现更好的效果。

此外，多尺度融合网络也使得监督信号更为容易地传播到浅层网络。由于融合了浅层网络的特征，对于浅层网络的一部分梯度不再需要经过深层网络的传递，有助于浅层网络更好地进行学习。

3.6 实现细节

本节将简要介绍网络实现中的细节部分，包括损失函数的使用、多层感知机的细节实现、优化器，学习率的选择和超参数设置。

3.6.1 损失函数

本文使用 Focal Loss [15] 作为网络的损失函数，Focal Loss 最早用于解决目标检测中正负样本不平衡和难例学习问题，对于语义分割中的类别不平衡和难分类样本同样有较好的效果。

focal loss 在传统交叉熵损失函数的基础上，针对类别不平衡增加了一个权重因子 α_t；针对难例学习提出了调节因子 $(1-p_t)^\gamma$，完整公式如下：

$$FL(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t) \tag{4}$$

其中，p_t 表示网络推理的置信概率

$$p_t = \begin{cases} p & if\ y=1 \\ 1-p & otherwise \end{cases} \tag{5}$$

置信概率 p_t 较小（难区分样本）时，调节因子接近 1，不影响损失的大小；置信概率 p_t 较大（易区分样本）时，调节因子接近 0，损失大幅下降，即通过减少易区分样本的损失间接增加难例的损失。

3.6.2 其他实现细节

使用 1×1 卷积 [14] 表示共享的多层感知机。为加速网络收敛，防止过拟合，本文在 1×1 卷积层后添加批规范化 [9] 层（batch normalization layer）并使用 LeakyRule [18] 作为网络的激活函数，使用截断正态分布进行权重的初始化。使用 Adam 优化器优化网络参数，学习率初始值为 0.01，使用阶梯式学习率衰减，每轮（epoch）下降 5%，最多训练轮次为 100 轮（epoch）。损失函数的聚焦参数 $\gamma=2$，标注数据为在所有数据中随机选择的 0.1%。

4 复现细节

4.1 与已有开源代码对比

本文复现的工作为 SQN [7]，该工作已有开源代码，本文复现和改进的算法与该算法的区别如下：

- SQN [7] 开源的代码采用 **Tensorflow** 深度学习框架编写，复现代码采用 **Pytorch** 框架编写。采用不同的框架复现后算法性能与原论文报告的性能接近。表为复现后的算法与原算法在 S3DIS 数据集上的部分结果。

Methods	mIoU(%)	car	bicycle	motorcycle	truck	other-ve.	person	bicyclist	motorcyclist	road	parking	sidewalk	other-gr.	building	fence	vegeta.	trunk	terrain	pole	traffic-sign
SQN(Paper)	50.8	92.1	25.3	30.1	36.7	26.0	36.4	39.3	7.2	90.5	56.8	72.9	19.1	84.8	53.3	80.8	59.1	67.0	44.5	44.0
SQN(reim.)	50.3	92.6	21.4	25.4	38.5	33.0	36.9	40.3	6.0	89.5	54.9	71.2	19.8	84.8	52.6	80.8	58.7	62.1	45.4	41.2

表 1. SQN 复现代码与原代码的性能对比

- 修改点查询网络中在不同层点云下查询的邻近点个数。在 SQN [7] 的点查询网络中，对于每一层的点云特征，将标注点（查询点）邻近的 3 个点的特征作为查询点的特征；而在改进算法中，考虑到在浅层时，网络的感受野较小，提取出的点云邻域信息较少，故在浅层的点云中，将查询点周围更多的邻近点的特征作为查询点的特征，在深层的点云中，将查询点周围更多的邻近点的特征作为查询点的特征。具体而言，如图2所示，对于网络中的四层点云，设置查询的邻近点数目为 $[16,4,2,1]$，而原论文中则为 $[3,3,3,3]$

- 为了减少网络参数，降低过拟合风险，将多个尺度的特征进行相加从而进行特征融合，而不是原文采用的拼接的方式。由于特征维度的改变，在网络结构上也需要进行简单地修改：将多尺度信息融合网络中第一层全连接网络从 928 维修改为 512 维。

- 考虑到数据集中的各个类别之间的不平衡，为了使数据集中样本较少的类别（难例）更好地学习，修改网络的损失函数：采用 Focal Loss 替代原代码中的 Cross-Entropy Loss。

5 实验结果分析

5.1 实验设置

本文的所有实验均在 Ubuntu 18.04 LTS、Intel(R) Xeon(R) Gold 6132 CPU 以及 Tesla V100 PCIe 32GB 的设备上进行，使用 Pytorch 框架进行代码实现。

使用室外点云数据集 SemanticKITTI [1] 进行实验。SemanticKITTI 数据集是一个自动驾驶点云数据集，其包含从德国的市中心、住宅区以及卡尔斯鲁厄周围的高速公路场景和乡村道路采集激光雷达点云数据。数据集由 22 个序列组成，每个序列由大量连续的点云场景帧构成。对于 SemanticKITTI 数据集的划分，将序列 07 及 910 作为训练集，序列 8 作为验证集，序列 1122 作为测试集。

5.2 评价指标

常用的语义分割评价指标有总体精度(Overall Accuracy, OA)和平均交并比(mean Intersection-over-Union, mIoU)。

总体精度（OA）描述预测正确的样本占所有样本的比例。平均交并比（mIoU）描述预测集合和真实值集合的交集与并集之间的比例。其计算公式如下：

$$\mathbf{mIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{i=0}^{k}p_{ij}+\sum_{i=0}^{k}p_{ji}-p_{ii}} \tag{6}$$

其中 p_{ij} 表示将 i 预测为 j，p_{ii} 与 p_{ji} 同理。

5.3 在 SemanticKITTI 数据集上的结果

表2报告了本文改进的算法在 SemanticKITTI 数据集（测试集）上的性能，表中的其余数据来源于 SQN [7]。其中粗体表示弱监督学习中表现最好的数据。

表 2. 不同方法在 SemanticKITTI 数据集上的定量结果

Setting	Methods	mIoU(%)	Params(M)	road	sidewalk	parking	other-ground	buliding	car	truck	bicycle	motorcycle	other-vehicle	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traffic-sign
Full supervision	PointNet [20]	14.6	3.00	61.6	35.7	15.8	1.4	41.4	46.3	0.1	1.3	0.3	0.8	31.0	4.6	17.6	0.2	0.2	0.0	12.9	2.4	3.7
	PointNet++ [21]	20.1	6.00	72.0	41.8	18.7	5.6	62.3	53.7	0.9	1.9	0.2	0.2	46.5	13.8	30.0	0.9	1.0	0.0	16.9	6.0	8.9
	SPG [11]	17.4	0.25	45.0	28.5	0.6	0.6	64.3	49.3	0.1	0.2	0.2	0.8	48.9	27.2	24.6	0.3	2.7	0.1	20.8	15.9	0.8
	SPLATNet [26]	18.4	0.80	64.6	39.1	0.4	0.0	58.3	58.2	0.0	0.0	0.0	0.0	71.1	9.9	19.3	0.0	0.0	0.0	23.1	5.6	0.0
	TangentConv [28]	40.9	0.40	83.9	63.9	33.4	15.4	83.4	90.8	15.2	2.7	16.5	12.1	79.5	49.3	58.1	23.0	28.4	8.1	49.0	35.8	28.5
	SqueezeSegV2 [32]	39.7	1.00	88.6	67.6	45.8	17.7	73.7	81.8	13.4	18.5	17.9	14.0	71.8	35.8	60.2	20.1	25.1	3.9	41.1	20.2	36.3
	DarkNet21Seg [1]	47.4	25.00	91.4	74.0	57.0	26.4	81.9	85.4	18.6	26.2	26.5	15.6	77.6	48.4	63.6	31.8	33.6	4.0	52.3	36.0	50.0
	DarkNet53Seg [1]	49.9	50.00	91.8	74.6	64.8	27.9	84.1	86.4	25.5	24.5	32.7	22.6	78.3	50.1	64.0	36.2	33.6	4.7	55.0	38.9	52.2
	RangeNet53++ [19]	52.2	50.00	91.8	75.2	65.0	27.8	87.4	91.4	25.7	25.7	34.4	23.0	80.5	55.1	64.6	38.3	38.8	4.8	58.6	47.9	55.9
	SalsaNext [3]	54.5	6.73	90.9	74.0	58.1	27.8	87.9	90.9	21.7	36.4	29.5	19.9	81.8	61.7	66.3	52.0	52.7	16.0	58.2	51.7	58.0
	LatticeNet [23]	52.2	-	88.8	73.8	64.6	25.6	86.9	88.6	43.3	12.0	20.8	24.8	76.4	57.9	54.7	34.2	39.9	60.9	55.2	41.5	42.7
	PolarNet [36]	54.3	14.00	90.8	74.4	61.7	21.7	90.0	93.8	22.9	40.2	30.1	28.5	84.0	65.5	67.8	43.2	40.2	5.6	61.3	51.8	57.5
	RandLA-Net [8]	55.9	1.24	90.5	74.0	61.8	24.5	89.7	94.2	43.9	47.4	32.2	39.1	83.8	63.6	68.6	48.4	47.4	9.4	60.4	51.0	50.7
	SqueezeSegV3 [34]	55.9	26.00	91.7	74.8	63.4	26.4	89.0	92.5	29.6	38.7	36.5	33.0	82.0	58.7	65.4	45.6	46.2	20.1	59.4	49.6	58.9
Weakly supervision	SQN(0.1%) [7]	50.8	1.05	**90.5**	72.9	56.8	19.1	84.8	92.1	36.7	**39.3**	**30.1**	26.0	80.8	59.1	**67.0**	36.4	25.3	**7.2**	53.3	44.5	44.0
	Ours(0.1%)	**53.3**	**0.89**	90.3	**73.5**	**58.4**	**25.1**	**87.1**	**93.4**	**39.6**	23.8	27.8	**33.2**	**82.9**	**64.3**	66.5	**44.8**	**43.5**	7.0	**57.0**	**46.8**	**46.7**

值得注意的是，本文改进的方法在仅使用 0.1% 的标记数据前提下超过了某些全监督算法，与先进的全监督算法 RandLA-Net [8] 接近。在同样的场景下，全监督算法需要 45000 个

标记数据，而本文改进的弱监督算法仅需要 45 个标记数据，根据 [7] 统计的数据，标记时间缩短了 98%，而性能是先进全监督算法的 95%。

与基线网络 SQN [7] 相比，本文改进的算法也取得了令人鼓舞的进步。使用相同比例的标记数据，本文改进的算法性能上比基线网络提高了 2.5%，且参数量相较于基线网络更少，这更有利于网络在小型设备上进行部署。

一些在 SemanticKITTI 数据集上语义分割的定性结果可视化如5所示，红圈标注出错误分类的例子。

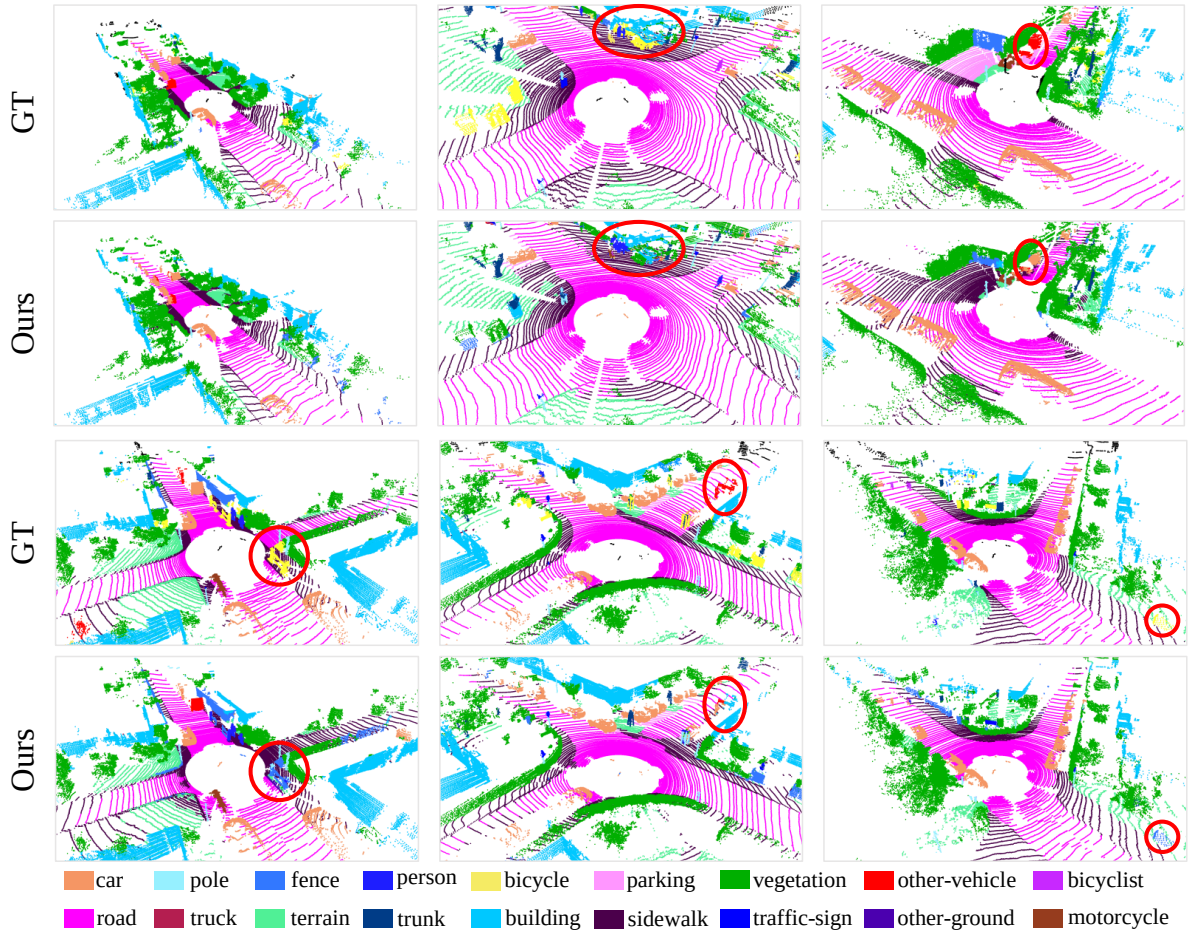

图 5. 本文改进方法在 SemanticKITTI 数据集 Sequence08 中的定性结果

6 总结与展望

6.1 总结

三维点云语义分割在自动驾驶、机器人等领域有广泛应用，但由于点云数据量大，结构复杂，对点云进行完全标注的成本太高，制约了点云的应用与发展。本文针对该问题重点研究了弱监督点云学习方法，并对基于点查询的弱监督点云语义分割算法进行改进，具体的研究内容和工作如下：

(1) 充分调研三维数据格式、点云数据格式的特点和其深度学习方法；充分调研并举例说明对点云数据集进行完全标注的庞大代价，并阐述弱监督学习的方法与必要性，最后简要介

绍多尺度信息融合网络。

(2) 充分调研本文研究内容的现状与相关工作。包括全监督、无监督和弱监督点云语义分割方法。

(3) 针对点云数据集完全标注成本过高的问题，本文复现了一个弱监督点云语义分割算法，并在此基础上进行改进。该算法由点特征提取网络、点查询网络和多尺度信息融合网络三部分构成。点特征提取网络采用 LFA [8] 模块进行信息聚合，采用随机采样减少点云的规模以增大网络的感受野；点查询网络将点的邻域特征替代单点的特征，将监督信号传递到更广的上下文空间中；多尺度信息融合使网络同时聚焦浅层和深层的信息，减少细节部分信息的丢失，提高小物体的分割效果，直接利用浅层信息进行预测也有利于梯度的传播，使浅层得到更加充分的训练。

(4) 进行实验验证算法性能。在 SemanticKITTI [1] 数据集上与先进的全监督、弱监督点云语义分割方法相比，改进后的算法平均交并比为 53.5%，接近甚至超越部分先进的全监督算法，超过大部分弱监督算法，并于改进前的算法相比平均交并比提高了 2.5%。

6.2 局限与展望

本文改进的方法虽在少量的标注下取得令人鼓舞的效果，但由于标注数据过少，监督信号过于稀疏，限制了网络向着更深，更复杂，更庞大的方向发展。直接的解决方案是通过多任务学习，设置多个损失函数，挖掘更多的监督信号。跨模态学习也是弱监督点云语义分割的另一个发展方向。将点云特征与其对应的图像特征进行融合（fusion），图像稠密的纹理信息与点云清晰的轮廓信息形成互补，为网络提供更详细更准确的监督信号。

参考文献

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019.

[2] Alexandre Boulch. Convpoint: Continuous convolutions for point cloud processing. *Computers & Graphics*, 88:24–34, 2020.

[3] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In *Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5–7, 2020, Proceedings, Part II 15*, pages 207–222. Springer, 2020.

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[5] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364, 2020.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Qingyong Hu, Bo Yang, Guangchi Fang, Yulan Guo, Ales Leonardis, Niki Trigoni, and Andrew Markham. Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds. In *European Conference on Computer Vision*, 2022.

[8] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[9] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[10] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised lidar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21705–21715, 2023.

[11] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4558–4567, 2018.

[12] Mengtian Li, Yuan Xie, Yunhang Shen, Bo Ke, Ruizhi Qiao, Bo Ren, Shaohui Lin, and Lizhuang Ma. Hybridcr: Weakly-supervised 3d point cloud semantic segmentation via hybrid contrastive regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14930–14939, 2022.

[13] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018.

[14] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[16] Minghua Liu, Yin Zhou, Charles R Qi, Boqing Gong, Hao Su, and Dragomir Anguelov. Less: Label-efficient semantic segmentation for lidar point clouds. In *European Conference on Computer Vision*, pages 70–89. Springer, 2022.

[17] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[18] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, GA, 2013.

[19] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4213–4220. IEEE, 2019.

[20] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[21] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[23] Radu Alexandru Rosu, Peer Schütt, Jan Quenzel, and Sven Behnke. Latticenet: Fast point cloud segmentation using permutohedral lattices. *arXiv preprint arXiv:1912.05905*, 2019.

[24] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. *Advances in Neural Information Processing Systems*, 32, 2019.

[25] Ziyang Song and Bo Yang. Ogc: Unsupervised 3d object segmentation from rigid dynamics of point clouds. *Advances in Neural Information Processing Systems*, 35:30798–30812, 2022.

[26] Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. Splatnet: Sparse lattice networks for point cloud processing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2530–2539, 2018.

[27] An Tao, Yueqi Duan, Yi Wei, Jiwen Lu, and Jie Zhou. Seggroup: Seg-level supervision for 3d instance and semantic segmentation. *IEEE Transactions on Image Processing*, 31:4952–4965, 2022.

[28] Maxim Tatarchenko, Jaesik Park, Vladlen Koltun, and Qian-Yi Zhou. Tangent convolutions for dense prediction in 3d. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3887–3896, 2018.

[29] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019.

[30] Ozan Unal, Dengxin Dai, and Luc Van Gool. Scribble-supervised lidar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2697–2707, 2022.

[31] Jiacheng Wei, Guosheng Lin, Kim-Hui Yap, Tzu-Yi Hung, and Lihua Xie. Multi-path region mining for weakly supervised 3d semantic segmentation on point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4384–4393, 2020.

[32] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *2019 international conference on robotics and automation (ICRA)*, pages 4376–4382. IEEE, 2019.

[33] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 9621–9630, 2019.

[34] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 1–19. Springer, 2020.

[35] Xun Xu and Gim Hee Lee. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13706–13715, 2020.

[36] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9601–9610, 2020.

[37] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014.

[38] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[39] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9939–9948, 2021.

[40] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16280–16290, 2021.