面向自动驾驶的三维自适应池化点云分割网络

王宗顺[1]，李策[1,2]，王鹏程[1]，肖利梅[1]，张栋[3,4]，马佳林[3,4,5]
（1. 兰州理工大学自动化与电气工程学院，730050，兰州；2. 兰州理工大学微电子现代产业学院，730050，兰州；3. 人机混合增强智能全国重点实验室，710049，西安；4. 西安交通大学人工智能与机器人研究所，710049，西安；5. 西安交通大学第二附属医院超声医学科，710299，西安）

摘要：针对自动驾驶场景中车载激光雷达点云规模大、分布稀疏不均，现有点云Transformer网络存在邻域搜索开销大、体素注意力计算量高及关键目标局部几何细节易丢失的问题，提出了一种三维自适应池化点云分割网络（3DAPT－Net）。首先，在非重叠体素窗口内利用三维自适应池化压缩键值序列，聚合多尺度上下文，以降低大规模点云的注意力计算开销；其次，采用点注意力分支聚合点级特征，补充道路边界、行人轮廓及远距离稀疏目标的细粒度几何信息；最后，融合点级特征与体素特征，形成兼顾精度与效率的点云分割网络。该网络在SemanticKITTI和ShapeNet数据集上的平均交并比分别为67.5%和87.4%；与Point Cloud Transformer网络相比，推理延迟降低54.9%。结果表明，3DAPT－Net网络在分割精度与计算效率之间取得了较好的平衡。

关键词：自动驾驶；点云分割；注意力机制；自适应池化
中图分类号：TP391.4 **文献标志码**：A
DOI：10.7652/xjtuxbXXXXXX001 **文章编号**：0253-987X（XXXX）XX-0001-11

3D Adaptive Pooling Point Cloud Segmentation Network for Autonomous Driving

WANG Zongshun[1], LI Ce[1,2], WANG Pengcheng[1], XIAO Limei[1], ZHANG Dong[3,4], MA Jialin[3,4,5]
(1. School of Automation and Electrical Engineering, Lanzhou University of Technology, Lanzhou 730050, China; 2. School of Microelectronics and Modern Industry, Lanzhou University of Technology, Lanzhou 730050, China; 3. State Key Laboratory of Human－Machine Hybrid Augmented Intelligence, Xi'an 710049, China; 4. Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China; 5. Department of Ultrasound, the Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: In autonomous driving scenarios, vehicle-mounted LiDAR point clouds are characterized by large scale and sparse, uneven distribution. To address the problems of high neighborhood search overhead, heavy voxel attention computation load, and easy loss of local geometric details of key objects in existing point cloud Transformer networks, a 3D adaptive pooling point cloud segmentation network (3DAPT-Net) is proposed. First, 3D adaptive pooling is utilized within non-overlapping voxel windows to compress key-value sequences and aggregate multiscale context, thereby reducing the attention computation overhead for large-scale point clouds. Second, a point attention branch is adopted to aggregate point-level features, supplementing the fine-grained geometric information of road boundaries, pedestrian contours, and distant sparse objects. Finally, point-level features and voxel features are fused to construct a point cloud segmentation network that balances both accuracy

收稿日期：2026-04-22。
网络出版日期：XXXX-XX-XX 网络出版地址：

and efficiency. The proposed network achieves mean intersection over union (mIoU) scores of 67.5% and 87.4% on the SemanticKITTI and ShapeNet datasets, respectively; compared with the Point Cloud Transformer network, it has the inference latency reduced by 54.9%. The results indicate that 3DAPT-Net achieves a favorable balance between segmentation accuracy and computational efficiency.

Keywords: autonomous driving; point cloud segmentation; attention mechanism; adaptive pooling

随着智能交通系统的快速发展,自动驾驶车辆的安全运行对环境感知的准确性和实时性提出了更高要求。然而,真实道路环境中交通参与者类型多、目标尺度和空间分布差异大,远距离目标信息不完整,给场景理解带来了较大困难[1]。与二维图像相比,车载激光雷达点云能够提供准确的三维空间结构和距离信息,在复杂光照、弱纹理和远距离目标感知中具有重要优势[2]。因而成为自动驾驶环境感知的重要传感器。为了使自动驾驶车辆理解激光雷达采集的三维场景,需要通过点云语义分割为每个三维点赋予语义类别,从而区分道路、车辆、行人、建筑物和植被等场景元素。但车载点云规模大且分布稀疏不均,难以满足实时处理要求,同时道路边界、行人和远距离小目标的几何信息容易在体素化过程中丢失[3-4]。因此,如何兼顾大规模点云的计算效率与细粒度几何表达,是自动驾驶环境感知亟待解决的问题。

针对上述问题,研究人员提出了基于点和基于体素的点云Transformer网络。Transformer[5]利用自注意力机制自适应聚合输入序列信息,适用于无序点云的特征建模,基于点的Transformer网络通常结合集合抽象(set abstraction, SA)模块[6-7]提取点云特征,但在自动驾驶场景中,SA过程依赖最远点采样(farthest point sampling, FPS)[8]和K近邻(k-nearest neighbors, KNN)[9]等操作,计算效率受到限制。相比之下,体素表示具有规则结构,便于并行计算和内存访问,更适合大规模点云处理。然而,Transformer直接作用于体素序列时,多头自注意力(multi-head self-attention, MHSA)的计算成本仍受序列长度制约;同时,体素化和降采样造成局部几何细节损失,影响道路、车辆和行人等关键类别的识别。

为了解决这一挑战,提出了一种面向自动驾驶点云分割的3D自适应池化Transformer网络(3DAPT-Net)。该网络融合点表示与体素表示的互补优势,通过3D自适应池化Transformer(3D APTFormer)模块在非重叠体素窗口内提取粗粒度体素特征,并利用点注意力模块逐点聚合细粒度特征。其中,3D APTFormer通过3D自适应池化注意力(3D adaptive pooling attention, 3D APAtten)将输入体素特征重塑到三维空间,并在不同尺度上进行自适应平均池化,获得多尺度池化特征序列。该序列长度小于原始体素序列,能够有效降低注意力计算复杂度,同时保留自动驾驶场景中的空间结构信息。

同时,点注意力模块用于弥补体素化带来的局部细节损失。该模块避免复杂的KNN和FPS操作,通过逐点注意力聚合点级上下文信息,保持高分辨率几何表达。点分支与体素分支融合后,3DAPT-Net网络兼顾计算效率与细节建模,更适用于自动驾驶大规模点云语义分割任务。在SemanticKITTI[10]和ShapeNet[11]数据集上进一步验证3DAPT-Net网络的泛化能力。实验结果表明,3DAPT-Net网络在SemanticKITTI[10]数据集上的平均交并比达到67.5%,同时ShapeNet[11]数据集上取得87.4%的性能相对于Point Cloud Transformer网络[12]推理速度提升54.9%。

1 相关研究

1.1 稀疏体素表征

在基于体素的方法中,点云通常被投影到体素网格,并利用3D卷积进行特征提取[13-15]。然而,自动驾驶场景点云规模庞大,高分辨率体素虽能保留更多细节,却会显著增加计算和存储开销,导致处理效率受限。为缓解这一问题,近期方法尝试引入条件去噪扩散模型进行点云上采样、采用渐进式网格变形实现自监督致密化,或通过高效序列化邻域映射替代耗时的KNN搜索,以降低时间和内存成本[16-18]。此外,体素化过程不可避免地造成信息损失。Wang等人提出基于3D体素树状结构的方法,仅对非空体素进行操作[19],虽然降低了部分冗余计算,但也增强了数据稀疏性,影响模型效率与表达能力。相比之下,3D自适应池化Transformer处理局部体素,通过多尺度自适应池化压缩输入序列长

度并捕获多尺度上下文信息，同时保持输入点云的空间几何结构，避免特征投影带来的空间信息丢失。

1.2　细粒度点特征学习

基于点云的方法直接以原始点集为输入，通过学习点间几何关系提取三维空间特征。PointNet[6]是首个直接在点云上学习的神经网络，它使用多层感知机（multilayer perceptron，MLP）提取点云特征，并取得了良好的性能。PointNet++通过获取不同层级的点云特征进一步提高了性能。此外，一些模型使用邻域特征池化[20-23]和图信息传播[24]等方法来提取点云特征。PointNeXt[25]认为改进网络训练策略可极大地提高网络性能。这些模型在处理点云的稀疏性和不规则性方面表现出了卓越的性能[25]。

近年来，Transformer[26]在自然语言处理领域的成功推动了3D点云处理的发展[27-28]。一些研究人员开始探索基于Transformer的点云学习架构，例如Point Transformer[29]模型引入点积自注意力机制和局部向量注意力来提取全局特征。Zhang等人提出了Patch Attention，通过自动学习较小的基集合来计算注意力特征，以解决自注意力在输入点特征上生成过大注意力图的问题[30]。由于点云的稀疏性和不规则性，当前方法在直接处理点云上取得了良好的性能，但该方法的数据构建成本已成为计算负担，尤其是在大规模点云数据集上。因此，设计高效的基于Transformer的点云分析架构已成为重要的研究方向。

1.3　多表示协同融合

与其他方法相比[22,31-33]，聚合点云和体素特征的方法结合了基于体素和基于点的方法的优势，使其能够提取深层特征，并以低时间和GPU内存消耗处理各类3D任务。PVCNN[15]和PVT[34]均利用点表示与体素表示的互补优势，在体素分支中提取局部结构特征，并在后续阶段与点级特征进行融合。基于PV-RCNN[35]，PV-RCNN++[36]通过中心扇区采样聚合模块实现了更快的处理速度。与现有的融合点和体素特征的方法相比，本文在体素化结构数据上提出了一种Transformer，其核心思想是通过3D窗口内的自适应池化压缩Transformer的输入序列长度，以减少注意力计算的复杂度。

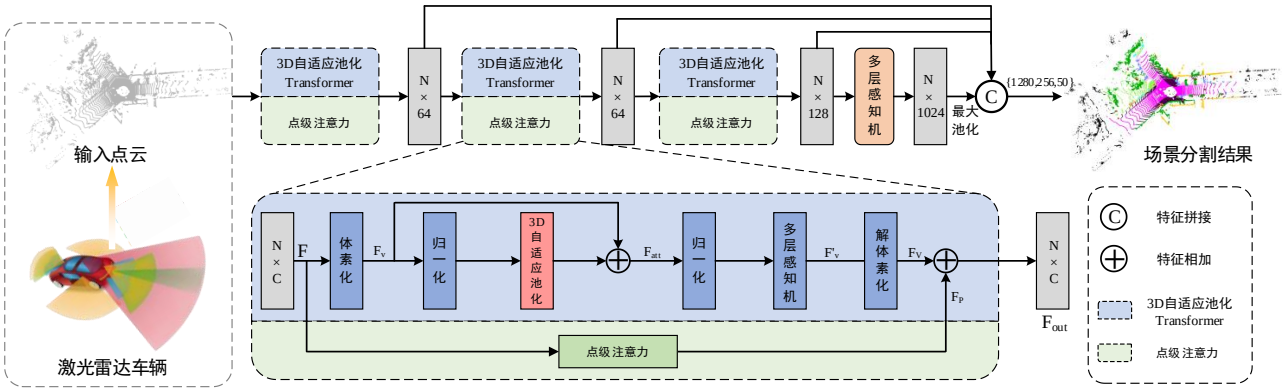

图1　所提3DAPT-Net整体网络框架图

Fig. 1　Overall network framework of the proposed 3DAPT-Net

2　3DAPT-Net网络

3DAPT-Net网络的整体结构如图1所示。该网络以原始点云为输入，编码阶段由3个特征提取模块组成，每个模块均包含3D自适应池化Transformer（3D APTFormer）和点注意力两个分支。其中，3D APTFormer分支在非重叠体素窗口内聚合多尺度体素上下文，点注意力分支直接提取点级特征，以补充体素化过程中损失的细粒度几何信息。随后，将两个分支的输出映射至相同的点级空间和通道维度，并通过逐元素相加进行融合。不同编码阶段的融合特征通过残差连接传递至解码端，以综合利用不同层级的局部特征和上下文信息。最后，解码器将融合特征映射至原始点云，预测每个点的语义类别。

2.1　3D自适应池化Transformer

多尺度池化能够在不同空间尺度上聚合局部特征，为点云场景理解提供多尺度上下文信息。然而，原始三维点云具有无序、稀疏和非均匀分布特性，难以直接进行规则的多尺度池化。为此，3D

APTFormer模块首先将点特征映射至规则体素空间，然后在非重叠体素窗口内引入3D自适应池化注意力，以压缩键和值的特征序列，并降低体素自注意力的计算量。

图1下方蓝色虚线框展示了3D APTFormer模块的结构。首先，对输入点特征进行体素化，得到体素特征$\boldsymbol{F}_{\mathrm{e}}$，其次，利用3D自适应池化注意力提取多尺度体素上下文，得到注意力输出$\boldsymbol{A}_{\mathrm{APA}}$，并将其与原始体素特征进行残差融合；最后，对融合特征依次进行归一化和特征编码，得到3D APTFormer模块的输出体素特征。上述过程表示如下：

$$\boldsymbol{F}_{\mathrm{a}}=\boldsymbol{F}_{\mathrm{e}}+\boldsymbol{A}_{\mathrm{APA}} \tag{1}$$

$$\boldsymbol{F}'_{\mathrm{v}}=\boldsymbol{E}[\boldsymbol{L}(\boldsymbol{F}_{\mathrm{a}})]. \tag{2}$$

式中，$\boldsymbol{F}_{\mathrm{e}}$为输入体素特征；$\boldsymbol{A}_{\mathrm{APA}}$为3D自适应池化注意力输出；$\boldsymbol{F}_{\mathrm{a}}$为注意力输出与原始体素特征进行残差融合后得到的特征；$\boldsymbol{L}(\cdot)$表示归一化操作；$\boldsymbol{E}(\cdot)$表示特征编码；$\boldsymbol{F}'_{\mathrm{v}}$为3D APTFormer模块的输出体素特征。

2.2 3D自适应池化注意力

图2给出了3D自适应池化注意力模块的输入序列构建过程。3D自适应池化注意力模块的输入通过对原始体素输入序列进行池化，并提取多尺度粗粒度体素特征，同时保留了底层3D结构。对于体素窗口内的输入特征为$\boldsymbol{F}_{\mathrm{v}}\in\mathbf{R}^{N_{\mathrm{v}}\times C}$ N_{V}表示体素窗口内的体素标记数，C表示特征通道数。首先，将输入体素特征重塑为$C\times D\times H\times W$的三维空间，$D$、$H$、$W$分别表示体素窗口的深度、高度和宽度，且满足$N_{\mathrm{V}}=D\times H\times W$然后，在重塑后的体素特征上执行不同输出尺度的3D自适应平均池化，并将每个尺度的池化结果转换为特征序列。第i个尺度的池化特征表示如下：

$$\boldsymbol{F}_{\mathrm{s}}^{i}=\mathrm{P}_{i}(\boldsymbol{F}_{\mathrm{v}}),\qquad i=1,2,...,n \tag{3}$$

式中，$\mathrm{P}_{i}(\cdot)$表示第i个尺度的3D自适应平均池化及序列重排操作；$\boldsymbol{F}_{\mathrm{s}}^{i}$表示第$i$个尺度的池化特征；$n$表示池化尺度数。设第$i$个尺度的池化输出尺寸为$D_i\times H_i\times W_i$，则该尺度池化特征的序列长度为$D_iH_iW_i$。将不同尺度的池化特征进行拼接和归一化，得到多尺度池化特征序列：

$$\boldsymbol{F}_{\mathrm{m}}=\mathrm{G}(\boldsymbol{F}_{\mathrm{s}}^{1},\boldsymbol{F}_{\mathrm{s}}^{2},...,\boldsymbol{F}_{\mathrm{s}}^{n}) \tag{4}$$

式中，$\mathrm{G}(\cdot)$表示多尺度特征拼接与归一化操作；$\boldsymbol{F}_{\mathrm{m}}$表示多尺度池化特征序列。该序列在不同空间尺度上聚合窗口内的体素响应，其序列长度小于原始体素特征的序列长度。

以原始体素特征$\boldsymbol{F}_{\mathrm{v}}$生成查询矩阵，以多尺度池化特征$\boldsymbol{F}_{\mathrm{m}}$生成键矩阵和值矩阵，表示如下：

$$\begin{cases}\boldsymbol{Q}=\boldsymbol{F}_{\mathrm{v}}\boldsymbol{\Theta}_{\mathrm{q}}\\ \boldsymbol{K}=\boldsymbol{F}_{\mathrm{m}}\boldsymbol{\Theta}_{\mathrm{k}}\\ \boldsymbol{V}=\boldsymbol{F}_{\mathrm{m}}\boldsymbol{\Theta}_{\mathrm{v}}\end{cases} \tag{5}$$

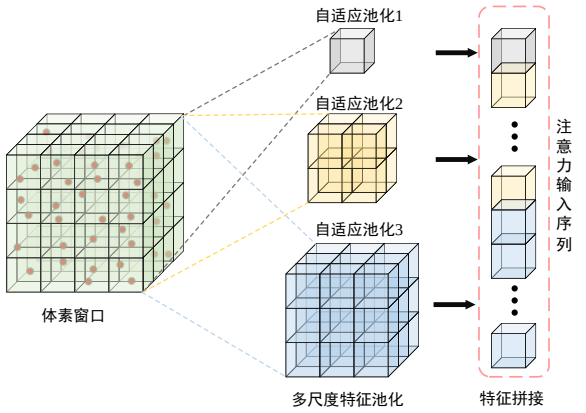

图2 3D自适应池化注意力的多尺度特征序列构建过程

Fig. 2 Construction of the multiscale feature sequence for 3D adaptive pooling attention

式中，$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$分别表示查询、键和值矩阵；$\boldsymbol{\Theta}_{\mathrm{q}}$、$\boldsymbol{\Theta}_{\mathrm{k}}$和$\boldsymbol{\Theta}_{\mathrm{v}}$分别表示查询、键和值的投影矩阵。采用$\boldsymbol{\Theta}$表示投影矩阵，3D自适应池化注意力输出表示如下：

$$\boldsymbol{A}_{\mathrm{APA}}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\top}}{\sqrt{d_{\mathrm{k}}}}\right)\boldsymbol{V} \tag{6}$$

式中，Softmax表示对相似度矩阵的每一行进行归一化；d_{k}表示查询矩阵和键矩阵的通道维度；$\boldsymbol{A}_{\mathrm{APA}}$表示3D自适应池化注意力输出。

与直接使用原始体素特征生成键和值相比，$\boldsymbol{F}_{\mathrm{m}}$同时包含多个空间尺度的上下文信息，并具有较短的序列长度。因此，3D自适应池化注意力能够减少窗口内冗余的体素标记交互，同时增强体素特征的多尺度上下文表达。

2.3　计算复杂度与有效性分析

将点云映射为一系列均匀排列且互不重叠的立方体体素窗口。设体素窗口的边长为B，则每个窗口内包含B^3个体素标记，体素特征的通道数为C。传统3D窗口自注意力需要计算窗口内全部体素标记之间的两两相关性，因此其主要计算复杂度为$O_1=O(B^6C)$。

提出的3D自适应池化注意力首先对窗口内体素特征进行多尺度3D自适应平均池化，并将不同尺度的池化特征拼接为池化序列$\boldsymbol{F}_{\mathrm{m}}$。设第$i$个池化尺度的输出尺寸为$H_i\times W_i\times D_i$，共有n个池化尺度，则池化序列长度表示如下：

$$N_{\mathrm{m}}=\sum_{i=1}^{n}H_iW_iD_i \tag{7}$$

式中，N_{m}表示多尺度池化特征序列长度；H_i、W_i、D_i分别表示第i个池化尺度输出特征的深度、高度和宽度。3D自适应池化注意力以窗口内的B^3个原始体素特征生成查询矩阵，以长度为N_{m}的多尺度池化特征生成键矩阵和值矩阵。因此3D自适应池化注意力的计算复杂度分别表示如下：

$$O_2=O(B^3N_{\mathrm{m}}C) \tag{8}$$

与O_1相比，O_2将窗口内原始体素标记之间的两两交互转化为原始标记与多尺度池化标记之间的交互，从而在保留原始体素空间位置查询能力的同时，降低了注意力计算量。

此外，多尺度自适应平均池化并非直接丢弃局部窗口内的体素响应，而是在不同空间尺度上对体素邻域进行结构化聚合，形成多尺度上下文摘要。因此，该模块能够在降低计算复杂度的同时，缓解直接下采样或稀疏化操作造成的空间几何结构损失。图3比较了窗口尺寸B从4到0时O_1与O_2的复杂度变化。可以看出，随着窗口尺寸增大，O_1快速上升，而O_2增长更加缓慢，说明3D自适应池化注意力在较大窗口下具有更好的计算效率。

2.4　点注意力

由于上述体素特征学习网络均基于融合局部窗口中的粗粒度邻域信息，这可能导致模型缺乏全局感受野。为了在低分辨率体素分支中捕获长程依赖关系，所提方法使用自注意力对整个点云进行全局上下文聚合，即计算每个点与所有其他点之间的注意力，以提取点级特征。自注意力公式表示

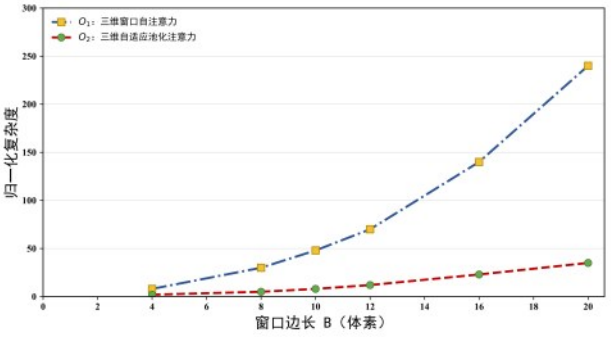

图3　O_1与O_2在不同窗口尺寸下的复杂度比较

Fig. 3　Complexity comparison of O_1 and O_2 under different window sizes

如下：

$$\boldsymbol{F}_{\mathrm{p}}=\boldsymbol{F}+\phi_{\mathrm{m}}\left[\mathrm{Softmax}\left(\frac{\boldsymbol{Q}_{\mathrm{p}}\boldsymbol{K}_{\mathrm{p}}^{\mathrm{T}}}{\sqrt{d_{\mathrm{p}}}}\right)\boldsymbol{V}_{\mathrm{p}}\right] \tag{9}$$

式中，$\boldsymbol{F}_{\mathrm{p}}$表示点注意力模块输出的点级特征；$d_{\mathrm{p}}$表示点分支查询矩阵和键矩阵的通道维度；$\phi_{\mathrm{m}}(\cdot)$表示特征映射；“+”表示残差连接；$Q_{\mathrm{p}}$、$K_{\mathrm{p}}$、$V_{\mathrm{p}}$由输入点特征$\boldsymbol{F}$通过独立线性映射生成点分支的查询矩阵、键矩阵和值矩阵。

2.5　特征融合

输入点云包含N_{p}个点，3D APTFormer输出的体素特征为$\boldsymbol{F}'_{\mathrm{v}}\in\mathbf{R}^{N_{\mathrm{v}}\times C_{\mathrm{v}}}$，点注意力模块输出的点级特征为$\boldsymbol{F}_{\mathrm{p}}\in\mathbf{R}^{N_{\mathrm{p}}\times C_{\mathrm{p}}}$。$N_{\mathrm{v}}$表示体素标记数；$C_{\mathrm{v}}$和$C_{\mathrm{p}}$分别表示体素特征与点特征的通道数。首先，根据点—体素索引映射，将体素特征解体素化至点级空间，表示如下：

$$\boldsymbol{F}_{\mathrm{d}}=\delta(\boldsymbol{F}'_{\mathrm{v}},\boldsymbol{I}_{\mathrm{pv}}),\qquad \boldsymbol{F}_{\mathrm{d}}\in\mathbf{R}^{N_{\mathrm{p}}\times C_{\mathrm{v}}} \tag{10}$$

式中，$\delta(\cdot)$表示解体素化操作；$\boldsymbol{I}_{\mathrm{pv}}$表示点—体素索引映射；$\boldsymbol{F}_{\mathrm{d}}$表示映射至点级空间的体素特征。

随后，将点级体素特征与点注意力特征映射至相同通道维度，并通过逐元素相加进行融合：

$$\boldsymbol{F}_{\mathrm{o}}=\phi_{\mathrm{v}}(\boldsymbol{F}_{\mathrm{d}})+\phi_{\mathrm{p}}(\boldsymbol{F}_{\mathrm{p}}),\qquad \boldsymbol{F}_{\mathrm{o}}\in\mathbf{R}^{N_{\mathrm{p}}\times C_{\mathrm{o}}} \tag{11}$$

式中，$\phi_{\mathrm{v}}(\cdot)$和$\phi_{\mathrm{p}}(\cdot)$分别表示体素特征与点特征的通道映射；C_{o}表示融合特征的通道数；“+”表示在相同点级空间和相同通道维度上逐通道相加；$\boldsymbol{F}_{\mathrm{o}}$表示点—体素融合特征。

通过上述融合，体素分支提供多尺度空间上下

文信息,点分支补充细粒度点级几何信息。融合特征随后被传递至下一编码阶段或输入解码器,用于预测原始点云中每个点的语义类别。

3 实 验

本节对所提出的3DAPT-Net网络进行评估,以验证模型的有效性和泛化能力。实验首先在SemanticKITTI[10]大规模室外点云数据集上评估模型面向自动驾驶场景的语义分割性能;其次在ShapeNet[11]部件分割数据集上进一步验证其在不同分割任务中的适用性,并通过消融实验分析关键模块对模型性能的贡献。

3.1 实现细节

所提方法基于PyTorch深度学习框架实现。本文在RTX 2080 GPU上测试模型推理延迟,并在4块RTX 3090 GPU上评估其他性能指标。体素大小设置为4,通过后续的消融实验确定自适应池化窗口的大小为12。使用Adam优化器来优化网络,其中$\beta_1=0.9$,$\beta_2=0.999$,学习率设置为1×10^{-4}。在设置中,为了降低过拟合风险并提高模型的泛化能力,在模型训练期间设置了0.5的Dropout率。在训练过程中,使用交叉熵损失函数,并将批大小设置为16。为了训练模型,进行了200个轮次的迭代。

3.2 面向自动驾驶点云场景语义分割实验

SemanticKITTI[10]数据集是自动驾驶大规模三维

点云数据集,该数据集包含21个序列,共计43 552帧密集标注的激光雷达扫描数据。每帧包含约10^5个点,空间范围达160 m×160 m×20 m。该数据集的原始三维点仅提供空间坐标信息,无颜色数据。在数据划分上,序列00至07以及09至10用于模型训练,序列08用于验证,序列11至21用于在线测试评估。

模型训练初始学习率设为0.01,每个epoch结束后衰减5%。所提方法以0.06 m为体素网格大小对每帧原始点云进行离线网格下采样,并为下采样点云建立空间索引,以支持后续快速邻域查询。训练和验证阶段,从单帧点云中裁取固定规模局部点集作为网络输入,每个样本包含45 056个点。测试阶段采用空间规则采样覆盖完整单帧点云,并将预测结果投影回原始点云,最终获得完整LiDAR帧的逐点语义标签。

为了验证所提网络在大规模室外场景下的语义分割能力,如表1所示,本文在SemanticKITTI[10]数据集上进行了全面评估。所提方法在分割精度与推理效率上均有所提升。在精度方面,本文模型的平均交并比达到67.5%,高于PointNet++[37]和RandLA-Net[39]等经典基于点的网络,且优于PVCNN[15]等体素或混合架构基准模型。该表现主要得益于3D APTFormer与点注意力模块的有效结合,使得网络在提取非重叠窗口内粗粒度上下文信息的同时,保留了细粒度的空间几何特征,增强了模型在复杂室外场景下的鲁棒性。

表1 模型在SemanticKITTI[10]上室外场景分割结果

Table 1 Outdoor Scene Segmentation Results of the Model on SemanticKITTI[10]

网络	类型	平均交并比/%	延迟/ms
PointNet++[37]	P	20.1	—
RangeNet++[38]	RV	52.2	830
RandLA[39]	P	53.9	550
SqueezeSegV3[40]	RV	55.9	—
KPConv[41]	P	58.8	—
SalsaNext[42]	RV	59.5	—
DRINet[43]	PV	65.5	—
Cylinder3D[44]	V	65.9	625
GASN[45]	V	66.0	—
RPVNet[46]	RPV	66.2	—
PVCNN[15]	PV	67.1	525
本 文	**PV**	**67.5**	**236**

注:"P"、"R"、"V"分别代表模型输入数据分别为点、距离图、体素。

此外,本文在图4中展示了不同模型在SemanticKITTI[10]数据集上的自动驾驶点云场景下的语义分割可视化结果,在"Ours vs. PVCNN"中进一步可视化了所提模型对比与PVCNN[15]更好的自动驾驶语义标签预测,并且使用绿色点进行标记。

3.3 点云部件分割实验

ShapeNet[11]是一个的点云部件分割数据集,包含50多个物体类别和数十万个3D模型。每个模型被划分为若干部件,并且每个部件都标记有唯一的ID。用于3D场景理解。ShapeNet[11]数据集具有高质量的3D模型、类别和部件注释、大规模数据以及多样的形状和结构等特点。使ShapeNet[11]用于点云分割、形状识别、形状生成等领域的研究。在分割模型训练中,从每个训练形状中采样2 048个点

作为输入。为了评估模型的性能，使用平均交并比指标，该指标计算 2 874 个测试模型每个部件的交并比并取平均值作为最终指标。在训练期间，将批大小设置为 16，并执行 200 个轮次的迭代。表 2 中，本文模型在平均交比指标方面优于 PVT[34]模型。PVT 在局部出窗

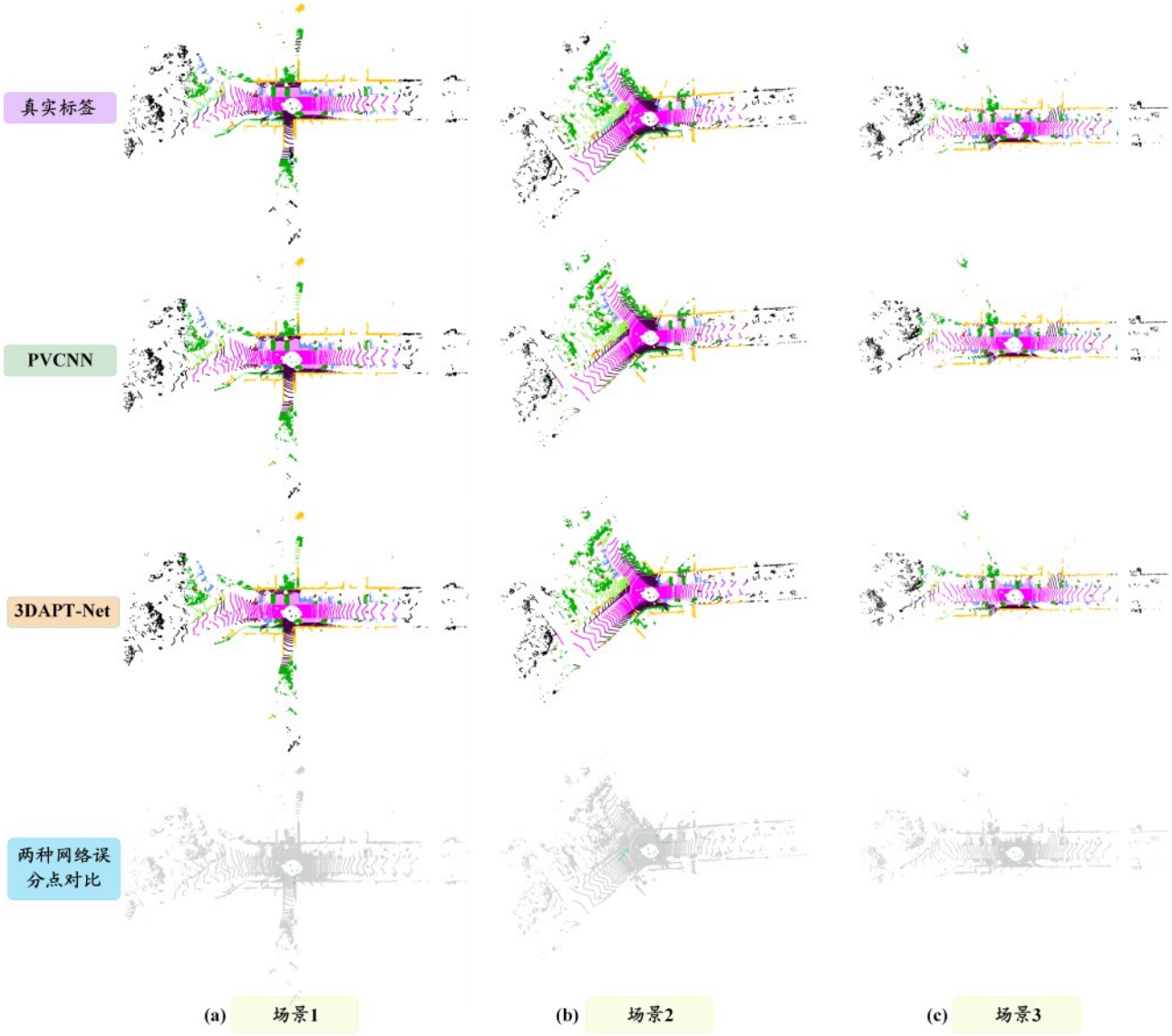

图 4　3DAPT-Net 在 SemanticKITTI[10]数据集上室外自动驾驶场景语义分割可视化结果

Fig. 4 Visualization results of semantic segmentation for outdoor autonomous driving scenes on the SemanticKITTI[10] dataset using 3DAPT-Net

口内对原始体素标记进行注意力计算，而 3D 自适应池化注意力首先生成多尺度池化序列，再利用该序列生成键和值，减少了窗口内冗余标记交互，并增强了多尺度上下文表达。

3.4　消融实验

为了理解模型中各设计决策的影响，本文对所提出的网络进行了全面的消融实验。实验评估了各种实验配置在 ShapeNet[11]部件分割数据集上的性能。

（1）　关于窗口尺寸选择的消融实验

为了确定自适应池化注意力的合适 3D 窗口尺寸，进行了消融实验，结果如表 3 所示。对于输入特征，将自适应池化比率固定为{2，3，4}，以评估不同窗口尺寸对模型的影响。表 3 中第 3 列显示了每个窗口尺寸在自适应池化后对应的 3 个输出尺寸。最终将窗口尺寸确定为 12，因为发现更大的窗口大小对模型性能的收益递减。

（2）　关于池化操作选择的消融实验

为了分析池化操作对模型性能的影响，本文在

表 2 模型在 ShapeNet[11] 数据集上的部件分割结果

Table 2 Part Segmentation Results of the Model on the ShapeNet[11] Dataset

网络	类型	平均交并比/%	延迟/ms
PointNet[6]	P	83.7	21.7
PointCNN[47]	P	86.1	145.6
PVCNN[15]	PV	86.1	90.7
PointASNL[48]	P	86.1	1023.2
PointMLP[49]	P	86.1	—
PAConv[50]	PV	86.3	—
PCT[12]	P	86.4	128.2
PVT[34]	PV	86.5	75.9
PointMT[51]	P	86.6	560.2
本 文	**PV**	**87.4**	**57.8**

注:“P”、“V”分别代表模型输入数据分别为点、体素。

保持其他参数不变的条件下,对最大池化和平均池化两种方式进行了对比。实验结果表明,采用最大池化时,模型的平均交并比为 86.2%;采用平均池化时,模型的平均交并比为 87.4%,较最大池化提升 1.2个百分点。由于两种池化方式不改变自适应池化后的序列长度,因此二者在计算复杂度上基本一致。综合分割精度表现,本文在模型的训练和推理中采用平均池化作为 3D 自适应池化注意力模块中的池化方式。

表 3 关于窗口尺寸选择的消融实验

Table 3 Ablation study on window size selection

编号	窗口尺寸/体素单元	输出尺寸/体素单元	平均交并比/%
1	4	2,1,1	85.2
2	8	3,2,1	86.5
3	10	4,3,2	87.2
4	**12**	**6,4,3**	**87.4**
5	20	10,6,4	86.1

注:窗口尺寸以体素网格单元计;多尺度池化输出尺寸表示3个池化分支输出特征图的空间分辨率。

(3) 关于不同阶段自适应池化影响的消融实验

为了验证多尺度自适应池化模块在网络不同阶段的作用,本文进行了消融实验,结果如表 4 所示。去除多尺度自适应池化的阶段,将其替换为单尺度池化操作,默认池化比率为3。

在网络的第1阶段使用多尺度自适应池化,而在第2和第3阶段使用单尺度池化。在相同参数下,模型在 ShapeNet[11] 数据集上实现了 86.1% 的平均交并比。随着多尺度自适应池化应用于多个阶段,性能逐渐提高到87.4%。这表明多尺度自适应池化操作在网络的每个阶段都发挥了关键作用。

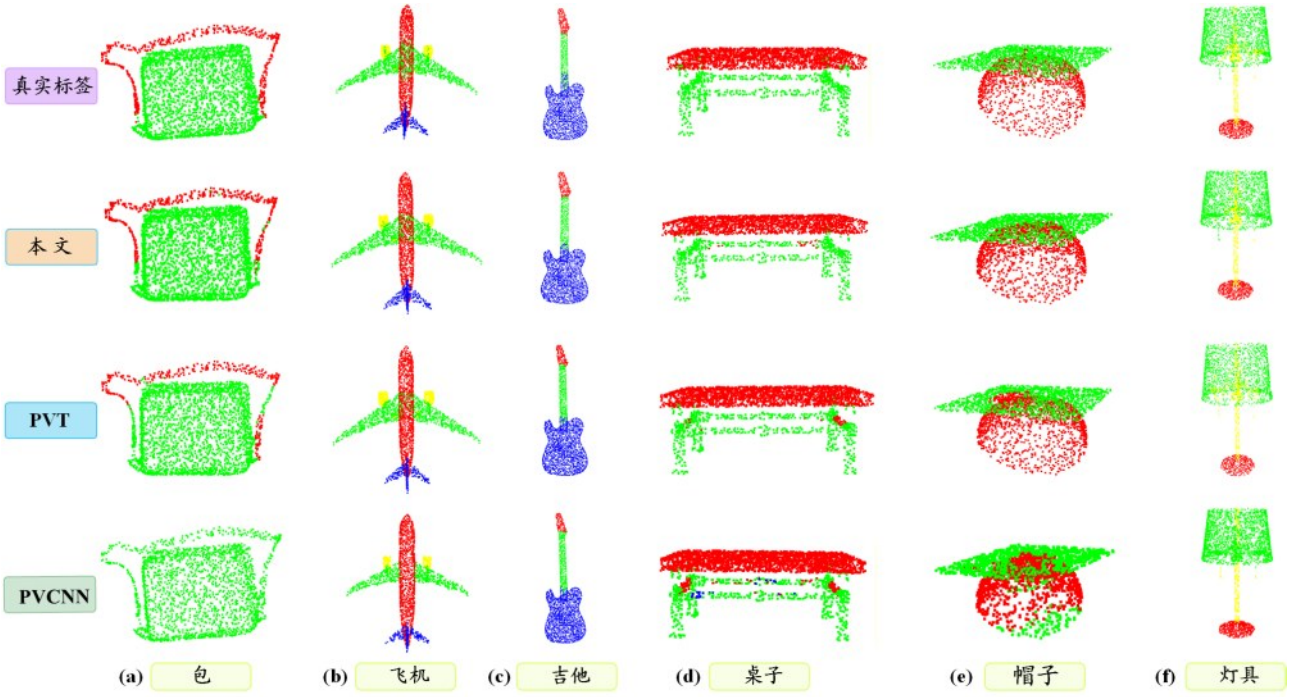

图5 3DAPT-Net在 ShapeNet[11] 数据集上的部分分割结果可视化

Fig. 5 Visualization of part segmentation results of 3DAPT-Net on the ShapeNet[11] dataset

表4　关于不同阶段自适应池化影响的消融实验

Table 4　Ablation study on the impact of adaptive pooling at different stages

编号	阶段1	阶段2	阶段3	平均交并比/%
1	✓			86.1
2	✓	✓		86.8
3	✓	✓	✓	**87.4**

4　结　论

本文围绕自动驾驶场景下大规模点云语义分割的高效建模问题,构建了融合点特征与体素特征的3DAPT-Net网络。通过在体素分支中引入自适应池化注意力机制,模型能够在较低计算开销下获取多尺度空间上下文;结合点注意力分支后,网络进一步增强了对局部几何细节的表达能力,从而提高了复杂交通场景中点云特征的判别性。在SemanticKITTI和ShapeNet数据集上的实验进一步表明,所提方法能够较好地平衡自动驾驶场景点云分割的精度与速度。后续工作将进一步结合多帧时序点云和轻量化部署策略,提升模型对远距离小目标、动态交通参与者和复杂道路边界的感知能力,为自动驾驶实时三维场景理解提供更稳定的算法支持。

参考文献:

[1] Betsas T, Georgopoulos A, Doulamis A, et al. Deep learning on 3D semantic segmentation: a detailed review[J]. Remote Sensing, 2025, 17(2): 298.

[2] Leng Wenfeng, Hu Chuan, Huang Yonghui, et al. KD-DiffSeg: knowledge distillation guided LiDAR-camera diffusion framework for 3D semantic segmentation[J]. Expert Systems with Applications, 2026, 322: 132272.

[3] 蔡子悦,袁振岳,庞明勇. 深度学习的点云语义分割方法综述[J]. 计算机工程与应用, 2025, 61(11): 22-30.
Cai Ziyue, Yuan Zhenyue, Pang Mingyong. Survey on deep-learning-based point cloud semantic segmentation [J]. Computer Engineering and Applications, 2025, 61(11): 22-30.

[4] Wang Zongshun, Li Ce, Ma Jialin, et al. Explicit geometric relationships under limited spatial reference points guide 3D visual grounding [J]. Information Processing & Management, 2026, 63(6): 104720.

[5] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2017: 6000-6010.

[6] Charles R Q, Su Hao, Kaichun Mo, et al. PointNet: deep learning on point Sets for 3D classification and segmentation [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 77-85.

[7] Li Wangkai, Li Zhaoyang, Pan Yuwen, et al. Adaptive augmentation-aware latent learning for robust LiDAR semantic segmentation[C]//ICLR 2026. New York, USA: ICLR, 2026: 1-36.

[8] Eldar Y, Lindenbaum M, Porat M, et al. The farthest point strategy for progressive image sampling [J]. IEEE Transactions on Image Processing, 1997, 6(9): 1305-1315.

[9] Cover Thomas M, Hart Peter E. Nearest neighbor pattern classification [J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.

[10] Behley J, Garbade M, Milioto A, et al. SemanticKITTI: a dataset for semantic scene understanding of LiDAR sequences[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2019: 9296-9306.

[11] Chang A X, Funkhouser T, Guibas L, et al. ShapeNet: an information-rich 3D model repository [PP/OL]. V1. arXiv (2015-12-09) [2026-04-22]. https://arxiv.org/abs/1512.03012.

[12] Guo Menghao, Cai Junxiong, Liu Zhengning, et al. PCT: point cloud transformer [J]. Computational Visual Media, 2021, 7(2): 187-199.

[13] Wang Haiyang, Shi Chen, Shi Shaoshuai, et al. DSVT: dynamic sparse voxel transformer with rotated sets[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2023: 13520-13529.

[14] Zhou Wei, Zhang Xiaodan, Hao Xingxing, et al. Multi point-voxel convolution (MPVConv) for deep learning on point clouds [J]. Computers & Graphics, 2023, 112: 72-80.

[15] Liu Zhijian, Tang Haotian, Lin Yujun, et al. Point-voxel CNN for efficient 3D deep learning [C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2019: 965-975.

[16] Zheng Xiao, Huang Xiaoshui, Mei Guofeng, et al.

Point cloud pre-training with diffusion models[C]// 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2024: 22935-22945.

[17] Qu Wentao, Shao Yuantian, Meng Lingwu, et al. A conditional denoising diffusion probabilistic model for point cloud upsampling [C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2024: 20786-20795.

[18] Liu Yanzhe, Chen Rong, Li Yushi, et al. SPU-PMD: self-supervised point cloud upsampling via progressive mesh deformation[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2024: 5188-5197.

[19] Wang Xuzhi, Feng Wei, Kong Lingdong, et al. NUC-Net: non-uniform cylindrical partition network for efficient LiDAR semantic segmentation [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2025, 35(9): 9090-9104.

[20] Zhang Jianhui, Luo Yizhi, Zhang Zicheng, et al. CamPoint: boosting point cloud segmentation with virtual camera[C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2025: 11822-11832.

[21] Dai Wenxia, Wu Wanmin, Hao Yuqi, et al. A topography-attentional network for ground filtering from ALS point cloud in forest environments[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2026, 19: 1842-1855.

[22] 鲁斌, 尚国栋. 基于多尺度感知和双模态融合的激光雷达点云语义分割[J]. 计算机应用研究, 2026, 43(5): 1322-1328.
Lu Bin, Shang Guodong. Multi-scale and dual-modality fusion for LiDAR point cloud semantic segmentation[J]. Application Research of Computers, 2026, 43(5): 1322-1328.

[23] 杨军, 王连甲. 结合位置关系卷积与深度残差网络的三维点云识别与分割[J]. 西安交通大学学报, 2023, 57(5): 182-193.
Yang Jun, Wang Lianjia. Recognition and segmentation of 3D point cloud through positional relation convolution in combination with deep residual network [J]. Journal of Xi′an Jiaotong University, 2023, 57(5): 182-193.

[24] Du Zijin, Liang Jianqing, Liang Jiye, et al. Graph regulation network for point cloud segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 7940-7955.

[25] Qian Guocheng, Li Yuchen, Peng Houwen, et al. PointNeXt: revisiting PointNet++ with improved training and scaling strategies[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2022: 23192-23204.

[26] Guang Jinzheng, Wu Shichao, Wang Yongru, et al. HTMNet: a hybrid transformer-mamba network for LiDAR-based 3D detection and semantic segmentation [J]. Expert Systems with Applications, 2026, 316: 131832.

[27] De Vries M, Naidoo R, Fourkioti O, et al. Interpretable point cloud classification using multiple instance learning[C]//2025 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2025: 22209-22220.

[28] Wang Zongshun, Li Ce, Feng Zhiqiang, et al. Rethinking static weights: language-guided adaptive weight adjustment for 3D visual grounding [J]. Knowledge-Based Systems, 2026, 338: 115467.

[29] Zhao Hengshuang, Jiang Li, Jia Jiaya, et al. Point transformer [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2021: 16239-16248.

[30] Zhang Cheng, Wan Haocheng, Shen Xinyi, et al. PatchFormer: an efficient point transformer with patch attention [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2022: 11789-11798.

[31] 唐友源, 张辉, 杜瑞, 等. 基于结构频谱感知框架的配电网点云语义分割[J]. 自动化学报, 2026, 52(4): 833-845.
Tang Youyuan, Zhang Hui, Du Rui, et al. Semantic segmentation of distribution network point clouds based on a structure spectrum-aware framework [J]. Acta Automatica Sinica, 2026, 52(4): 833-845.

[32] Zhang Weijian, Song Haichuan, Zhang Zhizhong, et al. from sparse semantics to rich instances: empowering label-efficient LiDAR panoptic segmentation via geometric priors [J]. Neural Networks, 2026, 200: 108767.

[33] 朱安迪, 达飞鹏, 盖绍彦. 对融合特征敏感的三维点云识别与分割[J]. 西安交通大学学报, 2024, 58(5): 52-63.
Zhu Andi, Da Feipeng, Ge Shaoyan. Recognition and segmentation of 3D point clouds sensitive to fusion

features [J]. Journal of Xi′an Jiaotong University, 2024, 58(5): 52-63.

[34] Zhang Cheng, Wan Haocheng, Liu Shengqiang, et al. Point-voxel transformer: an efficient approach to 3D deep learning [PP/OL]. V1. arXiv (2021-08-13) [2026-04-22]. https://arxiv.org/abs/2108.06076v1.

[35] Shi Shaoshuai, Guo Chaoxu, Jiang Li, et al. PV-RCNN: point-voxel feature set abstraction for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 10526-10535.

[36] Shi Shaoshuai, Jiang Li, Deng Jiajun, et al. PV-RCNN++: point-voxel feature set abstraction with local vector representation for 3D object detection[J]. International Journal of Computer Vision, 2023, 131(2): 531-551.

[37] Qi C R, Yi Li, Su Hao, 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. Red Hook, NY, USA: Curran Associates Inc., 2017: 5105-5114.

[38] Milioto A, Vizzo I, Behley J, et al. RangeNet ++: fast and accurate LiDAR semantic segmentation[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, NJ, USA: IEEE, 2019: 4213-4220.

[39] Hu Qingyong, Yang Bo, Xie Linhai, et al. RandLA-Net: efficient semantic segmentation of large-scale point clouds [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 11105-11114.

[40] Xu Chenfeng, Wu Bichen, Wang Zining, et al. SqueezeSegV3: spatially-adaptive convolution for efficient Point-Cloud segmentation [C]//Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 1-19.

[41] Thomas H, Qi C R, Deschaud J E, et al. KPConv: flexible and deformable convolution for point clouds [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2019: 6410-6419.

[42] Cortinhal T, Tzelepis G, Erdal Aksoy E. SalsaNext: fast, uncertainty-aware semantic segmentation of LiDAR point clouds [C]//Advances in Visual Computing. Cham: Springer International Publishing, 2020: 207-222.

[43] Ye Maosheng, Xu Shuangjie, Cao Tongyi, et al. DRINet: a dual-representation iterative learning network for point cloud segmentation[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2021: 7427-7436.

[44] Zhu Xinge, Zhou Hui, Wang Tai, et al. Cylindrical and asymmetrical 3D convolution networks for LiDAR segmentation [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2021: 9934-9943.

[45] Ye Maosheng, Wan Rui, Xu Shuangjie, et al. Efficient point cloud segmentation with geometry-aware sparse networks [C]//Computer Vision - ECCV 2022. Cham: Springer Nature Switzerland, 2022: 196-212.

[46] Xu Jianyun, Zhang Ruixiang, Dou Jian, et al. RPVNet: a deep and efficient range-point-voxel fusion network for LiDAR point cloud segmentation [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2021: 16004-16013.

[47] Li Yangyan, Bu Rui, Sun Mingchao, et al. PointCNN: convolution on X-transformed points[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2018: 828-838.

[48] Yan Xu, Zheng Chaoda, Li Zhen, et al. PointASNL: robust point clouds processing using nonlocal neural networks with adaptive sampling [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 5588-5597.

[49] Ma Xu, Qin Can, You Haoxuan, et al. Rethinking network design and local geometry in point cloud: a simple residual MLP framework [C]//ICLR 2022. New York, USA: ICLR, 2022: 1-15.

[50] Xu Mutian, Ding Runyu, Zhao Hengshuang, et al. PAConv: position adaptive convolution with dynamic kernel assembling on point clouds [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2021: 3172-3181.

[51] Zheng Qiang, Zhang Chao, Sun Jian. PointMT: efficient point cloud analysis with hybrid MLP-transformer architecture [J]. IEEE Transactions on Multimedia, 2025, 27: 6382-6396.

(编辑　曹康莹　杜秀杰)