摘要

激光雷达感知领域最近出现了一种趋势，即倾向于将多个任务统一到一个性能更优的强大网络中，而不是为每个任务使用单独的网络。在本文中，我们介绍了一种新的基于Transformer的激光雷达多任务学习范式。所提出的LiDARFormer利用跨空间全局上下文特征信息，并利用跨任务协同作用来提升激光雷达在多个大规模数据集和基准测试上的感知任务性能。我们新颖的基于Transformer的框架包括一个跨空间Transformer模块，它学习二维密集鸟瞰图（BEV）和三维稀疏体素特征图之间的注意力特征。此外，我们为分割任务提出了一个Transformer解码器，通过利用类别特征表示来动态调整学习到的特征。进一步地，我们在一个共享的Transformer解码器中结合分割和检测特征，并使用跨任务注意力层来增强和整合目标级和类别级特征。LiDARFormer在大规模的nuScenes和Waymo开放数据集上针对三维检测和语义分割任务进行了评估，并且在这两个任务上都优于所有先前发表的方法。值得注意的是，对于单模型纯激光雷达方法，LiDARFormer在具有挑战性的Waymo和nuScenes检测基准测试中分别达到了76.4%的L2 mAPH和74.3%的NDS的最先进性能。

1 简介

激光雷达点云检测与语义分割任务旨在预测目标级三维边界框和点级语义标签，这是自动驾驶感知中最基础的任务之一。随着大规模激光雷达点云数据集[3,46]的发布，将这些任务集成到单一框架中的研究兴趣激增。当前主流方法[19,66]依赖基于体素网络的稀疏卷积[63,14]，但不同任务仅通过共享底层特征连接，未考虑任务间高度相关的高层上下文信息。另一方面，最新研究[44,48,60]尝试融合包含体素级和点级信息的多视角特征，这类方法更侧重于利用局部点几何关系恢复细粒度细节。然而，在激光雷达感知任务中，如何高效提取和共享全局上下文信息在很大程度上仍未被充分探索。

与此同时，基于Transformer的网络结构[4,53,12,59,73]在2D图像检测和分割任务中开始展现出卓越性能。除了直接用Transformer编码器替代传统CNN[16,71]外，各类方法[4,80,12,28]探索利用Transformer解码器提取目标级或类别级特征表示，这些表示可作为特征学习的强上下文信息。这种Transformer解码器设计随后被应用于近期的激光雷达感知方法[78,1,37]。然而，当前用于激光雷达检测和分割任务的Transformer解码器仍在不同特征图上独立运行，尚未实现统一。

是否有可能开发一个基于Transformer的统一多任务激光雷达感知网络，使其具备学习全局上下文信息的能力？为实现这一目标，我们在基于体素的框架中引入了三个新颖组件。第一个组件是跨空间Transformer模块，用于增强三维稀疏体素空间与二维密集鸟瞰图（BEV）空间之间的特征映射——这两种空间常分别用于获取分割和检测任务的特征表示。第二个组件是基于Transformer的优化模块，作为分割解码器：该模块利用Transformer提取类别特征嵌入，并通过双向交叉注意力优化体素特征。最后，我们提出一种多任务学习结构，将分割和检测的Transformer解码器整合为统一的Transformer解码器，使网络能够通过跨任务注意力传递高层特征（如图1所示）。这三个创新组件共同构建了名为LiDARFormer的强大网络，适用于下一代激光雷达感知任务。

我们在两个具有挑战性的大规模激光雷达数据集上评估了我们的方法：nuScenes数据集[3]和Waymo开放数据集[46]。我们的方法在检测和语义分割任务上均树立了新的最先进标准，在nuScenes三维检测中实现了74.3%的NDS（归一化检测分数），在nuScenes语义分割中达到了81.5%的mIoU（平均交并比）。LiDARFormer在Waymo开放数据集检测集上也实现了76.4%的mAPH（航向精度加权平均精度），超越了所有先前方法。

我们的主要贡献总结如下：

• 提出跨空间Transformer模块，用于改善多任务网络中稀疏体素特征与密集BEV特征之间的特征迁移学习。

• 首次提出激光雷达跨任务Transformer解码器，桥接目标级和类别级特征嵌入的跨任务信息学习。
• 引入基于Transformer的粗到细网络，利用Transformer解码器为激光雷达语义分割任务提取类别级全局上下文信息。
• 我们的网络在两个主流大规模激光雷达基准测试中均实现了最先进的三维检测和语义分割性能。

2 相关工作

基于体素的激光雷达点云感知 与大多数直接在室外或室内点云数据中学习点级特征的点云网络[41,42,30,58,22,61,49,75]不同，激光雷达点云感知通常需要将大规模稀疏点云转换为三维体素图[77,81]、二维鸟瞰图（BEV）[64,27,74]或距离视图图[47,18,56,57,38,15]。得益于点云处理中三维稀疏卷积层[63,14]的发展，基于体素的方法在高性能和高效运行时间方面均占据主导地位。CenterPoint[69]和AFDet[20]采用无锚框设计，通过热图分类检测目标；Cylinder3D[81]利用圆柱形体素划分提取体素级特征；LargeKernel3D[10]表明，来自更大感受野的长距离信息可显著提升性能；LidarMultiNet[66]提出了一种统一不同激光雷达感知任务的多任务学习网络。

基于体素的方法由于投影或体素化过程中引入的信息丢失，不得不在准确性和复杂性之间进行权衡。为缓解量化误差，近期一些方法[48,44,68,60]提出融合多视角特征图的特征，将点级信息与二维BEV/距离视图和三维体素特征相结合。PVRCNN[44]和SPV-NAS[48]采用了两个并行的点级和体素级特征编码分支，并在每个网络块中连接这两种特征；RPVNet[60]则通过门控融合模块，在编码器-解码器分割网络中进一步融合了所有点、体素和距离图像特征。与这些专注于细粒度特征细节的方法不同，我们的方法旨在增强基于体素的网络中的全局特征学习。

分割优化 在图像领域，多种方法[29,82,8,72,70]采用多阶段从粗到细优化分割预测。ACFNet[72]提出注意力类别特征模块，基于粗分割图优化逐像素特征；OCR[70]进一步拓展该思路，利用逐像素特征与目标上下文表示的双向连接来丰富特征。相比之下，优化模块在点云语义分割中应用较少。

Transformer解码器
Transformer架构[50]近年来广受欢迎。基于二维Transformer主干网络[16,71]的发展，各类方法[80,76,53,12,59,73]被提出以解决二维检测与分割问题。根据输入来源，视觉Transformer可分为编码器[71,76,59]和解码器[4,53,12,70,73,28]，其中编码器通常作为特征编码网络替代传统神经网络，而解码器则用于为下游任务提取类别级或实例级特征表示。在激光雷达领域，已有若干检测方法[39,65,36,43,1,40,32,78]开始将Transformer解码器集成到现有框架中，其在端到端训练[39]和多帧[65,78]/模态[1,32]特征融合方面展现出巨大潜力。然而，如何在激光雷达分割任务中有效应用Transformer解码器仍是一个未被充分探索的领域，因此本文提出一种基于Transformer解码器的新型类别感知全局上下文优化模块，同时挖掘检测与分割解码器之间的协同作用。

3 Method

在本节中，我们介绍LiDARFormer的设计。如图2所示，我们的框架包含三个部分：（3.1）使用三维稀疏卷积的三维编码器-解码器主干网络；（3.2）在鸟瞰图（BEV）中提取大规模上下文特征的跨空间Transformer（XSF）模块；（3.3）从体素和BEV特征图中聚合类别级和目标级全局上下文信息的跨任务Transformer（XTF）解码器。我们的网络采用了LidarMultiNet[66]的多任务学习框架，但通过共享的跨任务注意力层进一步关联了分割和检测任务之间的全局特征。

3.1. 基于体素的激光雷达感知

激光雷达点云语义分割和目标检测旨在预测点云$P = {p_i|p_i \in \mathcal{R}^{3+c}}^N_{i=1}$中的逐像素语义标签$L = {l_i|l_i \in (1...K)}^N_{i=1}$和目标边界框$O = {o_i|o_i \in \mathcal{R}^7}^B_{i=1}$，其中$N$表示点的数量，$B$和$K$分别表示目标和类别的数量。每个点具有$(3 + c)$个输入特征，即三维坐标$(x, y, z)$、反射强度、激光雷达延伸度、时间戳等。每个目标由其三维位置、大小和方向表示。

体素化 我们首先将点云坐标(x, y, z)转换为体素索引$${\mathcal{I}*i=(\lfloor \frac{x_i}{s_x} \rfloor, (\lfloor \frac{y_i}{s_y} \rfloor,(\lfloor \frac{z_i}{s_z} \rfloor)}^N{i=1}$，其中$s$是体素大小。然后，我们使用一个简单的体素特征编码器，该编码器仅包含多层感知机(MLP)和最大池化层来生成稀疏体素特征表示$\mathcal{V} \in \mathbb{R}^{M \times C}$： $$ \mathcal{V}*j = \max{\mathcal{I}_i = \mathcal{I}*j} \left(\text{MLP}(p_i) \right), \quad j \in (1 \ldots M) $$ *其中M是唯一体素索引的数量。我们还通过多数投票生成每个稀疏体素的真实标签： $$ L^v_j = \arg \max{\mathcal{I}_i = \mathcal{I}_j} (l_i) $$*

基于稀疏体素的主干网络

我们采用VoxelNet[77]作为网络主干，其中体素特征在编码器中逐步下采样至原始尺寸的$\frac{1}{8}$。稀疏体素特征被投影到密集鸟瞰图（BEV）上，随后通过二维多尺度特征提取器提取全局信息。对于检测任务，我们在BEV特征图上附加检测头以预测目标边界框；对于分割任务，BEV特征被重新投影到体素空间，通过U-Net解码器将特征图上采样回原始尺度。我们使用体素级标签L^v监督模型，并在推理阶段通过去体素化步骤将预测标签投影回点级别。

3.2 跨空间Transformer

如图1所示，基于体素的激光雷达检测与分割通常需要主干网络分别在二维密集鸟瞰图（BEV）空间和三维稀疏体素空间提取特征表示。为解决融合这两个任务所学特征的挑战，先前的多任务网络[66]提出全局上下文池化模块，仅基于位置直接映射特征而未考虑稀疏性差异。相比之下，我们提出跨空间Transformer模块，利用可变形注意力增强空间间的特征提取以进一步扩大感受野。如图2所示，我们采用跨空间Transformer实现以下功能：1）将最后尺度的稀疏体素特征$\mathcal{F}^{sparse}*{in}\in\mathcal{R}^{C×M^{\prime}}$*转换为密集BEV特征（稀疏到密集）；2）将二维多尺度特征提取器输出的密集BEV特征$\mathcal{F}^{dense} \in \mathcal{R}^{(C× \frac{D}{dz})× \frac{H}{dx} ×\frac{W}{dy}}$*转换为稀疏体素特征$\mathcal{F}^{sparse}*{in}\in \mathcal{R}^{C×M^{\prime}}$（密集到稀疏），其中$d$为下采样率，$M^{\prime}$为编码器最后尺度的有效体素数。跨空间Transformer的结构如图3所示。具体而言，在图3a中，$\mathcal{F}^{dense}$按高度切分为$\mathcal{F}^{dense}*{3D} \in \mathcal{R}^{(C× \frac{D}{dz})× \frac{H}{dx} × \frac{W}{dy}}$，然后从$\mathcal{F}^{dense}*{3D}$中选取$\mathcal{F}^{sparse}*{in}$有效坐标$(u, v, h)$处的特征作为查询$Q*{3D}$，以预测$\mathcal{F}^{sparse}*{out}$。模块采用可变形注意力[80]作为自注意力层，用于挖掘密集特征图中的全局信息。由于二维多尺度特征提取器主要关注BEV层级信息，导致$\mathcal{F}^{sparse}$缺乏高度信息，因此我们设计多头多高度注意力模块以学习所有高度的特征：对于高度h处切片BEV特征图上位置为$\xi = (u, v)$的参考体素，可变形自注意力通过线性层学习所有头和高度下的BEV偏移$\Delta\xi$，并通过双线性插值从不同多高度切片BEV特征图中采样$\xi+ \Delta\xi$处的特征。多高度可变形自注意力的输出$\chi(p)$可表示为： $$ \chi (p) = \sum\limits*{i = 1}^{N_{head}}* {{W_i}\left[{\sum\limits_{j = 1}^{N_{height}}* {\sum\limits_{r = 1}^R {\sigma \left({{W_{ijr}}{q_p}} \right)W_i^{\prime}{x^j}\left({\xi + \Delta {\xi *{ijr}}} \right)} } } \right]} $$ 其中，$N*{\text{head}}$为头数，$N_{\text{height}} =\frac{D}{dz}$为高度层数，$W$为可学习权重，$R$为采样点数，$x^j$为多高度切片的BEV特征，$q_p$为位置$\xi$处的查询特征，${\sigma \left({{W_{ijr}}{q_p}} \right)}$为注意力权重。由于密集到稀疏的跨空间Transformer应用于二维特征提取器之后，不会影响已学习的二维BEV特征，因此对检测性能提升的影响有限。为扩大二维BEV特征提取器的感受野，我们以类似方式添加跨空间Transformer模块，将$\mathcal{F}^{\text{sparse}}_{\text{in}}$转换为密集BEV特征（如图3b所示），使输入二维多尺度特征提取器的BEV特征具备更丰富的上下文信息。

3.3 跨任务Transformer解码器

尽管目标检测和语义分割共享相关信息，但它们通常在两个独立的网络结构中进行学习。LidarMultiNet[66]的研究表明，通过共享中间特征表示，检测和分割性能均可得到提升。然而，在多任务网络的训练过程中，高层信息并未实现共享。为进一步挖掘多任务学习的协同效应，我们提出使用共享的Transformer解码器来桥接分割任务的类别级信息与检测任务的目标级信息。在本节中，我们首先介绍一种利用类别特征嵌入执行动态分割的新型分割解码器，然后探讨通过跨任务注意力将该分割解码器与传统检测解码器相连接的方法。

分割Transformer解码器

受二维图像分割中粗到细方法[72,70]的启发，我们提出一种类别感知特征优化模块，以增强分割任务的全局信息学习能力。该模块首先通过初始分割预测生成类别特征嵌入，然后利用带有双向交叉注意力的Transformer对体素特征表示和类别特征表示进行双向优化。优化后的类别特征表示还将作为动态卷积核应用于后续分割头中。

给定初始语义分割分数$y = \left\{ {pred_j\left| {pred_j \in {{\left[{0,1} \right]}^K}} \right.} \right\}_{j = 1}^M$及其编码特征表示$\mathcal{F} \in \mathcal{R}^{M×C}$（其中$M$为有效预测数），我们按以下方式生成类别特征嵌入$\varepsilon = \left\{ {{\varepsilon _k}\left| {k \in \left\{ {1 \ldots K} \right\}} \right.} \right\}$：${\varepsilon *k} = \frac{{\sum\nolimits*{j = 1}^M {pred_j\left[k \right] \cdot {\mathcal{F}*{j}} }}}{{\sum\nolimits*{j = 1}^M {pred_j\left[k \right]} }}$。在我们的跨任务Transformer中，使用粗预测及其对应的BEV特征初始化类别特征嵌入。类别特征嵌入ε基于每次扫描的粗分割结果封装了类别中心信息。假设同一类别的点在编码特征嵌入中具有相似或相关的特征，则学习到的类别特征可帮助网络区分分割头中易混淆的边缘点。

与[70]类似，我们提出使用Transformer解码器通过双向交叉注意力同步提取类别特征嵌入并优化原始体素特征。如图4所示，我们的Transformer结构针对体素特征$\mathcal{V} \in \mathcal{R}^{M×C}$和类别特征$\varepsilon \in \mathcal{R}^{K×C}$设计了两个并行分支。我们采用包含多头自注意力层、多头交叉注意力层和前馈层的标准Transformer解码器[50]，以ε作为初始查询嵌入来提取类别特征。在交叉注意力层中，查询Q_c是ε的线性投影，而键K_v和值V_v是$\mathcal{V}$的线性投影，其公式可表示为： $$ CrossAtt\left({\mathcal{V} \to \varepsilon } \right) = Softmax \left({\frac{{{Q_C}K_v^T}}{{\sqrt C }}} \right){V_\mathcal{v}} $$ 接下来，我们使用反向Transformer解码器将编码后的类别特征传递回体素特征。由于体素规模庞大，在体素分支中使用自注意力并不现实。相反，查询$\bf{Q}*mathcal{v}$来自$\mathcal{V}$的线性投影，键$\bf{K}c$和值$\bf{V}{c}$则是经过线性变换的类别特征ε。 $$ CrossAtt\left({\varepsilon^{\prime} \to \mathcal{V} } \right) = Softmax \left({\frac{{{Q*\mathcal{v}}K_C^T}}{{\sqrt C }}} \right){\bf{V}_\mathcal{c}} $$ 然后，将输出的体素特征$\mathcal{V}^{\prime}$与原始特征$\mathcal{V}^{r} = (\mathcal{V},\mathcal{V}^{\prime})$在分割头中进行拼接。

动态卷积核

传统分割网络采用包含卷积或线性层的分割头，将体素特征的通道数压缩至类别数进行预测。分割头中学习的权重在不同帧之间共享，因此难以适应场景的动态变化。受图像实例分割新趋势[54,53,12,28]启发，我们直接使用学习到的类别特征嵌入$\varepsilon^{\prime}$作为卷积核生成语义logits：$S = \frac{{\Phi \left({{V^r}} \right) \cdot {\varepsilon ^{'T}}}}{{\sqrt C }}\in ,\mathcal{R}^{M×K}$，其中$\Phi$是将体素特征通道数压缩至$C$的卷积层。

跨任务注意力

如图4所示，我们采用在CenterFormer[78]中已被充分研究的检测Transformer解码器，将目标级特征表示为从BEV中心提议初始化的中心查询嵌入。我们使用BEV特征初始化类别特征嵌入ε，将类别特征和中心特征拼接后输入共享Transformer解码器。在解码器中，检测和分割任务之间的信息通过跨任务自注意力层实现双向传递。由于内存限制，类别特征和中心特征分别从体素和BEV特征图中聚合信息。

4 实验

在本节中，我们展示了所提出方法在两个大规模公开激光雷达点云数据集上的实验结果：nuScenes数据集[3]和Waymo开放数据集[46]，这两个数据集均包含三维目标边界框标注和逐像素语义标签标注。我们还对模型的改进进行了详细的消融研究和深入分析，更多细节和可视化结果包含在补充材料中。

4.1 数据集

nuScenes数据集是由Motional开发的大规模自动驾驶数据集，包含1000个场景的20秒视频数据，每个场景由20Hz的Velodyne HDL-32E激光雷达传感器（32线）采集。nuScenes在以2Hz采样的每个关键帧中提供目标边界框标注和逐像素语义标签，共包含16个用于语义分割评估的类别，其中10个前景目标（“事物”）类别带有真实边界框标签，用于目标检测任务。nuScenes检测任务采用平均精度（mAP）和nuScenes检测分数（NDS）作为评估指标，语义分割则使用平均交并比（mIoU）。

Waymo开放数据集（WOD）包含约2000个场景的20秒视频数据，由64线激光雷达传感器以10Hz频率采集。尽管WOD为每帧提供目标边界框标注，但仅在部分以2Hz采样的关键帧中包含语义标注，共涉及23个语义类别，采用标准mIoU作为评估指标。目标边界框标注涵盖车辆、行人和骑行者3个类别，用于三维检测评估，主要指标为航向精度加权平均精度（APH）。真实目标分为两个难度等级：LEVEL_1（L1）对应激光雷达点数超过5个且不属于L2的目标；LEVEL_2（L2）对应激光雷达点数至少1个且最多5个，或手动标注为“困难”的目标。主指标mAPH L2综合考虑L1和L2目标进行计算。

4.2 实验设置

我们使用AdamW优化器结合单周期调度器对模型进行20个epoch的训练。大多数实验在8块Nvidia A100 GPU上进行，批量大小为16。在Waymo开放数据集上的多任务训练实验中，由于GPU内存限制，批量大小调整为8。对于nuScenes数据集，体素大小设置为[0.1, 0.1, 0.2]；Waymo开放数据集的体素大小为[0.1, 0.1, 0.15]。
在分割任务中，我们采用交叉熵损失与Lovasz损失[2]的组合对网络进行优化；检测任务则遵循[69]的方法，使用常见的中心热图分类损失和边界框回归损失。我们在输出的体素特征或BEV特征上添加辅助损失以监督分割预测，该损失用于初始化类别特征嵌入。所有损失通过多任务不确定性加权策略[25]进行融合。
在nuScenes中，我们将前9帧扫描的点云与当前点云拼接；Waymo开放数据集中则拼接前2帧扫描的点云。训练过程中采用了标准的数据增强策略[51,66]。更多网络细节和训练配置见补充材料。

4.3 主要成果

我们在nuScenes和WOD数据集上展示了检测与分割的基准测试结果。测试集中其他方法的所有结果均来自文献，其中大多数方法采用测试时增强（TTA）或集成方法来提升性能。除了多任务网络外，我们还提供了仅使用分割Transformer解码器训练的纯分割模型变体结果。

nuScenes数据集 在表1和表2中，我们将LiDARFormer与nuScenes测试集上的其他先进方法进行了对比。作为单模型，LiDARFormer实现了81.5%的mIoU、71.5%的mAP和74.3%的NDS，达到顶尖性能。值得注意的是，检测任务的结果大幅超越了所有先前方法，尤其是在mAP指标上。尽管LiDARFormer的分割性能仅比LidarMultiNet高0.1%，但相比之下，其无需第二阶段处理且可端到端训练。为了在不受测试时增强影响的情况下与其他方法公平比较，我们在表3中展示了nuScenes验证集上的性能：仅分割的LiDARFormer实现了81.7%的mIoU，而完整的LiDARFormer在保持检测性能（NDS 70.8%）的同时将mIoU进一步提升至82.7%，超越了所有先前的先进方法，与测试集结果一致。

Waymo开放数据集 表4显示了LiDARFormer在WOD测试集上的检测结果。LiDARFormer以76.4%的L2 mAPH达到了先进水平，甚至超越了摄像头-激光雷达融合方法及使用更多帧的方法。最后，表5报告了WOD验证集结果：我们基于开源代码复现了PolarNet和Cylinder3D的结果以作对比，仅分割的LiDARFormer在验证集上实现了71.3%的mIoU（L2级别），多任务模型在分割任务上比先前最佳多任务网络提升了0.3%；在竞争更激烈的检测任务中，我们的方法以76.2%的L2 mAPH达到最佳结果。

4.4 消融实验

Transformer结构对分割任务的影响 表6展示了仅针对分割任务训练时，我们提出的各组件的有效性。我们采用3.1节描述的网络作为基线模型，这一简单设计已能与当前其他先进方法竞争。添加分割Transformer解码器后，nuScenes和WOD的mIoU分别提升1.7%和0.3%；通过将前一帧点云与当前帧拼接，结果进一步提升2.5%和0.6%；跨空间Transformer也使mIoU分别提高0.9%和0.1%。

统一多任务Transformer解码器的影响 表7展示了我们提出的Transformer解码器在多任务网络中的改进。我们以LidarMultiNet[66]的第一阶段结果作为基线，在检测或分割分支中单独添加Transformer解码器均提升了两项任务的性能——由于多任务网络共享主干，单一任务的改进可促进特征表示学习。我们提出的共享Transformer解码器通过引入跨任务注意力学习，实现了更优的整体性能，跨空间Transformer模块进一步提升了性能（尤其是检测任务）。此外，表8评估了多任务网络的全景分割性能：即使未使用专门针对全景分割的第二阶段，我们的模型仍优于此前最佳方法LidarMultiNet，证明多任务Transformer解码器可生成更兼容两项任务的结果。

4.5 分析

分割解码器分析 表9对比了不同Transformer设计下我们方法的纯分割性能。移除任一方向的交叉注意力均导致性能下降，动态卷积核设计比传统分割头提升0.8% mIoU，而不使用辅助分割头初始化类别嵌入时性能降低0.3%。

跨空间Transformer分析 表10展示了跨空间Transformer（XSF）模块在检测和分割任务中的有效性。若用参数量相近的额外卷积层替代XSF，分割性能下降0.8%；在多任务模型中仅替换稀疏到密集的XSF时，分割性能基本不变，但检测性能显著下降。这表明密集到稀疏和稀疏到密集的XSF对两项任务的贡献不同。图5可视化了跨空间Transformer中的可变形偏移：传统直接映射方法仅利用相同位置的特征进行3D-2D空间特征传递，可能忽略密集2D BEV图中学习到的有用特征，而我们的方法能够在更广泛区域聚合相关特征。

5 结论

在本文中，我们提出了LiDARFormer，这是一种新颖且高效的多任务激光雷达感知框架。我们的方法通过强化体素特征表示，以更简洁有效的方式实现了检测与分割任务的联合学习。尽管LiDARFormer专为纯激光雷达输入设计，但其跨空间Transformer（XSF）和跨任务Transformer（XTF）模块可通过交叉注意力层轻松扩展至多模态和时序特征的学习。类似地，XSF可在可变形注意力模块中引入多尺度特征图，进一步提取更大感受野的上下文信息。LiDARFormer在竞争激烈的nuScenes和Waymo检测与分割基准中刷新了先进性能，我们相信这项工作将为该领域未来的创新研究提供启发。

A 网络细节

体素特征编码器 我们采用与[74]相同的设计将点云编码为体素特征图。首先，将每个体素内的点分组，并向点特征$P \in \mathcal{R}^{3+c+6}$附加6个额外特征，即对应体素的中心坐标$(x_v, y_v, z_v)$和点到中心的偏移量$(x - x_v, y - y_v, z - z_v)$。接下来，使用4个堆叠的多层感知机（MLP）层将点特征变换到高维

空间，随后通过稀疏最大池化层在每个有效体素中提取体素特征表示。每层MLP的通道数依次为[64, 128, 256, 256]。

跨空间Transformer（XSF）结构 我们应用2个堆叠的Transformer块，每个块包含4个头的可变形自注意力机制。在**密集到稀疏（Dense-to-Sparse）XSF**中，每个头的通道数为64；在**稀疏到密集（Sparse-to-Dense）XSF**中，每个头的通道数为32。两种XSF的前馈网络（FFN）通道数均为256。每层均采用**预归一化（pre-norm）**而非后归一化（post-norm）。

跨任务Transformer（XTF）结构 我们使用3个堆叠的Transformer解码器层，每层包含4个头的自注意力和交叉注意力机制。每个头的通道数为32，前馈网络（FFN）的通道数为64。

B 深度讨论

XSF模块深度分析 表11展示了密集到稀疏XSF中不同查询类型的效果。LiDARFormer采用有效体素对应BEV特征图的特征作为查询（Query），而"体素查询"直接使用稀疏特征图的体素特征，"嵌入查询"则将有效坐标的嵌入向量作为查询。实验显示，后两种方式分别导致mIoU下降0.3%和0.4%，这可能是由于二维多尺度特征提取器生成的BEV特征包含更丰富的上下文信息。

运行时间与模型规模 我们在Nvidia A100 GPU上评估了LiDARFormer的运行时间和模型大小。图6表明，多任务网络通过共享主干网络显著降低延迟。尽管与之前的两阶段多任务网络延迟相近，但我们的方法以端到端单阶段设计实现更优性能。此外，LiDARFormer的参数规模（77M）显著小于LidarMultiNet（131M）。

类别特征嵌入初始化分析 在跨任务Transformer解码器中，我们使用粗分割预测及其BEV特征初始化类别特征嵌入。表12显示，若改用体素特征初始化，分割任务性能提升但检测任务性能下降，说明BEV特征更有利于跨任务信息对齐。

跨任务Transformer可视化分析 通过对比基线模型（无Transformer模块），图7显示LiDARFormer的改进主要集中在不连续区域（如同一目标点被误分为不同类别），表明跨任务注意力有效增强了类别一致性。

极坐标表示的局限性 PolarNet[74]和Cylinder3D[81]证明极坐标在激光雷达分割中的潜力，但其通过模仿激光雷达扫描模式平衡体素点分布。然而，在Waymo数据集上的实验表明，将基线模型转换为极坐标（相同体素大小）导致mIoU下降1.4%。我们推测，较小的体素尺寸已缓解近程体素的点分布失衡，而极坐标在远程体素中引入的几何畸变反而降低性能。

nuScenes类别分割结果 表13展示了LiDARFormer在nuScenes测试集上的类别级性能。各分类的最佳结果分散于前五名方法中，反映出不同类别间的学习竞争（如"摩托车"类性能提升伴随"自行车"类下降）。如何处理相似类别间的竞争仍是待解决的问题。

C 定性结果

如图8、9所示，我们展示了LiDARFormer在nuScenes和WOD数据集上的定性结果。我们的方法能够在多样化的环境中生成准确的语义预测。在复杂城市道路场景中，模型清晰区分了车辆、行人、道路、植被等类别，尤其在目标边界模糊或点云稀疏区域（如远处车辆、遮挡物体）仍能保持较高的分割精度和检测框定位准确性。可视化结果表明，跨空间Transformer和动态卷积核设计有效提升了特征的空间一致性和类别判别力，验证了多任务学习框架对激光雷达点云理解的有效性。