引用格式：郑强文,吴升,魏婧卉.VoxTNT：基于多尺度Transformer的点云3D目标检测方法[J].地球信息科学学报,2025,27(6):1361-1380.
[Zheng Q W, Wu S, Wei J H. VoxTNT: A multi-scale transformer-based approach for 3D object detection in point clouds[J]. Journal of Geo-information Science, 2025,27(6):1361-1380.] **DOI:**10.12082/dqxxkx.2025.250122; **CSTR:**32074.14.dqxxkx.2025.250122

VoxTNT：基于多尺度Transformer的点云3D目标检测方法

郑强文[1],吴　升[2*],魏婧卉[1]

1. 福州大学计算机与大数据学院,福州 350100；2. 福州大学数字中国研究院(福建),福州 350100

VoxTNT: A Multi-Scale Transformer-based Approach for 3D Object Detection in Point Clouds

ZHENG Qiangwen[1], WU Sheng[2*], WEI Jinghui[1]

1. The College of Computer and Data Science, Fuzhou University, Fuzhou 350100, China; 2. The Academy of Digital China (Fujian), Fuzhou University, Fuzhou 350100, China

Abstract：[Background] Traditional methods, due to their static receptive field design, struggle to adapt to the significant scale differences among cars, pedestrians, and cyclists in urban autonomous driving scenarios. Moreover, cross-scale feature fusion often leads to hierarchical interference. **[Methodology]** To address the key challenge of cross-scale representation consistency in 3D object detection for multi-class, multi-scale objects in autonomous driving scenarios, this study proposes a novel method named VoxTNT. VoxTNT leverages an equalized receptive field and a local-global collaborative attention mechanism to enhance detection performance. At the local level, a PointSetFormer module is introduced, incorporating an Induced Set Attention Block (ISAB) to aggregate fine-grained geometric features from high-density point clouds through reduced cross-attention. This design overcomes the information loss typically associated with traditional voxel mean pooling. At the global level, a VoxelFormerFFN module is designed, which abstracts non-empty voxels into a super-point set and applies cross-voxel ISAB interactions to capture long-range contextual dependencies. This approach reduces the computational complexity of global feature learning from $O(N^2)$ to $O(M^2)$ (where $M << N$, M is the number of non-empty voxels), avoiding the high computational complexity associated with directly applying complex Transformers to raw point clouds. This dual-domain coupled architecture achieves a dynamic balance between local fine-grained perception and global semantic association, effectively mitigating modeling bias caused by fixed receptive fields and multi-scale fusion. **[Results]** Experiments demonstrate that the proposed method achieves a single-stage detection Average Precision (*AP*) of 59.56% for moderate-level pedestrian detection on the KITTI dataset, an improvement of approximately 12.4% over the SECOND baseline. For two-stage detection, it achieves a mean Average Precision (*mAP*) of 66.54%, outperforming the second-best method, BSAODet, which achieves 66.10%. Validation on the WOD dataset further confirms the method's effectiveness,

收稿日期：2025-03-14；**修回日期：**2025-04-18.

基金项目：公共数据开发利用科技创新团队(闽教科〔2023〕15号)。[**Foundation item:** Fujian Provincal Program for Innovative Research Team, Fujian ES [2023] No.15.]

作者简介：郑强文(1990—),男,福建龙岩人,博士生,主要从自动驾驶领域感知技术研究。E-mail: 593161522@qq.com

***通讯作者：**吴　升(1972—),男,福建松溪人,博士,教授,主要从事大数据分析与可视化、数字化规划、数字政府研究。E-mail: wusheng@fzu.edu.cn

achieving 66.09% mAP, which outperforms the SECOND and PointPillars baselines by 7.7% and 8.5%, respectively. Ablation studies demonstrate that the proposed equalized local-global receptive field mechanism significantly improves detection accuracy for small objects. For example, on the KITTI dataset, full component ablation resulted in a 10.8% and 10.0% drop in AP for moderate-level pedestrian and cyclist detection, respectively, while maintaining stable performance for large-object detection. **[Conclusions]** This study presents a novel approach to tackling the challenges of multi-scale object detection in autonomous driving scenarios. Future work will focus on optimizing the model architecture to further enhance efficiency.

Key words: intelligent transportation engineering; autonomous driving; point cloud; 3D object detection; voxel; transformer

***Corresponding author:** WU Sheng, E-mail: wusheng@fzu.edu.cn

摘要：【背景】传统方法因静态感受野设计较难适配城市自动驾驶场景中汽车、行人及骑行者等目标的显著尺度差异，且跨尺度特征融合易引发层级干扰。**【方法】**针对自动驾驶场景中多类别、多尺寸目标的3D检测中跨尺度表征一致性的关键挑战，本研究提出基于均衡化感受野的3D目标检测方法VoxTNT，通过局部-全局协同注意力机制提升检测性能。在局部层面，设计了Point-SetFormer模块，引入诱导集注意力模块（Induced Set Attention Block, ISAB），通过约简的交叉注意力聚合高密度点云的细粒度几何特征，突破传统体素均值池化的信息损失瓶颈；在全局层面，设计了VoxelFormerFFN模块，将非空体素抽象为超点集并实施跨体素ISAB交互，建立长程上下文依赖关系，并将全局特征学习计算负载从$O(N^2)$压缩至$O(M^2)$（$M<<N$，M为非空体素数量），规避了复杂的Transformer直接使用在原始点云造成的高计算复杂度。该双域耦合架构实现了局部细粒度感知与全局语义关联的动态平衡，有效缓解固定感受野和多尺度融合导致的特征建模偏差。**【结果】**实验表明，该方法在KITTI数据集单阶段检测下，中等难度级别的行人检测精度AP（Average Precision）值达到59.56%，较SECOND基线提高约12.4%，两阶段检测下以66.54%的综合指标mAP（mean Average Precision）领先次优方法BSAODet的66.10%。同时，在WOD数据集中验证了方法的有效性，综合指标mAP达到66.09%分别超越SECOND和PointPillars基线7.7%和8.5%。消融实验进一步表明，均衡化局部和全局感受野的3D特征学习机制能显著提升小目标检测精度（如在KITTI数据集中全组件消融的情况下，中等难度级别的行人和骑行者检测精度分别下降10.8%和10.0%），同时保持大目标检测的稳定性。**【结论】**本研究为解决自动驾驶多尺度目标检测难题提供了新思路，未来将优化模型结构以进一步提升效能。

关键词：智能交通工程；自动驾驶；点云；三维目标检测；体素；Transformer

1 引言

LiDAR点云驱动的3D目标检测对于自动驾驶[1-4]、机器人[5]、增强现实[6]、舰船检测[7-8]等领域具有重要作用。由于点云的稀疏、无序、不规则、高可变点密度等特性，如何学习高质量的点云3D特征面临较大挑战[9-11]。为了解决上述问题，研究人员提出了很多有效的方法，主要包括[1-3,12]：基于视图的方法[13-19]、基于体素的方法[10,20-25]、基于点的方法[26-32]、基于柱的方法[10,21,33-34]和基于图的方法[35-38]。但许多研究结果表明[1-3,39]，单独使用这些方法往往较难在效率和性能上取得较好的平衡。这些基于点云的3D检测方法中主要存在效率与精度均衡优化的问题，具体如下：

（1）基于视图、基于体素、基于柱的3D检测方法（如VoxelNet[10]、PointPillars[34]等），通常将不规则的3D点云数据转换成规则数据表示，然后再用MLP、CNN等基于深度学习的2D检测算法进行检测。这种方法简单直观，但将3D点云转换到2D规则表示，会丢失很多几何空间信息，并且感受野相对固定，因此会限制检测的性能。

（2）基于点和基于图的3D检测方法（如point-net[31]、point-gnn[35]等），能够直接处理点云数据并利用MLP、GNN等深度学习方法学习逐点特征。这类方法能够保留点云的原始结构，可以大幅提升检测精度，但是由于点云通常包含数以万计的空间点，导致计算速度缓慢和占用高额内存，代价较大，甚至无法保证实时检测。

尽管在复杂场景（如长尾分布场景）下涌现出CenterPoint[25]、DFAF3D[40]、BSAODet[41]、PG-RCNN[42]等优质方法，但在多类别、多尺寸目标检测任务中仍面临跨尺度表征一致性优化的挑战，特别是在实现小目标检测性能优化方面存在较大技术瓶颈。当前研究主要存在以下2个局限性：

（1）检测类别偏差问题：现有如SECOND[22]、PASS-PV-RCNN++[43]等方法[20,40,44-50]的研究焦点过度

集中于大尺度目标检测（如汽车类别），导致对行人、骑行者等小尺度目标的特征建模能力不足；

（2）模型跨尺度适用受限：针对不同类别目标的检测任务，现有PV-RCNN++[51]、PV-GNN[52]等方法[10,22,34,53-56]通常需要进行架构重构或调整较多参数（如特征维度、网络迭代层次、多尺度融合参数、损失函数等），这种策略不仅增加算法复杂度，还严重制约模型在开放场景中的自适应能力。

为解决上述方法存在的问题，一些研究使用混合方法[30,42-43,51-52,57-59]，即融合上述基于点、体素、柱、图等多种方法的优点并规避其缺点而设计相应的3D目标检测方案，实现了较好的检测性能。近年来，Transformer因其在自然语言领域的成功应用被引入3D目标检测领域[60-65]。研究者通过将Transformer与基于体素、基于柱、基于点或基于图的方法相结合，提出了一系列创新的解决方案[41,45,64,66-75]。例如，在注意力机制创新方面，ISAB[76]采用双交叉注意力模块替代传统自注意力机制，实现对任意大小的令牌簇计算自注意力，有效提高了汽车3D检测的性能、TED[74]通过变换等变性建模与高效网络设计，在自动驾驶3D检测任务中实现了精度与速度的平衡；在计算效率优化方面，SST[77]和SWFormer[78]将点云投影至柱状（Pillar）空间划分，利用局部2D窗口注意力机制降低计算复杂度，从而降低计算成本；在3D表征增强方面，VoxSeT[45]在体素内部使用交叉注意力机制，有效提升对遮挡目标的检测鲁棒性，DSVT[79]则通过动态体素划分策略，避免柱状投影导致的高度维度（Z轴方向）信息损失，实现了体素特征交互的注意力。尽管基于Transformer的3D检测框架通过自适应特征交互机制有效缓解了点云稀疏、无序、不规则等问题，但仍存在以下挑战：

（1）固定感受野限制小目标适应性。如以SST[77]和SWFormer[78]为代表的方法在固定尺寸的标记簇上执行自注意力计算，这种刚性感受野设定导致模型较难适配小目标的几何特性，尤其是当场景中存在显著尺寸差异的车辆、行人及骑行者目标时，检测性能易出现较大波动。

（2）层级间特征交互机制效率不足。以VoxSeT[45]为代表的通过跨区域交互机制提升了特征提取能力，但在传递信息时仍依赖传统固定感受野的前馈神经网络或3D稀疏卷积网络，较难获得动态长程依赖关系，限制深层的隐藏特征学习能力，导致复杂场景的3D检测适应性相对较差，小目标特征表达不足，影响检测稳定性。

为了解决上述效率与精度均衡优化和跨尺度表征一致性优化的问题，受到ISAB[76]和TNT（Transformer-in-Transformer）[80]2种方法的启发，本研究将3D点云空间进行体素分区，借鉴TNT[80]双重Transformer架构思想，使用ISAB[76]中约简的交叉注意力机制，提出了一种融合双重Transformer架构的3D检测方法，即VoxTNT，通过构建体素-Transformer耦合架构，设计均衡化局部和全局感受野的局部-全局特征协同学习策略，取得了较稳定的多尺度目标检测性能和较好的复杂场景鲁棒性。主要挑战是如何有效地将Transformer集成到体素分区下的点云检测框架中，并在保证推理速度的前提下学习高质量的点云特征。

2 VoxTNT整体框架与技术方法

2.1 整体框架

本研究遵循通用的3D目标检测流程，受到TNT（Transformer iN Transformer）[80]和ISAB（诱导集注意力块）[76]的启发，重点设计了3D特征学习网络，提出了VoxTNT（A Multi-scale Transformer-based Approach for 3D Object Detection in Point Clouds）方法，整体框架如图1所示，主要包含体素分区、位置嵌入模块、3D特征学习网络、点到BEV的特征变换和检测头几个部分。体素分区主要负责将原始3D点云空间划分为规则的体素网格，然后在每个体素网格内采样一定数量的点。位置嵌入模块主要用于将点云中点的相对位置嵌入注意力机制。3D特征学习网络是本研究的核心部分，在体素分区思想下，基于TNT[80]架构设计思路，体素局部区域内和全局区域间分别使用ISAB[76]交叉注意力机制设计的均衡化感受野的局部–全局特征协同学习的主干网络，PointSetFormer为该主干网络的核心组件。点到BEV的特征变换模块将学习的3D特征变换成更加密集的2D特征图表示，为后续的分类和边界框预测提供低维度的输入特征。检测头主要实现目标的分类和边界框预测。

VoxTNT的核心创新在于通过多层次注意力机制实现点云特征的高效建模，其亮点包括：① 动态局部建模：提出PointSetFormer模块，将点云处理转换为集合间交叉注意力学习，摒弃传统体素方法强制统一点数导致的填充或降采样缺陷，支持不同体

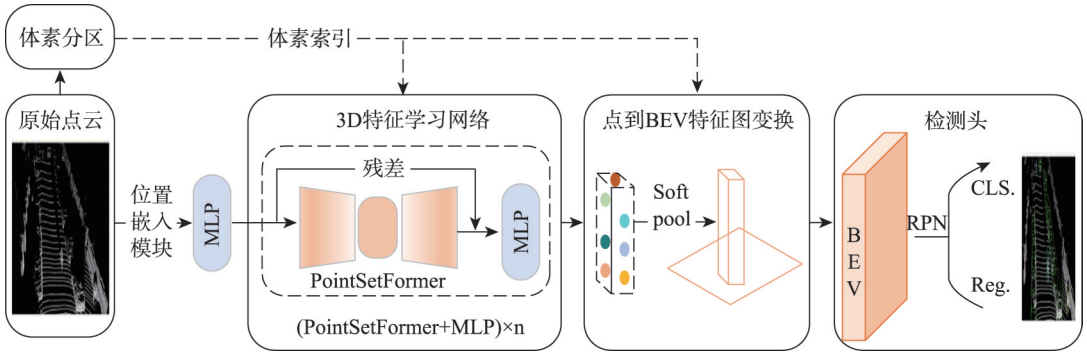


图1 VoxTNT总体框架

Fig. 1 The comprehensive architecture of VoxTNT

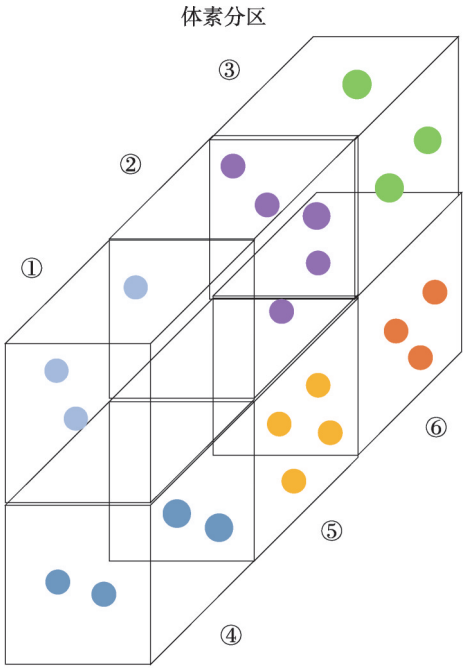
素点数的并行化计算，保留原始空间拓扑结构，通过两点注意力均衡局部感受野，充分捕捉细粒度几何特征；② 全局语义交互：PointSetFormer模块中设计VoxelFormerFFN组件，将非空体素抽象为超点集并构建跨体素交叉注意力，实现长程上下文信息传递与全局感受野均衡，动态聚焦关键区域以提升多尺度目标（如汽车与行人）的鉴别力；③ 层次化协同架构：通过编码器–解码器框架串联上述模块，在解码阶段融合局部细节与全局语义，抑制背景干扰，提升跨尺度一致性的检测性能，并在一定程度上兼顾效率与精度，为自动驾驶多尺度目标检测提供了新范式。

2.2 技术步骤及原理

（1）进行体素分区。本研究利用分区思想对原始点云进行体素分区（如图2所示，具体划分方法可以参看VoxelNet[10]），视每个体素为一个区域，在每个区域内采样T个点（根据实际需求，T可以是一个整数常量，也可以是变量）以在大幅减少点云数量的同时获得有效、稳定、均衡的关键点，为后续的推理节省计算开销，然后将每个区域内的点集合统一到同一个矩阵中（矢量化）以便于跨体素并行计算。

（2）构建3D特征学习网络。利用PointSetFormer的编码器对每个分区内的点进行特征编码并聚合体素局部特征，VoxelFormerFNN中间层作为M个非空体素间的信息传递网络以学习整个点云的全局特征，PointSetFormer解码器将全局交互后的体素特征逆向映射（广播）至每个体素内的T个点，并解码T个点的特征。3D特征学习模块是本研究的核心设计模块，将在下文详细介绍。

（3）进行点到BEV的特征变换。将学习到的3D特征投影成密集2D BEV特征图是3D检测的常



注：编号表示体素网格的序号，每个体素内采样的点用不同颜色标注以示区分。

图2 体素分区示意图

Fig. 2 Schematic diagram of voxel partitioning

用做法，以获得更高的召回率[22,45,53,58]。因此，本研究参考PointPillars[34]和VoxSeT[45]将主干网输出点云高维特征投影到BEV，为后续检测头中RPN网络提供输入特征。

（4）检测头。将2D BEV特征图进行特征优化并构建损失函数，2D BEV特征图通过检测头的RPN网络处理，即使用2个跨步（每个跨步包括3个卷积操作）的2D CNN网络进行卷积处理以增加特征密度，输出的卷积连接特征被传递到检测头以实现边界框的预测。为了提高VoxTNT主干网络特征学习能力，本研究借鉴PointRCNN[26]的做法，将前

景分割损失L_{seg}应用于输出特征，从而提高边界框检测的准确性。此外，本研究遵循基于锚的通用做法设计检测头，损失包括前景分割损失、边界框偏移回归损失、边界框方向预测损失、分类损失[10,22,34,81]。详细介绍可以参考PointRCNN[26]和Second[22]。

由于体素分区、位置嵌入模块、点到BEV特征图变换以及检测头都是通用模块，本研究不再详细介绍，后续将重点介绍3D特征学习网络的详细设计。此外，VoxTNT的3D特征学习网络不仅能用于单阶段的3D检测，也可以结合RoI[58,82]模块扩展至两阶段的3D检测。

需要特别说明的是，体素的特征可以由体素内的点特征聚合得到，为了保证提取每个体素特征的操作能够并行化，可以将跨体素的操作进行矢量化[45]。这种矢量化可以通过cuda[83]内核库中scatter_function[84]来实现，该函数在矩阵的不同段上执行对称约简，例如求和、最大值和平均值。将输入集视为单个矩阵，其中的每一行对应于一个逐点特征，并且其所属的体素可以通过体素坐标表来索引。

因此，给定点云的输入特征矩阵X：$\{X_i : i=1, \cdots, n\}$，通过对称函数$F(\cdot)$约简后的特征矩阵Y：$\{Y_j : j=1, \cdots, m\}$表示为：

$$Y = \left\{ F\left(\left\{ X_i : V_i = j \right\} \right) : j = 1, \cdots, m \right\} \tag{1}$$

式中：m表示非空体素区域的数量；V_i为体素索引。利用cuda[83]内核库中的散射函数scatter_function表示$F_{scatter}$，则上式可以表示为：

$$Y = F_{scatter}(X, V) \tag{2}$$

2.2.1 3D特征学习网络

3D特征学习网络是VoxTNT的核心模块。Transformer iN Transformer[80]通过嵌套Transformer结构与分层注意力机制，突破了传统方法在局部特征建模上的局限，实现了全局语义与局部细节的协同优化，为视觉Transformer的发展提供了新方向。本研究借鉴Transformer iN Transformer的架构思想设计了体素区域内局部交叉注意力机制和体素间全局交叉注意力机制的3D特征学习网络。具体来说，参照标准的Transformer网络层结构，VoxTNT特征学习主干网络由互连的多层感知器（MLP）和PointSetFormer（Point Set Transformer in voxel）模块组成，其中，PointSetFormer模块是3D特征学习网络的最关键部分。本研究使用批量归一化作为归一化层，每个PointSetFormer模块的输出都加入残差以获得最佳梯度流。点云特征的语义级别取决于PointSetFormer模块中体素大小，局部特征的学习由PointSetFormer编码器完成，全局特征的学习由VoxelFormerFFN中间层完成，PointSetFormer解码器实现局部和全局特征的融合。与逐步下采样和聚合逐点特征以进行上下文分组 提取特征的方法（如pointnet[31]、pointnet++[32]、votenet[85]等）不同，本研究的PointSetFormer骨干网是一种采用均衡化感受野机制下局部-全局特征协同策略的集合到集合转换的点云特征提取方法。此外，由于原始的自注意力机制没有关于序列中元素位置顺序的信息。因此，本研究为PointSetFormer和VoxelFormer模块引入了位置嵌入模块（Relative Position Embedding，PE模块），具体可以参看DETR[86]。

通常点云包含数万个点，如果直接使用注意力机制将耗费巨大的计算代价。为了减少计算量，PointSetFormer模块引入诱导集注意力块（ISAB）[76]作为本方法的注意力机制。ISAB利用由一组隐藏码（latent codes，L）诱导的2个约简的交叉注意力来近似集合中的完整自注意力。给定一个输入集$X \in \mathbb{R}^{n \times d}$和具有$k$个隐藏码的集合$L \in \mathbb{R}^{k \times d}$，则诱导集注意力块的输出集$O \in \mathbb{R}^{n \times d}$可以表示为：

$$H = CrossAttention1(L, X) \in \mathbb{R}^{k \times d} \tag{3}$$

$$\tilde{H} = FFN(H) \in \mathbb{R}^{k \times d} \tag{4}$$

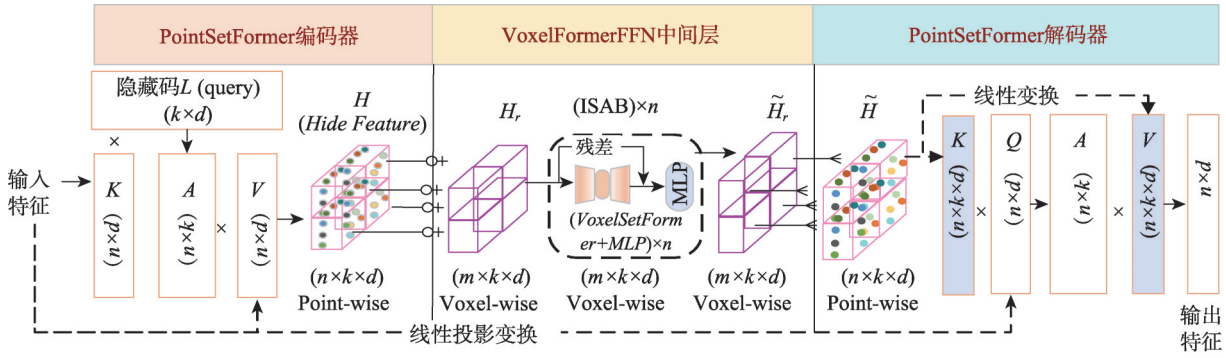
$$O = CrossAttention2(X, \tilde{H}) \in \mathbb{R}^{n \times d} \tag{5}$$

式中：$CrossAttention1$利用隐藏码L将输入集X变换成隐藏特征H；然后利用逐点前馈神经网络FFN将H变换成深层空间特征$\tilde{H}$；接着$CrossAttention2$关注输入集X到隐藏的深层特征$\tilde{H}$得到输出集O。值得注意的是，基于自注意力的诱导集注意力可以用低秩投影来近似，因此对输入集执行以隐藏码L作为聚类中心的k聚类操作可以替代完整的自注意力，并且$k \ll n$，因此可以大幅减少计算量。

PointSetFormer模块的PointSetFormer编码器（式（3））、VoxelFormerFFN中间层（式（4））和PointSetFormer解码器（式（5））3个部分的设计，具体如图3所示。可以看出，PointSetFormer模块是运用Transformer的一种实例，首先将输入集变换到隐藏空间，再通过VoxelFormerFFN中间层学习更深层次的隐藏特征，最后通过解码器生成输出集。

（1）PointSetFormer编码器

在编码器设计中，核心在于通过潜在空间编码机制实现特征表达，将原始点云进行3D体素网格



注：H、$\tilde{H}_r$、$\tilde{H}_r$和H是不同状态下的体素特征。L、Q、K、V等为交叉注意力变量。

图3 PointSetFormer模块设计

Fig. 3 Architecture of the PointSetFormer module

划分后，给每个非空体素分配一组隐藏码$L\in\mathbb{R}^{k\times d}$（相当于Transformer中的$Q$查询向量）对每个体素单元内的空间特征进行隐式表征，其中特征编码过程通过局部交叉自注意力机制实现。具体来说，将前一个模块的输入特征线性变换分别得到键$K\in\mathbb{R}^{n\times d}$和值$V\in\mathbb{R}^{n\times d}$，并对键$K\in\mathbb{R}^{n\times d}$和隐藏码$L\in\mathbb{R}^{k\times d}$执行交叉注意力得到注意力矩阵$A\in\mathbb{R}^{n\times k\times d}$。然后，通过体素坐标索引$V$对注意力矩阵$A$进行体素归一化以获得$\tilde{A}$，再与值$V$相乘，产生隐藏特征$H$。最后，本研究根据体素索引$V$对隐藏空间中的深层特征$H$进行体素缩减并得到体素局部聚合特征$H_r$。具体计算可以表示为：

$$\tilde{A}=\mathrm{Softmax}_{\mathrm{scatter}}(A,V),A=KL^{T} \tag{6}$$

$$H=\tilde{A}^{T}V \tag{7}$$

$$H_r=\mathrm{Sum}_{\mathrm{scatter}}(H,V) \tag{8}$$

式中：$\mathrm{Softmax}_{\mathrm{scatter}}$和$\mathrm{Sum}_{\mathrm{scatter}}$是scatter_function[84]库中的scatter库函数。

（2）VoxelFormerFFN中间层

PointSetFormer编码器仅对同一体素内的点集合执行局部交叉自注意力计算，这种设计虽然有效降低了计算复杂度，但也引入了体素间特征交互的挑战。受TNT[80]框架启发，本研究提出VoxelFormerFFN（Voxel-Level Transformer feed-forward network）模块以替代传统前馈网络（FFN）。该模块作为PointSetFormer编码器与解码器之间的中间层，其核心创新在于再次引入ISAB[76]机制实现全局层面（非空体素之间）的特征交互。具体而言，VoxelFormerFFN将每个非空体素区域建模为特征节点，将编码器输出的体素局部聚合特征H_r作为输入集合，通过构建跨体素的交叉自注意力机制（即VoxelFormer）实现全局上下文建模。相较于传统方法，本设计具有2个显著优势：首先，通过建立非空体素间的两两注意力关联，有效均衡了点云特征的全局感受野分布，这对存在显著尺度差异的复杂3D场景检测任务尤为重要；其次，该机制通过整合长程空间依赖关系，能够有效融合不同距离体素间的互补特征信息，从而提取更具鉴别力的深度全局特征表示。

PointSetFormer将约简后的非空体素聚合特征H_r作为输入（将体素视为特征节点），并根据H_r对应的体素坐标C_r构造相对位置编码。给定VoxelFormerFFN权重W，通过ISAB[45]算子计算得到的隐藏特征$\tilde{H}_r\in\mathbb{R}^{m\times k\times d}$可以写为：

$$\tilde{H}_r=ISAB\left(\left(H_r,C_r\right);W\right) \tag{9}$$

如果需要多次迭代ISAB算子（每次迭代都会加入残差），例如迭代3次，给定权重W_1、W_2和W_3，则得到的$\tilde{H}_r$可以改写为：

$$\tilde{H}_r=ISAB((\sigma(ISAB((\sigma(ISAB((H_r,C_r);W_1)),C_r);W_2)),C_r);W_3) \tag{10}$$

式中：σ表示非线性激活函数。VoxelFormerFFN因其能捕获理想的归纳偏差和全局上下文信息而成为PointSetFormer模块的最核心部分。

需要说明的是，VoxelFormerFFN组件内的VoxelFormer结构与PointSetFormer基本一致，不同点在于VoxelFormer编码器和解码器的中间层采用传统的FFN（Feed-Forward Network）实现，计算规则也遵循上述公式，只需要把VFF算子替换成FFN即可。此外，PointSetFormer中编码器使用的隐藏码称为局部隐藏码L_l，VoxelFormerFFN中VoxelForm-

er使用的隐藏码称为全局隐藏码L_g，L_l和L_g的不同组合对检测结果有较大影响，具体将在后续的消融实验中进行详细讨论。

（3）PointSetFormer解码器

在解码器的设计中，本研究将上述具有丰富隐藏特征的$\tilde{H}_r$作为输入，再重构输出集。具体来说，首先利用体素索引V把每个体素的隐藏特征广播（broadcast）到对应体素区域内的点，产生与输入集长度相同的$\tilde{H} \in R^{n \times k \times d}$。然后，使用线性投影从输入集和隐藏特征分别生成查询和键-值对。给定查询$Q \in R^{n \times d}$、键$K \in R^{n \times k \times d}$和值$V \in R^{n \times k \times d}$的矩阵，解码器输出的$O$如下：

$$\boldsymbol{A} = \left[A_1, \cdots, A_n \right] = \left[K_1 Q_1^T, \cdots, K_n Q_n^T \right] \quad (11)$$

$$\tilde{\boldsymbol{A}} = \left[Soft\max\left(A_1 \right), \cdots, Soft\max\left(A_n \right) \right] \quad (12)$$

$$O = \left[O_1, \cdots, O_n \right] = \left[\tilde{\boldsymbol{A}}_1^T V_1, \cdots, \tilde{\boldsymbol{A}}_n^T V_n \right] \quad (13)$$

式中：$\boldsymbol{A}$为查询Q和键K的注意力矩阵；$\tilde{\boldsymbol{A}}$为$\boldsymbol{A}$的归一化矩阵。

2.2.2 扩展至两阶段检测

本研究在架构设计上与经典体素特征编码（VFE）方法[10,53]存在本质区别：虽然两者均基于体素单元进行局部特征提取，但VFE通过逐点全连接层将点云特征压缩为单一向量，而PointSetFormer则通过隐藏空间交叉注意力构建动态特征交互网络。这种差异使VoxTNT的核心模块PointSetFormer具备更强的架构兼容性，可灵活应用于各类检测框架。

为验证方法的通用性，本研究将PointSetFormer分别与Voxel R-CNN[53]和Part-A^2-free[23]相结合以将其扩展至两阶段检测。实验结果表明，无论是单阶段还是两阶段架构，本研究方法均展现出显著优势。为了更清楚地说明本研究的贡献，本研究在实验环节同时报告了PointSetFormer在单阶段和两阶段3D检测中的性能。

3　实验设计与结果分析

3.1 数据集选择

3.1.1 数据集介绍

本研究主要基于KITTI[87]和1.2.0版本的Waymo Open Dataset (WOD)[88]两大主流数据集进行实验验证，其选取依据如下：KITTI作为经典基准数据集适用于算法基础性能验证，而WOD凭借大规模复杂场景数据可评估模型在实际环境中的鲁棒性。

KITTI数据集作为自动驾驶领域最广泛应用的基准之一，由德国卡尔斯鲁厄理工学院（KIT）与芝加哥丰田技术研究院（TTIC）联合开发，提供包含城市、乡村及高速场景的7 481组训练样本和7 518组测试样本。该数据集集成RGB图像与LiDAR点云双模态数据，涵盖15 000帧激光雷达扫描及同步视觉数据（总时长1.5 h），包含汽车、卡车、行人等8类交通目标的200 000个三维边界框标注。其特色在于依据目标尺寸、遮挡率及截断程度将检测难度划分为三级：简单级（高度≥40像素/无遮挡/截断≤15%）、中级（高度≥25像素/部分遮挡/截断≤30%）和困难级（高度≥25像素/严重遮挡/截断≤50%）。本研究重点针对汽车、行人及骑行人员3类核心目标，以中级难度（最贴近实际驾驶工况）作为主要评估基准，通过该分级体系系统性验证算法在复杂场景下的鲁棒性。更详细的介绍可以参见https://www.cvlibs.net/datasets/kitti/index.php。

WOD是由Waymo公司发布的多模态自动驾驶基准数据集，包含798个训练序列和202个验证序列，覆盖旧金山、凤凰城等地的城市与郊区多气候场景。数据通过5个高分辨率相机（水平视场360°）及5组激光雷达（4短程+1中程）同步采集，提供约23万次LiDAR扫描（以包含距离/强度信息的360°距离图形式存储）和100万张图像（总时长6.4小时），配套包含车辆姿态、传感器标定等元数据。其3D标注涵盖车辆、行人、骑行人员及交通标志四类目标（共1 200万实例），其中前3类为评测核心指标。每个连续场景持续约20 s，附带精确的2D/3D边界框标注，为时序感知算法研究提供重要支撑。更详细的介绍可以参见https://waymo.com/open/。

为了规避过拟合，尤其是KITTI训练和验证样本都小于4 000，因此需要对训练数据进行数据增强。本研究选择了简单常用的全局缩放、全局旋转、全局翻转和扰动的方法[10,22,35,58]，如表1所示。

表1　数据增强技术及使用概率一览表

Tab. 1　List of data enhancement technologies and probability of use

方法	是否使用	概率
全局缩放	✓	0.25
全局旋转	✓	0.25
全局翻转	✓	0.25
扰动	✓	0.25

3.1.2 评估指标

（1）KITTI评估指标

本研究遵循KITTI官方的通用评估协议，包含汽车（Car）、行人（Pedestrian）和骑行者（Cyclist）3个类别的3D检测，每个类别包括简单、中等和困难3个级别并用平均精度(AP)表示。其中，Car类的*IoU*阈值为0.7，Pedestrian和Cyclist类的*IoU*阈值为0.5。评价指标为：在11/40个不同召回阈值下计算每个类别的AP（Average Precision），所有类别下AP的平均值称为mAP（mean Average Precision）。

本研究分别在验证集和测试集中评估了VoxTNT方法。在验证集上的评价方法为：遵循先前研究[10,45,54]的一般惯例，在验证集上使用11个召回阈值（R11）下的AP（Average Precision）对VoxTNT方法评估结果，并与经典的单阶段基线方法进行比较。在测试集上的评价方法为：在40个不同召回阈值（R40）下计算*AP*和*mAP*，将VoxTNT扩展至两阶段检测，并与相关方法进行比较。

（2）WOD评估指标

WOD采用双维度评估体系：基础指标沿用基于P-R曲线面积的AP计算法（区别于KITTI的11/40点插值近似法），同时创新性引入朝向感知的加权平均精度APH指标。APH在传统*AP*基础上，通过目标朝向误差的惩罚系数修正评分权重，有效量化算法对车辆行驶方向等关键运动参数的预测精度。该设计使评估体系既保持基础检测性能的可比性，又强化对自动驾驶决策关键要素的评估针对性。车辆、行人和骑行者类别的*IoU*阈值分别为0.7、0.5和0.5。上述2个指标考虑2个难度级别：1级（L1，具有超过5个LiDAR点的框盒Box）和2级（L2，具有至少1个LiDAR点的框盒Box）。

3.2 实验过程

3.2.1 KITTI数据集

在KITTI数据集下基于体素（柱）的检测方法中，体素大小[0.16 m, 0.16 m, 4 m]和[0.32 m, 0.32 m, 4 m]是典型的2种尺寸设置[22,34,45]。为了便于表述，本研究将前面2种体素大小简称为*V*=0.16和*V*=0.32。诱导集注意力块中的隐藏码*L*的长度通常设置为8或者16[45,76]，PointSetFormer和VoxelFormer都使用了隐藏码分别称为局部隐藏码L_l和全局隐藏码L_g。本研究分别在*V*=0.16和*V*=0.32两种体素尺寸下测试了不同局部隐藏码L_l和全局隐藏码L_g组合对VoxTNT检测性能的影响。

（1）模型设置

针对不同检测目标（汽车、行人、骑行者）设置差异化的体素参数与隐藏码，并迭代4层PointSetFormer网络。关键参数如表2所示。

模型结构配置：PointSetFormer层迭代4次（VoxelFormer同设置），体素沿X/Y轴逐层加倍（如第1层*V*=0.32 → 第2层*V*=0.64），特征维度逐层为16、32、64、128，位置编码（PE）带宽长度为64。

点的初始状态是一个7维的特征（$x, y, z, r, \delta x, \delta y, \delta z$），并使用MLP将其映射到$R^{16}$已完成初始特征编码。其中$(x, y, z)$为点的3D坐标，$r$为反射强度值，$(\delta x, \delta y, \delta z)$为点坐标与体素区域内每个点坐标平均值之差。

（2）训练和推理

训练和推理的相关参数配置如表3所示。

本研究将Second[22]中的锚定设置应用在单阶段模型中。对于两阶段模型，本研究分别应用Voxel R-CNN[53]和Part-A^2-free[23]的锚定设置。需要指出的是，无论是单阶段还是两阶段检测，汽车、行人和骑行者检测最终的anchor大小和匹配标准遵循数据集官方标准。对于其他参数设置，本研究遵循OpenPCDet[89]工具箱的默认设置。

3.2.2 WOD数据集

针对WOD数据集的超大规模特性（其数据规模显著超越KITTI数据集达20倍以上，单帧点云的

表2 模型参数配置

Tab. 2 Model parameter configuration

检测类别	体素大小	局部和全局隐藏码(单阶段)	点云范围	其他配置
汽车	0.32	KITTI: (8, 16)	KITTI: *X*: [0 m, 69.12 m], *Y*: [−39.68 m, 39.68 m], *Z*: [−3 m, 1 m] WOD: *X*: [−74.24 m, 74.24 m], *Y*: [−74.24 m, 74.24 m], *Z*: [−2 m, 4 m]	在KITTI中，使用前景图像分割范围
行人	KITTI: 0.16 WOD: 0.32	KITTI:(8, 8)	KITTI: *X*: [0 m, 69.12 m], *Y*: [−39.68 m, 39.68 m], *Z*: [−3 m, 1 m] WOD: *X*: [−74.24 m, 74.24 m], *Y*: [−74.24 m, 74.24 m], *Z*: [−2 m, 4 m]	在KITTI中，使用前景图像分割范围
骑行者	0.32	KITTI:(16, 16)	KITTI: *X*: [0 m, 69.12 m], *Y*: [−39.68 m, 39.68 m], *Z*: [−3 m, 1 m] WOD: *X*: [−74.24 m, 74.24 m], *Y*: [−74.24 m, 74.24 m], *Z*: [−2 m, 4m]	在KITTI中，使用前景图像分割范围

表3 训练参数配置

Tab. 3 Training parameter configuration

检测类型	单阶段检测	两阶段检测
适用任务	通用检测	汽车/骑行者检测（结合 Voxel R-CNN） 行人检测（结合 Part-A2-free）
GPU硬件	1×RTX 4090	1×RTX 4090
训练周期	120 epochs	汽车/骑行者：110 epochs 行人：100 epochs
优化策略	Adam	沿用对应方法（Voxel R-CNN / Part-A2-free）
学习率	初始 0.015，单周期衰减	初始 0.015，单周期衰减
动量阻尼范围	[0.85, 0.95]	[0.85, 0.95]
批大小	2	2
权重衰减	0.01	0.01
隐藏码设置	–	KITTI:(16, 8) WOD: (4, 4)
其他参数	–	汽车/骑行者：其他参数与 Voxel R-CNN[47]一致 行人：其他参数与 Part-A2-free[58]一致

空间覆盖范围更是达到KITTI的6倍），本研究通过以下策略实现实验设计的优化平衡：首先，在模型训练阶段严格遵循OpenPCDet[89]框架的标准实验协议，采用分层采样策略选取约20%的训练样本进行参数优化，既保证数据分布的完整性又有效控制计算成本；其次，在性能验证环节，严格遵循其双难度等级（L1/L2）评估协议，采用平均精度（AP）和精度-航向角联合指标（APH）作为核心评价指标，通过官方验证集的全量测试确保结果的可比性与可信度。

特别地，在方法验证层面引入单阶段检测约束机制：将VoxTNT方法的实验设计限定于单阶段检测架构，避免采用两阶段检测范式可能引发的计算与存储资源超载风险。这种层次化实验设计策略既满足大规模点云数据处理的效率要求，又为3D目标检测算法在资源受限环境下的可行性研究提供了参考。

在WOD数据集上，本研究分别选择沿X、Y、Z轴落在[−74.24 m, 74.24 m]、[−74.24 m, 74.24 m]、[−2 m, 4 m]范围内的LiDAR点云，体素尺寸设置为[0.32, 0.32, 6.0]。为了便于表述，将体素尺寸简称为V=0.32。诱导集注意力块中的隐藏码L的长度通常设置为8或者16[45]，但鉴于实验条件限制，本研究对VoxTNT的局部隐藏码L_l和全局隐藏码L_g统一设置为4。

（1）模型设置

在车辆、行人和骑行者检测中，第一PointSetFormer层的体素大小为V=0.32，每个体素采样24个点，迭代4次，体素大小沿X轴和Y轴逐层加倍，每层对应的特征维度分别为16、32、64和128。每个PointSetFormer块包含编码器、VoxelFormerFFN中间层和解码器。

VoxelFormerFFN也包含4次VoxelFormer算子迭代。与Set transformer[76]和VoxSeT[45]一样，本研究将PointSetFormer和VoxelFormer中PE模块的带宽长度设置为64。

（2）训练和推理

使用1个RTX 4090 GPU进行了65个周期的端到端训练，并使用Adam策略进行优化。学习率设置为0.001 5，随着单周期策略而衰减，动量的阻尼范围为[0.85, 0.95]。此外，批大小和权重衰减分别设置为2和0.01。

锚框（anchor）设置、后处理等其他参数设置，本研究遵循OpenPCDet[89]工具箱的默认设置。

3.3 KITTI数据集对比实验及结果分析

3.3.1 单阶段3D检测结果对比实验

本研究针对当前3D目标检测领域存在的三大评估瓶颈问题展开系统分析：① 很多性能较强的方法大多依赖多模态数据融合，显著增加硬件部署成本；② 检测类别单一化，很多研究仅关注某一类目标进行检测，如汽车检测；③ 评估指标异质性，不同工作采用11或40个召回阈值评估精度（AP），导致跨研究可比性降低。为建立公平评估基准，本实验设计遵循2个核心原则：首先，在KITTI验证集上统一采用11个召回阈值的标准AP计算协议；其次，严格筛选仅使用点云且同时支持汽车（Car）、行

人(Pedestrian)、骑行者(Cyclist)3类检测的单阶段基线方法进行对比。竞争对手主要是VoxelNet[10]、SECOND[22]、PointPillars[34]和VoxSeT[45],因为它们分别代表3类被广泛使用的特征学习网络。其中,VoxelNet是体素法的先驱,SECOND是体素法的重要改进(引入稀疏卷积大幅提升计算效率),PointPillars是基于柱的先驱方法,VoxSeT是集合到集合变换的先驱方法。

VoxTNT单阶段检测结果与上述单阶段基线方法的对比如表4所示。从表中实验结果对比发现:在简单场景(如只包含汽车的场景)3D检测中,可以采用体素尺寸、局部隐藏码和全局隐藏码的最佳组合,分别检测汽车、行人和骑行者,即:对于汽 车检测,采用V=0.32、L_l=16、L_g=8;对于行人检测,采用V=0.16、L_l=8、L_g=8;对于骑行者检测,采用V=0.32、L_l=16、L_g=16。表中实验结果表明,在汽车、行人和骑行者的3D检测中获得了具有竞争力的性能。其中,在行人检测中优势显著,在简单、中等和困难3个级别中分别领先最先进的基线方法VoxSeT 5.14、4.82和2.81。在简单和中等难度级别的骑行者检测中分别领先最先进的基线方法VoxSeT 1.59和1.96。尽管汽车检测精度略低于当前最优方法,但仍在0.5误差范围内保持第一梯队性能。

针对复杂多目标场景提出2种优化方案,为实际应用提供灵活选择空间:

(1)统一参数模型:采用平衡配置(V=0.32、L_l=16、L_g=16)。在KITTI数据集上取得显著效果:在行人和骑行者检测中实现了全面领先,例如在中等和困难级别的行人检测中分别领先最先进的基线方法VoxSeT 2.47和1.82;在简单和中等级别的骑行者检测中分别领先最先进的基线方法VoxSeT 1.59和1.96;在综合指标*mAP*中达到70.41,领先最先进的基线方法VoxSeT 0.8。汽车检测虽略低于最先进基线方法SECOND,但总体差距都在1以内,仍保持第一梯队性能。

(2)多模型融合方案:在算力允许时,可针对不同类别分别训练最优参数模型(汽车V=0.32、L_l=16,L_g=8、行人V=0.16、L_l=8、L_g=8、骑行者V=0.32、L_l=16、L_g=16),通过并行检测提升整体精度。实验表明该方案可使综合指标*mAP*达到71.56较次优方法VoxSeT[45]提升1.89。

上述实验结果及分析表明,本研究提出的方法实现了具有竞争力的3D检测性能,因为在保证汽车检测的前提下,实现了小目标检测性能的显著提升。本研究通过多尺度自适应检测机制,在KITTI基准上实现了检测精度与目标尺度的有效平衡。在维持汽车检测高精度的同时,显著优化了小目标检测性能。这种提升主要得益于局部–全局特征编码的协同作用,局部和全局感受野的均衡化使网络能够学习目标物理尺寸差异,在复杂场景中实现跨尺度目标的均衡检测。

3.3.2 两阶段3D检测结果对比实验

为了进一步验证VoxTNT框架中3D特征学习的有效性,本研究通过将其扩展至两阶段3D检测架构进行了系统性实验验证。具体而言,本研究采用VoxTNT与Voxel R-CNN[53]相结合的架构实现汽车和骑行者检测,同时结合Part-A^2-free[23]框架完成行人检测。将KITTI测试集中两阶段检测结果提交至KITTI官方并返回测试结果如表5所示。在测试集上,竞争对手是在KITTI基准官方排行榜中2023—2025年基于点云的至少涵盖2个类别检测

表4 VoxTNT与经典的基线方法在KITTI验证集上单阶段3D检测对比

Tab. 4 Comparison of VoxTNT and classic baseline methods for single-stage 3D detection on KITTI validation set

方法	汽车			行人			骑行者			*mAP*
	简单	中等	困难	简单	中等	困难	简单	中等	困难	
VoxelNet[10]	81.97	65.46	62.85	57.86*	53.42*	48.87*	67.17	47.65	45.11	58.93
SECOND[22]	**88.61**	**78.62**	**77.22**	56.55	52.98	47.73	80.58*	67.15*	63.10*	68.06*
PointPillars[34]	86.46*	77.28	74.65	57.75	52.29	47.90	80.04	62.61	59.52	66.50
VoxSeT[45]	88.45	78.48	77.07	60.62	54.74	50.39	84.07	68.11	65.14	69.67
VoxTNT(最佳组合)	88.52	78.20*	77.03*	**65.76**	**59.56**	**53.20**	**85.66**	**70.07**	**66.04**	**71.56**
增量	−0.09	−0.42	−0.35	+5.14	+4.82	+2.81	+1.59	+1.96	+0.90	
VoxTNT(统一参数模型)	87.58	77.83	76.48	**61.12**	**57.21**	**52.21**	**85.66**	**70.07**	**66.04**	**70.47**
增量	−1.03	−0.79	−0.90	+0.50	+2.47	+1.82	+1.59	+1.96	+0.90	

注:排前3的方法分别以粗体、下划线和*号突出。

表5 VoxTNT与基于点云的主要方法在KITTI测试集上两阶段3D检测对比

Tab. 5 Comparison of VoxTNT and Point Cloud based 3D Detection Methods in Two Stages on KITTI test set

方法	汽车			行人			骑行者			*mAP*
	简单	中等	困难	简单	中等	困难	简单	中等	困难	
PV-RCNN++[51]	90.14	81.88	77.15*	–	–	–	82.22	67.33	60.04	–
DFAF3D[40]	88.59	79.37	72.21	47.58	40.99	37.65	82.09	65.86	59.02	63.71
BSAODet[41]	88.89	81.74	77.14	**51.71**	43.63	**41.09**	82.65*	67.79	60.26	66.10
PG-RCNN[42]	89.38*	82.13*	**77.33**	–	–	–	82.77	67.82*	61.25	–
PASS-PV-RCNN++[43]	87.65	81.28	76.79	47.66	41.95	38.90*	80.43	68.45	60.93*	64.89*
PV-GNN_Cyc&Ped[52]	–	–	–	48.78*	42.00*	36.91	78.58	62.54	55.28	–
VoxTNT	**90.51**	81.74*	77.22	50.92	**43.87**	40.53	**83.37**	**68.53**	**62.13**	**66.54**
增量	+0.37	−0.39	−0.11	−0.79	+0.24	−0.56	+0.60	+0.08	+0.88	+0.44

注：排前3的方法分别以粗体、下划线和*号突出。

的最先进方法，包括PV-RCNN++[51]、DFAF3D[40]、BSAODet[41]、PG-RCNN[42]、PASS-PV-RCNN++[43]和PV-GNN[52]。实验结果表明：

（1）在骑行者检测任务中，本方法在所有难度级别（简单、中等、困难）均取得最优结果；

（2）在汽车检测（简单级别）和行人检测（中等级别）任务中保持领先地位；

（3）综合评估指标方面，本方法以66.54的mAP值（所有类别平均精度）位居榜首，较次优方法PV-GNN[52]提升0.44 mAP。

值得注意的是，这种跨类别的性能优势不仅体现在单一目标检测任务中，更反映在不同尺寸、不同类型目标的综合检测能力上。这种卓越表现验证了本研究提出的局部–全局特征协同学习机制的有效性：通过创新性地在体素邻域内部采用约简的交叉注意力进行局部特征聚合，同时在跨体素区域间实施约简的注意力机制来捕获全局上下文信息，该方法成功实现了多尺度感受野的动态平衡。实验结果表明，这种双路径特征学习融合策略显著提升了算法在复杂城市场景中应对多尺度目标检测和跨类别识别挑战的能力，特别是在处理交通密集区域的遮挡（中等难度级别）目标和小尺度目标时表现出更强的鲁棒性。此外，该特征学习框架具备良好的架构兼容性，其与不同检测器的结合均能产生正向增益，这为未来3D目标检测研究提供了新的技术路径。

3.4 WOD数据集对比实验及结果分析

为验证VoxTNT方法在WOD上的泛化性能，本研究设计了对比实验框架：选取与本研究训练数据规模相当（均采用20%训练集）的经典单阶段3D检测模型（CenterPoint[25]、SECOND[22]和PointPillars[34]）作为基准，实验数据严格遵循OpenPCDet[89]官方发布的基准测试结果。如表6（AP指标）和表7（APH指标）所示，评估体系包含3个目标类别在2个难度等级下的6项核心指标——车辆（Vec_L1/Vec_L2）、行人（Ped_L1/Ped_L2）、骑行者（Cyc_L1/Cyc_L2），每项指标均包含平均精度（AP）与航向角加权精度（APH）双重度量。实验分析表明：

（1）小目标检测优势：在行人与骑行者检测任务中，全面领先SECOND[22]和PointPillars[34]基线方法，例如L2级别的行人和骑行者检测达到63.86 AP（分别领先SECOND和PointPillars基线11.6%和9.8%）和63.35 AP（分别领先SECOND和PointPillars基线15.24%和19.1%）。

表6 VoxTNT与经典基线方法在WOD验证集上的单阶段3D检测对比(AP指标)

Tab. 6 Comparison of VoxTNT and classical baseline methods for single-stage 3D detection on WOD validation set (AP metric)

方法	Vec_L1	Vec_L2	Ped_L1	Ped_L2	Cyc_L1	Cyc_L2	*mAP*
SECOND[22]	70.96	62.58	65.23	57.22	57.13	54.97	61.35
PointPillars[34]	70.43	62.18	66.21	58.18	55.26	53.18	60.91
CenterPoint[25]	**71.33**	**63.16**	72.09	**64.27**	**68.68**	**66.11**	**67.61**
VoxTNT	69.74	61.32	**72.49**	63.86	65.76	63.35	66.09

注：排前2的方法分别以粗体和下划线突出。

表7 VoxTNT与经典基线方法在WOD验证集上的单阶段3D检测对比(APH指标)

Tab. 7 Comparison of VoxTNT and classical baseline methods for single-stage 3D detection on WOD validation set (APH metric)

方法	Vec_L1	Vec_L2	Ped_L1	Ped_L2	Cyc_L1	Cyc_L2	*mAPH*
SECOND[22]	70.34	62.02	54.24	47.49	55.62	53.53	57.21
PointPillars[34]	69.83	61.64	46.32	40.64	51.75	49.80	53.33
CenterPoint[25]	**70.76**	**62.65**	**65.49**	**58.23**	**67.39**	**64.87**	**64.90**
VoxTNT	69.19	60.83	62.29	54.71	64.46	62.09	57.21

注:排前2的方法分别以粗体和下划线突出。

(2)跨类别一致性:虽然车辆检测指标略差于CenterPoint[25]、SECOND[22]和PointPillars[34] 3个基线方法,但差距幅度控制在3%以内,并且在3个类别的综合指标*mAP*和*mAPH*领先SECOND[22]和PointPillars[34],例如综合指标*mAP*达到66.09分别超越SECOND和PointPillars基线7.7%和8.5%,表明方法在提升小目标检测的基础上具备跨类别的稳定检测能力。

本研究的实验结果从多维角度验证了方法的泛化性能优势,具体体现在2个核心维度:① 基于空间稀疏性先验设计的非空体素域内–域间双模注意力机制,创新性地构建了局部几何细节与全局上下文特征的协同学习范式。通过动态调整注意力场的空间分布,实现跨尺度目标感受野的动态适配,尤其是增强了典型小尺度目标(如骑行者)的几何表征能力;② 提出的统一特征表征框架通过参数共享机制与自适应特征融合策略,有效解决了传统方法因目标类型特异性导致的参数冗余问题。这些实验结果充分验证了本研究均衡化局部和全局感受野调控机制的理论优越性,其创新的局部–全局特征协同策略为3D点云目标检测提供了新的范式,特别是在复杂交通场景的多尺度目标感知方面展现出显著的工程应用价值。

3.5 消融实验

3.5.1 KITTI数据集

本研究主要在KITTI数据集中做了一系列消融实验,以了解VoxTNT中不同部分的作用。

(1)局部和全局隐藏码不同组合的影响

本研究在KITTI验证集上针对单阶段3D检测中体素尺寸(Voxel Size, V)与隐藏码(Local/Global Latent Codes, L_l/L_g)的参数耦合效应展开系统性分析。平均精度AP基于11个召回阈值生成,如表8和表9所示,通过极差分析(Range)、平均值(Avg)和平均绝对偏差(Mean Absolute Deviation, MAD)统计指标,揭示了不同参数组合对多类目标检测的差异化影响规律。

当V=0.16的时候:对于汽车检测,从极差和平均绝对误差值上看,隐藏码组合的敏感性较低,表明相对任意的局部与全局隐藏码的协同可有效捕捉车辆宏观结构特征,但当L_l=8、L_g=16时效果相对较好;对于行人检测,从极差和平均绝对误差值上

表8 KITTI验证集上不同隐藏码组合的影响(V=0.16)

Tab. 8 The impact of different combinations of latent codes on KITTI validation set (V=0.16)

体素尺寸	L_l	L_g	汽车			行人			骑行者		
			简单	中等	困难	简单	中等	困难	简单	中等	困难
V=0.16	8	8	87.58	76.16	70.32	65.76	59.56	53.20	84.43	66.08	63.86
	8	16	88.57	77.12	71.15	62.42	56.20	50.95	82.34	65.97	62.43
	16	8	86.92	75.74	70.50	63.79	57.90	52.63	81.23	64.14	60.69
	16	16	87.60	76.15	70.76	63.16	57.17	51.97	86.00	68.99	64.70
	Range		1.65	1.38	0.83	3.34	3.36	2.25	4.77	4.85	4.01
	Avg		87.67	76.29	70.68	63.78	57.71	52.19	83.50	66.30	62.92
	MAD		0.45	0.41	0.27	0.99	1.02	0.73	1.72	1.35	1.36
	最佳组合	汽车(L_l=8、L_g=16) 行人(L_l=8、L_g=8) 骑行者(L_l=16、L_g=16)	88.57	77.12	71.15	65.76	59.56	53.20	86.00	68.99	64.70

表9　KITTI验证集上不同隐藏码的影响(V=0.32)

Tab. 9　The impact of different combinations of latent codes on KITTI validation set (V=0.32)

体素尺寸	L_l	L_g	汽车			行人			骑行者		
			简单	中等	困难	简单	中等	困难	简单	中等	困难
V=0.32	8	8	88.20	78.20	76.92	60.21	55.68	51.07	84.93	66.51	63.51
	8	16	88.52	78.20	77.03	58.01	53.29	48.93	85.06	71.03	64.95
	16	8	87.81	78.06	76.66	62.27	56.58	51.98	81.97	64.90	62.24
	16	16	87.58	77.83	76.48	61.12	57.21	52.12	85.66	70.07	66.04
	Range		0.94	0.37	0.55	4.26	3.92	3.19	3.69	6.13	3.80
	Avg		88.03	78.07	76.77	60.40	55.69	51.03	84.41	68.12	64.19
	MAD		0.33	0.13	0.20	1.29	1.21	1.05	1.22	2.42	1.31
	最佳组合	汽车(L_l=8、L_g=16) 行人\骑行者(L_l=16、L_g=16)	88.52	78.20	77.03	61.12	57.21	52.12	85.66	70.07	66.04

看，呈现显著参数敏感性，当L_l=8、L_g=8时效果最好，反映对称隐藏码结构对小尺度目标细部特征的表征优势，但其他组合下的结果也具有一定竞争力；对于骑行者检测，从极差和平均绝对误差值上看，呈现显著参数敏感性，当L_l=16、L_g=16时效果最佳，验证了深层隐藏码对复杂运动姿态建模的有效性。因此，当V=0.16的时候，本研究建议：针对只有汽车的简单场景，采用L_l=8、L_g=16的组合会得到相对较好的结果；针对只有行人的简单场景，采用L_l=8、L_g=8的组合会得到相对较好的结果；针对只有骑行者的简单场景，采当L_l=16、L_g=16的组合可以获得最好的结果；针对具有多尺寸、多类别目标的复杂场景，可以采用L_l=16、L_g=16组合。

当V=0.32的时候，从极差和平均绝对误差值上看，汽车检测保持参数低敏感性，而行人与骑行者检测呈现显著参数敏感性，进一步表明小目标检测更依赖隐藏码的长度与组合。根据对结果对比，本研究建议：针对只有行人的简单场景，采用L_l=8、L_g=16的组合会得到最佳效果；对行人和骑行者进行检测，采用L_l=16、L_g=16的组合会得到相对较好的结果；针对具有多尺寸、多类别目标的复杂场景，可以采用当L_l=16、L_g=16组合。

（2）体素尺寸的影响

本文将V=0.16和V=0.32下的最优检测方案和平均值进行对比如表10所示。对于汽车检测，无论是平均值，还是最优检测方案，体素尺寸V=0.32下的检测效果都更好，但总体上属于同一级别的检测；与汽车检测相反，体素V=0.16下的行人检测效果都更好，且差距较大；对于骑行者检测，在体素尺寸V=0.32下的检测效果更好，但差距相对行人检测较小。

因此，结合局部和全局隐藏码组合方案，最终本研究建议：对于汽车检测，应该采用V=0.32、L_l=8，L_g=16；对于行人，应该采用V=0.16、L_l=8、L_g=8；对于骑行者检测，应该采用V=0.32、L_l=16、L_g=16；针对具有多尺寸、多类别目标的复杂场景，可以采用V=0.32、L_l=16、L_g=16的组合。

针对实际应用场景提出两级部署策略：① 资源受限场景：采用通用配置实现多目标联合检测；

表10　在KITTI验证集上最佳检测方案和平均值对比

Tab. 10　Comparison of best detection scheme and average value on KITTI validation set

体素尺寸	对比项	汽车			行人			骑行者		
		简单	中等	困难	简单	中等	困难	简单	中等	困难
V=0.16	最佳组合	88.57	77.12	71.15	65.76	59.56	53.20	86.00	68.99	64.70
V=0.32	最佳组合	88.52	78.20	77.03	61.12	57.21	52.12	85.66	70.07	66.04
最佳: 汽车\骑行者(V=0.32) 行人(V=0.16)		88.52	78.20	77.03	65.76	59.56	53.20	85.66	70.07	66.04
V=0.16	Avg	87.67	76.29	70.68	63.78	57.71	52.19	83.50	66.30	62.92
V=0.32	Avg	88.03	78.07	76.77	60.40	55.69	51.03	84.41	68.12	64.19

② 高算力场景：建立多模型级联系统，通过单类别专用检测器组合提升检测性能。

上述分析表明，通过动态调节体素粒度与隐藏码长度的组合关系，可使检测系统适应不同场景需求，这为自动驾驶感知模块的软硬件协同设计提供了重要理论依据。

（3）PointSetFormer和VoxelFormerFFN组件的影响

为验证3D特征学习框架中核心组件的协同作用，本研究在KITTI验证集上设计了系统的消融实验（表11），由于在实际应用的时候，VoxelFormer-FFN属于PointSetFormer的内置模块，所以本研究无需对VoxelFormerFFN进行单独的消融实验。实验聚焦于PointSetFormer架构的三级特征学习机制——编码器（局部特征提取）、VoxelFormerFFN中间件（全局特征交互）与解码器（特征融合）的耦合效应，揭示以下关键发现：① 全组件配置：在V=0.16/V=0.32的体素配置下，同时使用PointSetFormer和VoxelFormerFFN组件能明显提高检测性能，尤其是在行人和骑行者检测中效果更明显，这说明PointSetFormer组件的编码器使用约简的交叉注意力机制学习体素区域内的局部特征、PointSetFormer组件的中间件VoxelFormerFFN使用约简的交叉注意力机制实现跨体素的全局特征学习和PointSetFormer组件的解码器实现局部与全局特征融合的点云3D特征学习方式是有效的；② 移除VoxelFormerFFN：只使用PointSetFormer组件，但不使用VoxelFormerFFN组件，会明显降低检测精度，尤其是在行人和骑行者检测中下降幅度更大（例如，在体素大小V=0.16下，中等难度级别的行人和骑行者检测精度分别下降了2.89%和4.75%），这说明所提出的VoxelFormerFFN中间件通过均衡全局感受野实现的跨体素信息传递与交换对提升检测性能至关重要，并且对小目标的检测性能提升更明显；③ 全组件移除：PointSetFormer和VoxelFormerFFN组件都没有使用的情况下，在行人和骑行者检测中下降特别显著（例如，在体素大小V=0.16下，中等难度级别的行人和骑行者检测分别下降了10.80%和10.04%），这进一步说明本研究提出的均衡化感受野机制下局部–全局特征协同策略的有效性，尤其影响小目标几何表征。

根据消融实验可知：① PointSetFormer和VoxelFormerFFN这2个关键组件对不同尺寸、不同类别目标的3D检测性能都有显著的提升；② 与骑行者和行人检测相比，汽车检测对局部–全局特征协同的依赖性相对较低。

3.5.2 WOD数据集

由于受硬件显存限制（只有单张RTX 4090 GPU），无法进行局部/全局隐藏码的各种不同组合（如8/8、8/16等）影响，也无法测试更多体素尺寸下的检测结果，因此，本研究着重研究了VoxTNT中关键组件PointSetFormer和VoxelFormerFFN对3D检测的性能影响（表12）。由于在实际应用的时候，VoxelFormerFFN属于PointSetFormer的内置模块，所以无需对VoxelFormerFFN进行单独的消融实验。实验聚焦于PointSetFormer架构的三级特征学习机制——编码器（局部特征提取）、VoxelFormer-FFN中间件（全局特征交互）与解码器（特征融合）的耦合效应，揭示以下关键发现：

（1）全组件配置：同时使用PointSetFormer和VoxelFormerFFN组件与都不使用相比，大幅提升了车辆、行人和骑行者的检测精度，这说明本研究设

表11 在KITTI验证集上VoxTNT中各组件消融结果比较

Tab. 11 Comparison of ablation results of various components in VoxTNT on KITTI validation set

体素尺寸	组件		汽车			行人			骑行者		
	PointSet Former	Voxel Former FNN	简单	中等	困难	简单	中等	困难	简单	中等	困难
V=0.16	✓	✓	**88.57**	77.12	71.15	**65.76**	**59.56**	**53.20**	**86.00**	**68.99**	**64.70**
	✓		87.82	77.61	76.26	65.16	57.84	51.90	85.09	65.71	63.33
	×	×	87.79	77.30	71.42	59.07	53.13	48.40	81.55	62.06	58.78
V=0.32	✓	✓	**88.52**	**78.20**	**77.03**	**61.12**	**57.21**	**52.12**	**85.66**	**70.07**	**66.04**
	✓		87.59	77.79	76.57	59.89	54.57	48.82	82.50	63.62	62.28
	×	×	87.04	76.77	73.30	53.76	46.69	42.81	81.25	63.99	60.21

注：加粗的表示最佳值。

表12　关键组件在WOD验证集上的消融结果对比

Tab. 12　Comparison of ablation results of key components in VoxTNT on WOD validation set

组件		Vec_L1 (*AP*)	Vec_L1 (*APH*)	Vec_L2 (*AP*)	Vec_L2 (*APH*)	Ped_L1 (*AP*)	Ped_L1 (*APH*)
PointSetFormer	VoxelFormerFFN						
✓	✓	**69.74**	**69.19**	**61.32**	**60.83**	72.49	62.29
✓		69.63	69.10	61.23	60.75	**72.66**	**62.45**
×	×	64.84	64.21	56.72	56.15	63.39	46.12
PointSetFormer	VoxelFormerFFN	Ped_L2 (*AP*)	Ped_L2 (*APH*)	Cyc_L1 (*AP*)	Cyc_L1 (*APH*)	Cyc_L2 (*AP*)	Cyc_L2 (*APH*)
✓	✓	63.86	54.71	**65.76**	**64.46**	**63.35**	**62.09**
✓		**64.03**	**54.86**	65.38	63.99	62.99	61.65
×	×	54.99	39.91	56.86	53.86	54.70	51.83

注：加粗数值表示最佳值。

计的均衡化局部和全局感受野实现的局部–全局特征协同学习策略是有效的。

同时使用PointSetFormer和VoxelFormerFFN组件在目标检测任务中展现出差异化性能表现。实验数据表明，该设计提升了车辆与骑行者类别的检测精度，但在行人检测中出现轻微的衰减。这种现象可归因于全局隐编码长度缩减至4时，导致行人细粒度运动模式的表征能力下降。

（2）移除VoxelFormerFFN：只使用PointSetFormer组件，但不使用VoxelFormerFFN组件，在目标检测任务中展现出差异化性能表现。实验数据表明，未使用VoxelFormerFFN组件会降低车辆与骑行者类别的检测精度，但在行人检测中出现轻微的提升。这种现象可归因于全局隐编码长度缩减至4时，导致行人细粒度运动模式的表征能力下降。

（3）全组件移除：PointSetFormer和VoxelFormerFFN组件都没有使用的情况下，检测效果出现大幅度下降（例如，在L2难度级别的车辆、行人和骑行者检测的*AP*分别下降了7.5%、13.9%和13.7%。），这进一步说明本研究提出的均衡化局部感受野和可变全局感受野机制下局部–全局特征协同策略的有效性，尤其影响小目标几何表征。

根据上述WOD中的消融实验可知：① PointSetFormer和VoxelFormerFFN 2个关键组件对不同尺寸、不同类别目标的3D检测性能都有显著的提升，但VoxelFormerFFN组件对行人检测精度产生了微小的负影响，这主要是实验条件限制导致参数压缩无法实现更优结果所致；② 与汽车检测相比，骑行者和行人检测对均衡化的局部特征学习的依赖性相对较高，这进一步说明了本研究的方法能够兼顾小目标的检测，从而实现跨尺度的3D检测。

4　结论与展望

4.1　结论

本研究聚焦城市自动驾驶场景下基于点云的三维目标检测方法，针对现有方法在跨尺度表征一致性方面仍然存在的挑战，基于局部–全局特征协同学习架构，研究了融合双重Transformer的3D目标检测方法VoxTNT，并通过实验分别验证了其有效性与泛化能力，主要研究结论如下：

（1）VoxTNT通过构建局部–全局特征协同学习策略，实现了点云局部几何特征与全局上下文信息的协同学习，其体素分区与矢量化并行计算显著降低了计算复杂度（从$O(N^3)$降至$O(N)$），同时避免了传统体素方法中零填充与信息丢失的问题。有效地兼顾了效率与精度的平衡，并缓解了体素特征编码方法在长尾分布场景下小目标检测不足的问题。实验表明，VoxTNT在KITTI数据集的小目标检测任务上实现了较好的改进，例如：在单阶段检测下，中等难度级别行人检测的精度*AP*值达到59.56，较SECOND基线提高约12.4%；扩展至两阶段检测框架，实现汽车、行人及骑行者检测的综合指标*mAP*达到66.54，排名第一。此外，在WOD数据集中充分验证了方法的泛化能力，例如综合指标*mAP*达到66.09分别超越SECOND和PointPillars基线7.7%和8.5%，表明方法在提升小目标检测的基础上具备跨类别的稳定检测能力。

（2）均衡全局感受野设计。本研究创新性地提出将非空体素建模为超点集结构，并利用简约的交叉注意力实现全局感受野的均衡化。该方法突破传统前馈神经网络（FFN）与3D稀疏卷积的固定感受野局限，通过建立跨尺度的长程上下文依赖关

系,实现点云特征的各尺度感受野均衡。具体而言,通过自注意力权重的自适应分配机制,增强全局远近空间关联建模能力,有效解决多尺度目标的特征表征失衡问题。在KITTI数据集中的消融实验表明,未使用VoxelFormerFFN组件(均衡化的全局感受野),会明显降低检测精度,尤其是在行人和骑行者检测中下降幅度更大(例如,在体素大小 V=0.16下,中等难度级别的行人和骑行者检测精度分别下降了2.89%和4.75%),这说明通过均衡全局感受野实现的跨体素信息传递与交换对提升检测性能至关重要,并且对小目标的检测性能提升更明显。此外,在WOD上的消融实验显示未使用VoxelFormerFFN组件会降低车辆与骑行者类别的检测精度(例如L2级别的骑行者检测下降了0.36 AP),进一步验证了均衡全局感受野的设计是有效的。

4.2 展望

本研究在实验性能上表现优异,但受限于跨体素矢量化操作的高密度点云采样策略与双重注意力架构的设计,仍存在时延与内存效率的瓶颈。为此,围绕如何进一步降低模型复杂度、提高计算效能和提升模型自适应能力等方面,未来将进一步开展的后续研究工作包括:

(1)更轻量化的模型算法设计

复杂Transformer结构虽在特征提取精度上表现出色,但计算成本高昂,实时部署面临挑战。尽管体素化和约简交叉注意力机制已优化效率,仍需进一步探索轻量化设计。未来研究将聚焦于Transformer网络结构的精简优化,在降低内存占用的同时显著提升检测速度。例如,采用轻量化Transformer变体(如EfficientFormer)替代传统架构,并结合硬件感知训练(Hardware-Aware Training)实现端到端的性能优化。

(2)更复杂场景的泛化性验证

本研究实验主要基于KITTI和WOD数据集,尽管在复杂城市场景中展现了较强的鲁棒性,但对低光照、雨雾等恶劣天气条件下的点云稀疏性及噪声干扰问题尚未充分验证。这些场景中,点云数据的稀疏性和噪声特性会对模型的感知能力产生显著影响,而现有方法在这些方面的适应性仍需进一步优化。未来研究将探索结合环境感知模块(如气象条件感知)或通过数据增强与合成技术模拟恶劣天气场景,提升模型在极端环境下的鲁棒性与泛化能力。

(3)多模态融合的检测方法

本研究主要关注单一LiDAR点云数据的处理,尚未充分整合相机图像的纹理信息,这在一定程度上限制了模型在复杂场景下的感知能力。尽管LiDAR点云提供了精确的深度信息,但在遮挡严重或远距离目标检测中,点云数据的稀疏性可能导致关键特征丢失,进而影响检测精度。未来研究将致力于开发高效的多模态融合策略,通过结合相机图像的丰富纹理与颜色信息,弥补单一模态的不足。具体而言,可探索特征级融合方法,利用注意力机制动态调整LiDAR与相机数据的权重,增强对遮挡目标的识别能力;同时,决策级融合策略可通过多传感器协同优化检测结果,提升系统鲁棒性。

(本研究源代码可以在 https://github.com/yujianxinnian/VoxTNT下载)

AI使用说明:本文使用了AI技术对英文摘要进行润色。

利益冲突: Conflicts of Interest

所有作者声明不存在利益冲突。

All authors disclose no relevant conflicts of interest.

作者贡献: Author Contributions

郑强文和吴升参与方法设计;郑强文和魏婧卉参与实验设计;郑强文完成实验操作;郑强文和魏婧卉参与实验结果分析;郑强文完成论文初稿;郑强文和吴升参与论文的写作和修改;吴升提供资助经费。所有作者均阅读并同意最终稿件的提交。

ZHENG Qiangwen and WU Sheng contributed to methodology design; ZHENG Qiangwen and WEI JInghui contributed to experimental design; ZHENG Qiangwen performed the experiments; ZHENG Qiangwen and WEI Jinghui analyzed the experimental results; ZHENG Qiangwen drafted the manuscript; ZHENG Qiangwen and WU Sheng contributed to writing and revising the manuscript; WU Sheng provided funding. All the authors have read the last version of manuscript and consented for submission.

参考文献(References):

[1] Mao J G, Shi S S, Wang X G, et al. 3D object detection for autonomous driving: A comprehensive survey[J]. International Journal of Computer Vision, 2023, 131(8): 1909-1963. DOI:10.1007/s11263-023-01790-1

[2] Qian R, Lai X, Li X R. 3D object detection for autono-

mous driving: A survey[J]. Pattern Recognition, 2022,130: 108796. DOI:10.1016/j.patcog.2022.108796

[3] Zamanakos G, Tsochatzidis L, Amanatiadis A, et al. A comprehensive survey of LIDAR-based 3D object detection methods with deep learning for autonomous driving [J]. Computers & Graphics, 2021, 99: 153-181. DOI: 10.1016/j.cag.2021.07.003

[4] Lang B, Li X, Chuah M C. BEV-TP: End-to-end visual perception and trajectory prediction for autonomous driving[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(11): 18537-18546. DOI: 10.1109/TITS.2024.3433591

[5] Zhang A, Eranki C, Zhang C, et al. Toward robust robot 3-D perception in urban environments: The UT campus object dataset[J]. IEEE Transactions on Robotics, 2024, 40: 3322-3340. DOI: 10.1109/TRO.2024.3400831

[6] Shreyas E, Sheth M H, Mohana. 3D object detection and tracking methods using deep learning for computer vision applications[C]//2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT). IEEE, 2021: 735-738. DOI:10.1109/rteict52294.2021.9573964

[7] 张尧,张艳,王涛,等.大场景SAR影像舰船目标检测的轻量化研究[J].地球信息科学学报,2025,27(1):256-270. [Zhang Y, Zhang Y, Wang T, et al. Lightweight research on ship target detection in large-scale SAR images[J]. Journal of Geo-information Science, 2025, 27(1): 256-270.] DOI:10.12082/dqxxkx.2025.240574

[8] 高定, 李明, 范大昭, 等. 复杂背景下轻量级SAR影像船舶检测方法[J]. 地球信息科学学报, 2024,26(11):2612-2625. [Gao D, Li M, Fan D Z, et al. A ship detection method from lightweight SAR images under complex backgrounds[J]. Journal of Geo-information Science, 2024,26 (11):2612-2625.] DOI:10.12082/dqxxkx.2024.230544

[9] Guo Y L, Wang H Y, Hu Q Y, et al. Deep learning for 3D point clouds: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(12): 4338-4364. DOI:10.1109/TPAMI.2020.3005434

[10] Zhou Y, Tuzel O. VoxelNet: End-to-end learning for point cloud based 3D object detection[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018:4490-4499. DOI:10.1109/CVPR.2018.00472

[11] Bello S A, Yu S S, Wang C, et al. Review: Deep learning on 3D point clouds[J]. Remote Sensing, 2020, 12(11): 1729. DOI:10.3390/rs12111729

[12] Song Z Y, Liu L, Jia F Y, et al. Robustness-aware 3D object detection in autonomous driving: A review and outlook[J]. IEEE Transactions on Intelligent Transportation Systems, 2024,25(11):15407-15436. DOI:10.1109/TITS.2024.3439557

[13] Yang B, Luo W J, Urtasun R. PIXOR: Real-time 3D object detection from point clouds[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018:7652-7660. DOI:10.1109/CVPR.2018.00798

[14] Chen X Z, Ma H M, Wan J, et al. Multi-view 3D object detection network for autonomous driving[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017:6526-6534. DOI:10.1109/CVP R.2017.691

[15] Liang M, Yang B, Wang S L, et al. Deep continuous fusion for multi-sensor 3D object detection[M]//Computer Vision - ECCV 2018. Cham: Springer International Publishing, 2018:663-678. DOI:10.1007/978-3-030-01270-0_39

[16] Liang M, Yang B, Chen Y, et al. Multi-task multi-sensor fusion for 3D object detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019:7337-7345. DOI:10.1109/CVPR.2019.00752

[17] Li J Z, Yang L, Shi Z, et al. SparseDet: Towards efficient multi-view 3D object detection via sparse scene representation[J]. Advanced Engineering Informatics, 2024, 62: 102955. DOI:10.1016/j.aei.2024.102955

[18] Chen Y Q, Li N Y, Zhu D D, et al. BEVSOC: Self-supervised contrastive learning for calibration-free BEV 3-D object detection[J]. IEEE Internet of Things Journal, 2024,11 (12):22167-22182. DOI:10.1109/JIOT.2024.3379471

[19] Yang L, Zhang X Y, Yu J X, et al. MonoGAE: Roadside monocular 3D object detection with ground-aware embeddings[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(11): 17587-17601. DOI: 10.1109/TITS.2024.3412759

[20] Kuang H W, Wang B, An J P, et al. Voxel-FPN: Multi-scale voxel feature aggregation for 3D object detection from LIDAR point clouds[J]. Sensors, 2020, 20(3): 704. DOI:10.3390/s20030704

[21] He C H, Zeng H, Huang J Q, et al. Structure aware single-stage 3D object detection from point cloud[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 11870-11879. DOI: 10.1109/CVPR42600.2020.01189

[22] Yan Y, Mao Y X, Li B. SECOND: Sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337. DOI: 10.3390/s18103337

[23] Shi S S, Wang Z, Shi J P, et al. From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021,43(8):2647-2664. DOI:10.1109/TPAMI.2020.2977026

[24] Ye M S, Xu S J, Cao T Y. HVNet: Hybrid voxel network for LiDAR based 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition

(CVPR). IEEE, 2020: 1628-1637. DOI: 10.1109/CVPR42600.2020.00170

[25] Yin T W, Zhou X Y, Krahenbuhl P. Center-based 3D object detection and tracking[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 11779-11788. DOI: 10.1109/cvpr46437.2021.01161

[26] Shi S S, Wang X G, Li H S. PointRCNN: 3D object proposal generation and detection from point cloud[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 770-779. DOI: 10.1109/cvpr.2019.00086

[27] Yang Z T, Sun Y N, Liu S, et al. 3DSSD: Point-based 3D single stage object detector[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020: 11037-11045. DOI: 10.1109/cvpr42600.2020.01105

[28] Qi C R, Liu W, Wu C X, et al. Frustum PointNets for 3D object detection from RGB-D data[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018:918-927. DOI:10.1109/CVPR.2018.00102

[29] Wang Z X, Jia K. Frustum ConvNet: Sliding Frustums to aggregate local point-wise features for amodal 3D object detection[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019: 1742-1749. DOI:10.1109/IROS40897.2019.8968513

[30] Yang Z T, Sun Y N, Liu S, et al. STD: Sparse-to-dense 3D object detector for point cloud[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 1951-1960. DOI:10.1109/iccv.2019.00204

[31] Charles R Q, Hao S, Mo K C, et al. PointNet: Deep learning on point sets for 3D classification and segmentation[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 77-85. DOI:10.1109/CVPR.2017.16

[32] Qi C R, Yi L, Su H, et al. PointNet++: Deep hierarchical feature learning on point sets in a metric space[EB/OL]. 2017: 1706.02413. https://arxiv.org/abs/1706.02413v1

[33] Luo Z P, Zhang G J, Zhou C Q, et al. TransPillars: Coarse-to-fine aggregation for multi-frame 3D object detection [C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2023: 4219-4228. DOI:10.1109/WACV56688.2023.00421

[34] Lang A H, Vora S, Caesar H, et al. PointPillars: Fast encoders for object detection from point clouds[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 12689-12697. DOI: 10.1109/CVPR.2019.01298

[35] Shi W J, Rajkumar R. Point-GNN: Graph neural network for 3D object detection in a point cloud[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020:1708-1716. DOI:10.1109/cvpr42600.2020.00178

[36] Meraz M, Ansari M A, Javed M, et al. DC-GNN: Drop channel graph neural network for object classification and part segmentation in the point cloud[J]. International Journal of Multimedia Information Retrieval, 2022,11(2):123-133. DOI:10.1007/s13735-022-00236-7

[37] Xiong S M, Li B, Zhu S. DCGNN: A single-stage 3D object detection network based on density clustering and graph neural network[J]. Complex & Intelligent Systems, 2023,9(3):3399-3408. DOI:10.1007/s40747-022-00926-z

[38] Zarzar J, Giancola S, Ghanem B. PointRGCN: Graph convolution networks for 3D vehicles detection refinement [EB/OL]. 2019: 1911.12236. https://arxiv.org/abs/1911.12236v1

[39] Wang X, Li K Q, Chehri A. Multi-sensor fusion technology for 3D object detection in autonomous driving: A review[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(2): 1148-1165. DOI: 10.1109/TITS.2023.3317372

[40] Tang Q S, Bai X Y, Guo J T, et al. DFAF3D: A dual-feature-aware anchor-free single-stage 3D detector for point clouds[J]. Image and Vision Computing, 2023, 129: 104594. DOI:10.1016/j.imavis.2022.104594

[41] Xiao W P, Peng Y, Liu C, et al. Balanced sample assignment and objective for single-model multi-class 3D object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023,33(9):5036-5048. DOI: 10.1109/TCSVT.2023.3248656

[42] Koo I, Lee I, Kim S H, et al. PG-RCNN: Semantic surface point generation for 3D object detection[C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2023: 18096-18105. DOI: 10.1109/ICCV51070.2023.01663

[43] Chen S T, Zhang H L, Zheng N N. Leveraging anchor-based LiDAR 3D object detection via point assisted sample selection[EB/OL]. 2024:2403.01978. https://arxiv.org/abs/2403.01978v1

[44] Feng X Y, Du H M, Fan H H, et al. SEFormer: Structure embedding transformer for 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023,37(1):632-640. DOI:10.1609/aaai.v37i1.25139

[45] He C H, Li R H, Li S, et al. Voxel set transformer: A set-to-set approach to 3D object detection from point clouds [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022:8407-8417. DOI:10.1109/CVPR52688.2022.00823

[46] Chen C, Chen Z, Zhang J, et al. SASA: Semantics-augmented set abstraction for point-based 3D object detection[J]. Proceedings of the AAAI Conference on Artifi-

cial Intelligence, 2022,36(1):221-229. DOI:10.1609/aaai.v36i1.19897

[47] Xie T, Wang L, Wang K, et al. FARP-net: Local-global feature aggregation and relation-aware proposals for 3D object detection[J]. IEEE Transactions on Multimedia, 2023,26:1027-1040. DOI:10.1109/TMM.2023.3275366

[48] Xia Q M, Chen Y D, Cai G R, et al. 3-D HANet: A flexible 3-D heatmap auxiliary network for object detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023,61:5701113. DOI:10.1109/TGRS.2023.3250229

[49] Dong Z C, Ji H, Huang X F, et al. PeP: A Point enhanced Painting method for unified point cloud tasks[EB/OL]. 2023:2310.07591. https://arxiv.org/abs/2310.07591v2.

[50] Zheng W, Tang W L, Jiang L, et al. SE-SSD: Self-ensembling single-stage object detector from point cloud[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 14489-14498. DOI:10.1109/cvpr46437.2021.01426

[51] Shi S S, Jiang L, Deng J J, 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. DOI:10.1007/s11263-022-01710-9

[52] Fei H, Zhao J, Zhang Z, et al. PV-GNN: point-voxel 3d object detection based on graph neural network[J]. PREPRINT (Version 1) available at Research Square, 2024. DOI:10.21203/rs.3.rs-4598182/v1

[53] Deng J J, Shi S S, Li P W, et al. Voxel R-CNN: Towards high performance voxel-based 3D object detection[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021,35(2):1201-1209. DOI:10.1609/aaai.v35i2.16207

[54] Shi S, Guo C, Jiang L, et al. PV-RCNN: point-voxel feature set abstraction for 3d object detection[C]// Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). IEEE, 2020: 10526-10535. DOI:10.1109/CVPR42600.2020.01054

[55] Shenga H L, Cai S J, Liu Y, et al. Improving 3D object detection with channel-wise transformer[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021: 2723-2732. DOI: 10.1109/ICCV48922.2021.00274

[56] 孔德明, 李晓伟, 杨庆鑫. 基于伪点云特征增强的多模态三维目标检测方法[J]. 计算机学报, 2024,47(4):759-775. [Kong D M, Li X W, Yang Q X. Multimodal 3D object detection method based on pseudo point cloud feature enhancement[J]. Chinese Journal of Computers, 2024,47(4):759-775.] DOI:10.11897/SP.J.1016.2024.00759

[57] Liu Z, Tang H, Lin Y, et al. Point-voxel CNN for efficient 3d deep learning[J]. Advances in neural information processing systems, 2019, 32. DOI: 10.48550/arXiv.190 7.03739

[58] Chen Y L, Liu S, Shen X Y, et al. Fast point R-CNN[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019:9774-9783. DOI:10.1109/iccv.2019.00987

[59] Vora S, Lang A H, Helou B, et al. PointPainting: Sequential fusion for 3D object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020:4603-4611. DOI:10.1109/cvpr42600.2020.00466

[60] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021:9992-10002. DOI:10.1109/ICCV48922.2021.00986

[61] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020. DOI:10.48550/arXiv.2010.11929

[62] 彭颖,张胜根,黄俊富,等.基于自注意力机制的两阶段三维目标检测方法[J].科学技术与工程,2024,24(25):10825-10831. [Peng Y, Zhang S G, Huang J F, et al. Two-stage 3D object detection method based on self-attention mechanism[J]. Science Technology and Engineering, 2024,24(25):10825-10831.] DOI:10.12404/j.issn.1671-1815.2400232

[63] 鲁斌,杨振宇,孙洋,等.基于多通道交叉注意力融合的三维目标检测算法[J].智能系统学报,2024,19(4):885-897. [Lu B, Yang Z Y, Sun Y, et al. 3D object detection algorithm with multi-channel cross attention fusion[J]. CAAI Transactions on Intelligent Systems, 2024,19(4):885-897.] DOI:10.11992/tis.202305029

[64] 张素良,张惊雷,文彪.基于交叉自注意力机制的LiDAR点云三维目标检测[J].光电子·激光,2024,35(1):75-83. [Zhang S L, Zhang J L, Wen B. LiDAR point cloud 3D object detection based on cross self-attention mechanism[J]. Journal of Optoelectronics·Laser, 2024,35(1):75-83.] DOI:10.16136/j.joel.2024.01.0593

[65] 刘明阳,杨啟明,胡冠华,等.基于Transformer的3D点云目标检测算法[J].西北工业大学学报,2023,41(6):1190-1197. [Liu M Y, Yang Q M, Hu G H, et al. 3D point cloud object detection algorithm based on Transformer[J]. Journal of Northwestern Polytechnical University, 2023, 41(6):1190-1197.] DOI:10.1051/jnwpu/20234161190

[66] Zhao H S, Jiang L, Jia J Y, et al. Point transformer[C]// 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021:16239-16248. DOI:10.1109/ICCV48922.2021.01595

[67] Pei Y, Zhao X, Li H, et al. Clusterformer: Cluster-based transformer for 3D object detection in point clouds[C]//

2023 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2023:6641-6650. DOI:10.1109/ICCV51070.2023.00613

[68] Ren S Y, Pan X, Zhao W J, et al. Dynamic graph transformer for 3D object detection[J]. Knowledge-Based Systems, 2023,259:110085. DOI:10.1016/j.knosys.2022.110085

[69] Lu B, Sun Y, Yang Z Y. Voxel graph attention for 3-D object detection from point clouds[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 5023012. DOI:10.1109/TIM.2023.3301907

[70] Ai L M, Xie Z Y, Yao R X, et al. MVTr: Multi-feature voxel transformer for 3D object detection[J]. The Visual Computer, 2024, 40(3): 1453-1466. DOI: 10.1007/s00371-023-02860-8

[71] Hoang H A, Bui D C, Yoo M. TSSTDet: Transformation-based 3-D object detection via a spatial shape transformer [J]. IEEE Sensors Journal, 2024,24(5):7126-7139

[72] Dong Y P, Kang C X, Zhang J L, et al. Benchmarking robustness of 3D object detection to common corruptions in autonomous driving[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 1022-1032. DOI: 10.1109/CVPR52729.2023.00105

[73] 刘慧,董振阳,田帅华.融合点云和体素信息的目标检测网络[J].计算机工程与设计,2024,45(9):2771-2778. [Liu H, Dong Z Y, Tian S H. Object detection network fusing point cloud and voxel information[J]. Computer Engineering and Design, 2024,45(9):2771-2778.] DOI:10.16208/j.issn1000-7024.2024.09.029

[74] Wu H, Wen C L, Li W, et al. Transformation-equivariant 3D object detection for autonomous driving[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023,37(3):2795-2802. DOI:10.1609/aaai.v37i3.25380

[75] Guo S, Cai J Y, Hu Y Z, et al. LCASAFormer: Cross-attention enhanced backbone network for 3D point cloud tasks [J]. Pattern Recognition, 2025,162:111361. DOI:10.1016/j.patcog.2025.111361

[76] Lee J, Lee Y, Kim J, et al. Set transformer: A framework for attention-based permutation-invariant neural networks [C]// International Conference on Machine Learning (ICML). PMLR, 2019: 3744-3753. DOI: 10.48550/arXiv.1810.00825

[77] Fan L, Pang Z Q, Zhang T Y, et al. Embracing single stride 3D object detector with sparse transformer[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 8448-8458. DOI: 10.1109/CVPR52688.2022.00827

[78] Sun P, Tan M X, Wang W Y, et al. SWFormer: Sparse window transformer for 3D object detection in point clouds [M]//Computer Vision-ECCV 2022. Cham: Springer Nature Switzerland, 2022:426-442. DOI:10.1007/978-3-031-20080-9_25

[79] Wang H Y, Shi C, Shi S S, et al. DSVT: Dynamic sparse voxel transformer with rotated sets[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 13520-13529. DOI: 10.1109/CVPR52729.2023.01299

[80] Han K, Xiao A, Wu E, et al. Transformer in transformer [J]. Advances in Neural Information Processing Systems, 2021,34:15908-15919

[81] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149. DOI: 10.1109/TPAMI.2016.2577031

[82] Te G S, Hu W, Zheng A M, et al. RGCNN: Regularized graph CNN for point cloud segmentation[C]//Proceedings of the 26th ACM International Conference on Multimedia. ACM, 2018:746-754. DOI:10.1145/3240508.3240621

[83] Corporation N. NVIDIA documentation hub[EB/OL]. [9-15]. https://docs.nvidia.com/#all-documents.

[84] PyTorch. Torch. scatter[EB/OL]. [5-24]. https://pytorch.org/docs/2.3/generated/torch.scatter.html#torch.scatter.

[85] Qi C R, Litany O, He K M, et al. Deep Hough voting for 3D object detection in point clouds[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019:9276-9285. DOI:10.1109/iccv.2019.00937

[86] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[M]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020:213-229. DOI:10.1007/978-3-030-58452-8_13

[87] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 3354-3361. DOI: 10.1109/CVPR.2012.6248074

[88] Sun P, Kretzschmar H, Dotiwalla X, et al. Scalability in perception for autonomous driving: Waymo open dataset [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020:2443-2451. DOI:10.1109/cvpr42600.2020.00252

[89] Team O D. OpenPCDet: An open-source toolbox for 3d object detection from point clouds[EB/OL]. (2024-12-30) [10.1]. https://github.com/open-mmlab/OpenPCDet.

■ 本文图文责任编辑：黄光玉 蒋树芳