中图法分类号： **文献标识码：** A **文章编号：** 1006-8961(XXXX)XX-0001-16

论文引用格式： Li Xing, Zheng Yuhui. 3D object detection with pseudo-image features enhanced by dynamic spatial global attention[J/OL]. Journal of Image and Graphics, XXXX:1-16. DOI: 10.11834/jig.250583. (李幸, 郑钰辉. 动态空间全局注意力增强的伪图像特征3D目标检测[J/OL]. 中国图象图形学报, XXXX:1-16. DOI: 10.11834/jig.250583.)[DOI:10.11834/jig.250583]

动态空间全局注意力增强的伪图像特征3D目标检测

李幸[1], 郑钰辉[2]

1. 南京信息工程大学计算机学院、软件学院, 南京 210044; 2. 青海师范大学计算机学院, 西宁 810008

摘 要： **目的** 针对三维目标检测任务中, 点云经柱体化映射为二维伪图像时易导致的几何细节丢失, 以及伪图像特征提取阶段建模不足的问题, 提出一种动态空间全局注意力增强的伪图像特征3D目标检测模型。**方法** 首先, 设计动态空间全局注意力机制, 通过为每个样本生成自适应卷积核实现针对性的局部结构建模, 并结合全局自注意力以增强空间特征交互与长程依赖建模能力, 从而在稀疏伪图像上恢复更多几何信息。其次, 构建高效多尺度伪图像特征融合网络, 在兼顾细粒度局部特征和全局语义的同时, 通过多尺度特征融合整合浅层空间细节与深层语义。最后, 提出双重感知动态上采样, 在DySample的基础上采用边缘感知与小目标显著性增强双通路设计, 优化上采样过程并提升空间细节恢复与目标检测性能。**结果** 在KITTI(Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago)和DAIR-V2X-V两个公开数据集上的模型评估结果表明, 与基线模型PointPillars相比, 本文方法在KITTI数据集中等难度下对汽车、行人和骑行者类别的平均精度(average precision, AP)分别提升2.51%、6.07%和10.33%, 平均精度均值(mean average precision, mAP)提升6.3%; 在DAIR-V2X-V数据集中等难度下对汽车、行人和骑行者类别的AP分别提升0.07%、6.48%和8.70%, mAP提升5.09%。**结论** 所提方法通过利用动态空间全局注意力机制、采用高效多尺度特征融合策略及双重感知动态上采样设计, 有效缓解柱体化映射带来的几何信息损失, 增强伪图像的空间结构与语义表达能力, 从而显著提升稀疏点云场景下的三维目标检测性能。

关键词： 计算机视觉; 自动驾驶; 三维目标检测; 注意力机制; 特征融合; 上采样

3D object detection with pseudo-image features enhanced by dynamic spatial global attention

Li Xing[1], Zheng Yuhui[2]

1. School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China; 2. School of Computer Science, Qinghai Normal University, Xining 810016, China

Abstract: Objective In recent years, autonomous driving technology has developed rapidly, with its core goal being to reduce the driver's workload and enhance driving safety. In autonomous driving systems, 3D object detection serves as a critical component of the environmental perception module, and its performance directly affects the accuracy and reliability of subsequent path planning and decision-making. In current 3D object detection tasks, LiDAR-based methods have become a research focus in both academia and industry because they can directly capture high-precision spatial geometric information and exhibit strong robustness to environmental disturbances such as lighting changes and adverse weather conditions. Among various LiDAR-based detection methods, pillar-based approaches have been widely applied in practical

收稿日期： 2025-11-18; **修回日期：** 2026-03-20

基金项目： 江苏省前沿技术研发计划资助(BF2024070)

Supported by: Frontier Technologies R&D Program of Jiangsu (BF2024070)

autonomous driving scenarios due to their superior computational efficiency and real-time performance. These methods employ a unique pillarization process: the original unstructured point cloud data is first discretized into a series of regularly arranged vertical pillars, and then the 3D pillar features are aggregated and projected onto a 2D plane to generate pseudo-image representations. This transformation converts sparse and irregular 3D point clouds into structured 2D grid data, enabling the use of mature 2D convolutional neural networks for efficient feature extraction and object detection, thereby significantly reducing the computational complexity of 3D detection. However, during the process of mapping point clouds to 2D pseudo-images, the unavoidable information compression and dimensionality reduction often lead to the loss of fine-grained geometric details in 3D space. This issue is particularly severe in sparse point cloud regions or scenes containing small objects. Moreover, in the feature extraction stage of pseudo-images, conventional 2D convolutional neural networks are limited by their fixed local receptive fields and thus struggle to fully capture long-range dependencies and local features, resulting in insufficient feature representation and further constraining detection performance. To address these challenges, this paper proposes a 3D object detection model with pseudo-image features enhanced by dynamic spatial global attention. **Method** First, a dynamic spatial global attention mechanism is designed. This module generates sample-adaptive convolutional kernels to achieve targeted modeling of local geometric structures. Meanwhile, a global self-attention mechanism is introduced to enhance spatial feature interaction and long-range dependency modeling. The combination of local and global representations helps preserve more geometric details in pseudo-images, effectively mitigating information loss during the pillarization process, and improves contextual awareness for small or occluded objects. Second, an efficient multi-scale pseudo-image feature fusion network is constructed. The network adopts a three-stage hierarchical feature extraction structure to balance fine-grained local details and high-level semantic information, and incorporates a feature pyramid network for multi-scale feature fusion. Specifically, during the feature fusion stage, a top-down upsampling and lateral connection strategy is employed to unify features of different scales to the same resolution, followed by channel-wise fusion. This effectively integrates shallow spatial details with deep semantic representations, enhancing the overall representational capability of pseudo-image features and significantly improving their suitability for downstream 3D detection tasks. Finally, a dual perception dynamic upsampling module is proposed. Based on DySample, an edge-awareness branch and a small-object saliency enhancement branch are introduced to collaboratively optimize the spatial detail recovery during upsampling. This design improves the modeling performance for small objects and complex boundaries in 3D object detection. **Result** To systematically evaluate the performance of the proposed method, comprehensive comparison experiments were conducted on two public datasets, KITTI and DAIR-V2X-V. On the KITTI dataset, the proposed model was compared with various representative 3D object detection methods, including both multi-modal and LiDAR-based approaches. The compared methods include AVOD, F-PointNet, VeloFCN, SECOND, Associate-3Det, VoxelNet, PointRCNN, HS-Pillar, TANet, CenterPoint, VoxelNeXt, MAT-PointPillars, PointBi-FPN, XPillars and PointPillars. Experimental results show that, compared with the baseline model PointPillars, the proposed method achieves AP improvements for the car category of 1.31%, 2.51%, and 2.76% at easy, moderate, and hard difficulty levels, respectively; for pedestrian detection, the improvements are 6.96%, 6.07%, and 6.10%; and for cyclist detection, the improvements are 10.62%, 10.33%, and 9.81%. At the moderate difficulty level, the proposed algorithm achieves a mAP of 65.44%, which is 6.3 percentage points higher than the baseline model. Similarly, on the DAIR-V2X-V test set, the proposed method was compared with several representative 3D object detection methods, including Part-A2, SECOND, PointRCNN, and PointPillars. The experimental results indicate that, compared with the baseline PointPillars, the proposed method improves AP for the car category by 0.08%, 0.07%, and 0.30% at easy, moderate, and hard difficulty levels, respectively; for pedestrian detection, the improvements are 7.81%, 6.48%, and 6.04%; and for cyclist detection, the improvements are 7.45%, 8.70%, and 8.64%. The overall mAP for all categories reaches 44.37%, which is 5.09 percentage points higher than that of the baseline model. **Conclusion** To address the issues of geometric detail loss during the pillar-based projection of point clouds into pseudo-images and insufficient modeling during pseudo-image feature extraction, this paper proposes a 3D object detection model with pseudo-image features enhanced by dynamic spatial global attention. First, a dynamic spatial global attention mechanism is introduced, which combines dynamic convolution kernel generation with a global self-attention mechanism to effectively improve the spatial representation capability of pseudo-image

features. Second, an efficient multi-scale pseudo-image feature fusion network is designed to enhance the model's ability to extract multi-scale features while preserving fine-grained local details and global semantic information. Finally, a dual-perception dynamic upsampling module is proposed, which guides the sampling point distribution through a dual-path design of edge awareness and small-object saliency enhancement, optimizing the upsampling process and improving spatial detail recovery and detection performance. Experimental results on the KITTI and DAIR-V2X-V datasets demonstrate that the proposed algorithm achieves superior detection performance. Ablation studies and visualization results further verify the effectiveness of each module. Although the proposed method significantly improves detection accuracy, the dynamic spatial global attention computation and multi-scale feature fusion mechanism result in relatively high model complexity and large parameter size. Future work will explore lightweight model designs to further enhance the algorithm's real-time performance and robustness, and to improve its applicability in diverse and complex scenarios.

Key words: computer vision; autonomous driving; 3D object detection; attention mechanism; feature fusion; upsampling

0 引　言

随着深度学习与自动驾驶技术(Ku等,2018)的快速发展,三维目标检测作为环境感知系统的重要组成部分,旨在根据来自传感器的一种或多种输入数据,预测场景中物体的位置、大小和类别,为路径规划与决策控制提供可靠的空间信息(Wang等,2016)。相较于传统的二维目标检测方法,基于深度学习的三维目标检测能够更充分地刻画目标的几何结构与空间分布特征,已成为实现高精度环境感知与智能决策的重要支撑技术。

目前,基于深度学习的三维目标检测方法主要可分为三类:基于图像的方法、基于激光雷达的方法以及基于多模态融合的方法。

基于图像的方法(Chen等,2016;Li等,2020)依赖单目或双目图像信息,通常在二维图象中检测目标,再通过深度估计或几何投影将检测结果映射到三维空间。该类方法硬件成本低、部署灵活,但其性能易受光照变化、遮挡以及视差误差的影响,尤其在远距离场景下检测精度显著下降,难以满足自动驾驶场景对高鲁棒性与高精度的要求。

基于激光雷达的方法利用点云提供的精确三维几何信息,具备高精度测距与强抗干扰能力,已成为三维目标检测的主流技术路线。然而,点云数据本身具有稀疏性、无序性和非结构化的特点(龚靖渝等,2023),无法直接由传统卷积神经网络进行处理。根据点云处理方式的不同,该类方法可进一步划分为基于点的方法和基于体素的方法(陶帅兵 等,2021)。基于点的方法以PointNet(Charles 等,2017)为代表,直接从原始点云中学习几何特征,检测精度高。然而,其直接对点云中的每个点进行处理,计算资源消耗较大,难以满足对实时性要求高的大规模自动驾驶场景。Yu等人(2022)提出的Point-BERT模型借助点云Transformer结构,显著增强了对长程依赖与复杂上下文的建模能力,但其二次复杂度带来的高延迟与显存占用,仍使其难以满足实时自动驾驶的推理需求。基于体素的方法则将稀疏点云转换为规则体素网格,然后使用3D卷积进行特征提取。典型方法VoxelNet(Zhou和Tuzel,2018)首次引入体素特征编码和3D卷积神经网络(convolutional neural networks,CNN)进行特征学习,但由于点云分布稀疏,体素网格中存在大量的空体素,三维卷积会造成较高的计算与存储开销。为克服该问题,后续研究主要沿两个方向推进:一是改进三维体素特征的学习效率,例如SECOND(sparsely embedded convolutional detection)(Yan 等,2018)引入稀疏卷积以加速处理;VoxelNeXt(Chen等,2023)进一步提出了无锚框的纯稀疏卷积网络,简化了检测流程。二是对体素表征进行降维,将其投影至二维平面形成伪图像特征表示,从而利用成熟的2D卷积神经网络提升处理速度。其中,PointPillars(Lang等,2019)算法创新性地将点云划分为点柱并映射为二维伪图像,随后利用2D卷积进行特征提取,实现了精度与速度的良好平衡。此后,研究者们致力于增强伪图像的表征能力:CaDDN(Reading等,2021)通过深度分布建模强化了空间感知;PillarNeXt(Li等,2023)则在单尺度柱体特征图上引入ASPP模块扩大感受野,并通过头部轻量级上采样恢复细节,有效缓解了柱体编码带来的信息丢失。尽管上述研究均取得了显著进展,但基于体素的方法仍面临核心挑战:首先,点云经柱体化映射为二维伪图像的过程中,不可避

免地会因信息压缩和维度缩减而丢失三维空间中的细粒度几何信息,这一问题在稀疏点云与小尺度目标上影响尤为严重。其次,在伪图像特征提取阶段,传统二维卷积神经网络受限于其固定的局部感受野,难以充分建模点云空间中关键的长程依赖关系和局部精细几何结构,导致特征表达能力不足,从而制约了对小尺度目标和复杂边界形状的检测性能。

基于多模态融合的方法(Ku 等,2018)融合激光雷达的空间结构信息与图像的语义纹理信息,利用模态互补性进一步提高检测性能。但该类方法对传感器间的精确标定与时序同步依赖较强,模型结构复杂,计算与存储开销较大,且在极端光照或恶劣天气条件下,某一模态的性能退化可能影响系统整体鲁棒性。尽管近期研究 SparseInteraction(Xu 等,2024)等通过引入稀疏语义引导以优化雷达与相机融合的策略,其性能仍依赖于跨模态的精确关联,且引入了额外的模态交互复杂性,这使其在追求高鲁棒性与高效率的车载系统中部署仍面临挑战。

综上所述,在自动驾驶对实时性、鲁棒性与高精度的综合需求驱动下,解决基于体素的方法中存在的几何细节丢失与特征建模能力不足问题,显得尤为迫切。为此,本文提出一种动态空间全局注意力增强的伪图像特征 3D 目标检测模型。本文的主要贡献如下:

1)提出一种新颖的动态空间全局注意力机制,通过生成样本特定卷积核实现自适应的局部结构建模,同时结合全局自注意力机制,增强空间特征交互与长程依赖建模能力。

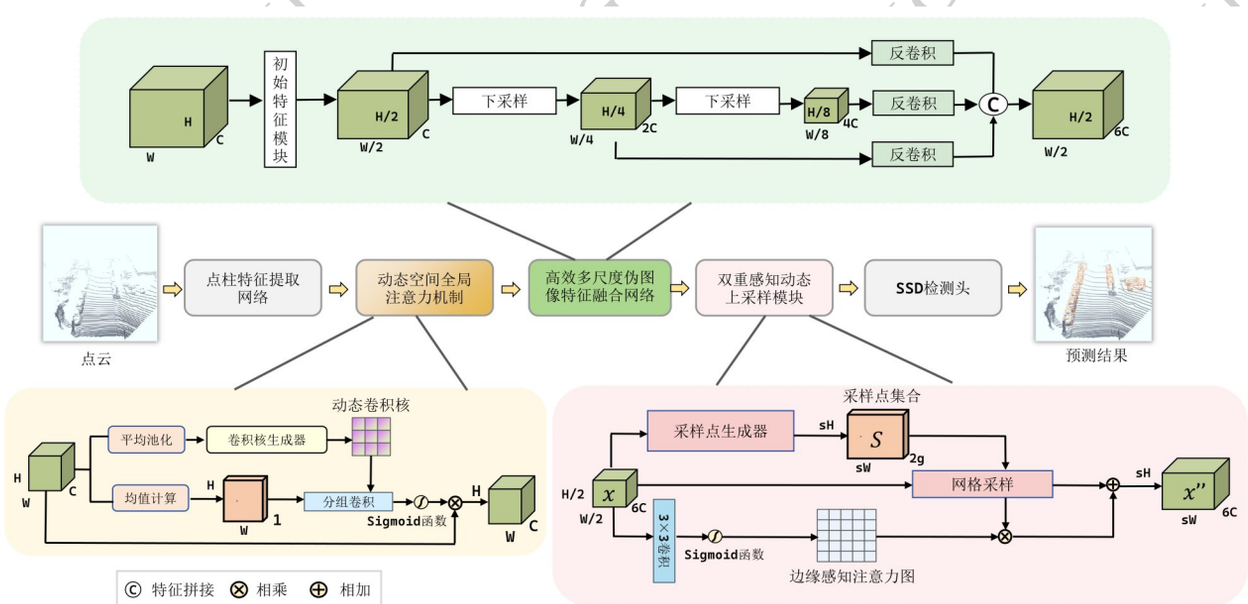
2)设计了高效多尺度伪图像特征融合网络,在兼顾细粒度局部特征和全局语义的同时,采用自顶向下的上采样与横向连接策略,将不同尺度的特征统一至相同分辨率后进行通道级融合,从而有效整合浅层空间细节与深层语义表示。

3)设计了双重感知动态上采样模块,采用边缘感知与小目标显著性增强双通路设计,优化上采样过程并提升空间细节恢复与目标检测性能。

1 模型方法

1.1 整体框架

所提模型整体框架如图 1 所示。该模型以原始无序三维点云为输入,最终输出包含目标的三维边界框等检测结果。整体网络结构主要由点柱特征提取网络、动态空间全局注意力机制(dynamic spatial global attention, DSGA)、高效多尺度伪图像特征融合网络(efficient multi-scale pseudo-image feature fusion network, EMPFN)、双重感知动态上采样模块(dual perception dynamic upsampling module, DuSample)以及单阶段目标检测头(single shot detector, SSD)(Liu 等,2016)组成。



具体而言,首先将原始无序点云输入网络,通过

点柱特征编码将三维点云映射为规则化的二维伪图像表示，从而便于后续的卷积与注意力机制处理。在此基础上，通过动态空间全局注意力机制，增强伪图像特征的空间建模能力，捕捉全局依赖关系。随后，利用高效多尺度伪图像特征融合网络对不同尺度的特征进行融合与交互，从而充分利用跨层次信息，提升特征表达能力。接着，双重感知动态上采样模块通过采用边缘感知与小目标显著性增强的双通路设计，优化上采样过程并实现细节恢复，为检测头提供高质量的特征输入。最后，上采样后的特征被送入SSD目标检测头，直接预测三维目标的边界框及类别，实现准确且高效的三维目标检测。

图1 所提模型整体框架

Fig. 1 The overall framework of the proposed model

1.2 动态空间全局注意力机制

在三维目标检测任务中，点云数据通过柱体化映射为二维伪图像特征的过程中，往往伴随几何细节的丢失，这对小目标和遮挡目标的检测性能产生了显著影响。传统空间注意力方法（Wang 等，2020）通常采用固定结构的卷积或池化操作生成注意力图，其感受野受限，难以捕捉跨区域上下文信息。尽管

这类方法能够在一定程度上突出显著区域，但由于缺乏对输入特征的自适应性，在处理三维场景中的小目标和遮挡目标时，往往无法有效平衡特征增强与背景抑制之间的关系。为改进这一局限，研究者们提出了基于输入特征的动态建模方法（Chen 等，2020），通过自适应生成卷积核来调整注意力分布，

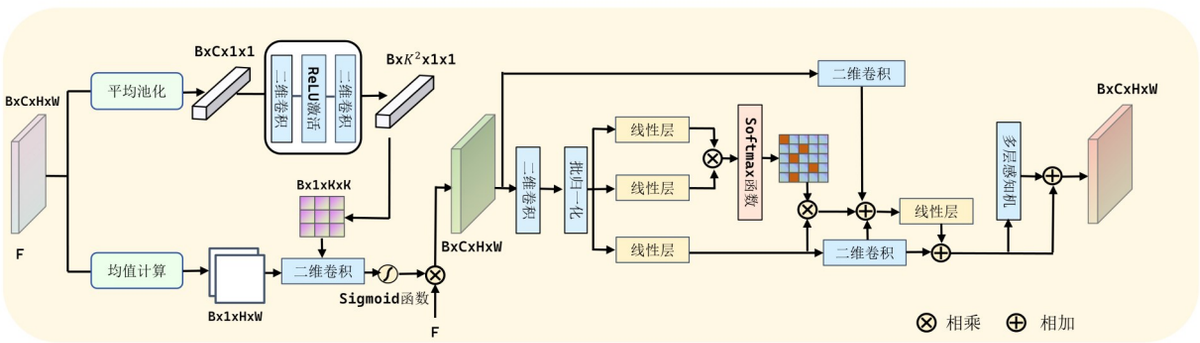


图2 动态空间全局注意力机制

Fig. 2 Dynamic spatial global attention

而增强小目标的显著性并削弱局部背景噪声。然而，此类方法仍然主要依赖局部动态卷积，缺乏对全局依赖关系的显式建模。在存在大范围干扰或需进行长程语义推理的三维场景中，模型易出现误检与漏检现象。

针对三维目标检测中局部几何细节保留与全局上下文感知的双重需求，本文提出了一种动态空间全局注意力机制（dynamic spatial global attention，DSGA），其整体结构如图2所示。该模块通过动态卷积与全局自注意力的协同设计，旨在解决投影过程中的细节损失与远距离依赖建模问题。具体而言，通过动态卷积自适应增强局部几何特征，弥补投影过程中的细节损失；同时利用自注意力机制实现跨区域的全局信息交互，增强对遮挡或小目标的上下文感知能力，从而有效提升模型在复杂背景下的特征判别能力。设输入的伪图像特征为 $\boldsymbol{F} \in \mathbf{R}^{B\times C\times H\times W}$，其中 B 为批次大小，C 为通道数，H、W 为特征图的高和宽。DSGA模块首先对其进行两种全局信息提取：

$$\boldsymbol{F}_{\text{avg}} = \text{AvgPool}(\boldsymbol{F}), \quad \boldsymbol{F}_{\text{mean}} = \text{Mean}(\boldsymbol{F}) \tag{1}$$

式中，AvgPool(·)表示全局平均池化操作，用于获取通道级全局上下文信息 $\boldsymbol{F}_{\text{avg}}$；Mean(·)表示沿空间维度求均值，用以生成压缩后的空间特征 $\boldsymbol{F}_{\text{mean}}$。

随后，将包含通道信息的 $\boldsymbol{F}_{\text{avg}}$ 输入至一个由两个1×1卷积层与ReLU激活函数组成的轻量级卷积模块，以生成自适应卷积核参数：

$$\boldsymbol{K}_{\text{dyn}} = \text{Conv}_{1\times1}\left(\text{ReLU}\left(\text{Conv}_{1\times1}(\boldsymbol{F}_{\text{avg}})\right)\right) \tag{2}$$

接着，将动态卷积核 $\boldsymbol{K}_{\text{dyn}}$ 与 $\boldsymbol{F}_{\text{mean}}$ 进行卷积操作，并通过Sigmoid激活函数获得空间注意力权重。该

权重作用于原始输入特征 $\boldsymbol{F}$，实现局部特征的自适应增强：

$$\boldsymbol{F}_{\mathrm{dyn}}=\sigma\left(\mathrm{Conv}\left(\boldsymbol{F}_{\mathrm{mean}},\boldsymbol{K}_{\mathrm{dyn}}\right)\right)\odot\boldsymbol{F} \tag{3}$$

式中，$\sigma(\cdot)$表示Sigmoid激活函数，$\odot$表示逐元素乘法。该过程使模型能够依据输入内容动态强化信息丰富的结构区域，有效改善传统卷积在复杂几何结构建模时的局限性。

在获得局部增强特征$\boldsymbol{F}_{\mathrm{dyn}}$后，结合全局自注意力机制实现跨区域特征的长程依赖建模。首先，通过对$\boldsymbol{F}_{\mathrm{dyn}}$进行线性变换得到查询矩阵$\boldsymbol{Q}$、键矩阵$\boldsymbol{K}$和值矩阵$\boldsymbol{V}$：

$$\boldsymbol{Q}=\boldsymbol{W}_{Q}\boldsymbol{F}_{\mathrm{dyn}},\boldsymbol{K}=\boldsymbol{W}_{K}\boldsymbol{F}_{dyn},\boldsymbol{V}=\boldsymbol{W}_{V}\boldsymbol{F}_{\mathrm{dyn}} \tag{4}$$

式中，$\boldsymbol{W}_{Q},\boldsymbol{W}_{K},\boldsymbol{W}_{V}$分别表示可学习的线性变换参数矩阵，用于特征映射的降维与特征空间变换。

随后，计算特征间的相关性矩阵以建模全局上下文关系：

$$\boldsymbol{A}=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{T}}{\sqrt{d_{k}}}\right) \tag{5}$$

式中，d_k为键矩阵$\boldsymbol{K}$的特征维度，用作缩放因子以稳定训练。通过该注意力矩阵，模型能够捕获远距离像素之间的依赖关系，实现全局语义建模。最终，将注意力加权的结果与值矩阵$\boldsymbol{V}$相乘得到全局增强特征：

$$\boldsymbol{F}_{\mathrm{att}}=\boldsymbol{A}\boldsymbol{V} \tag{6}$$

为有效融合局部增强特征与全局上下文信息，采用多层感知机进行特征变换，并引入残差连接以确保训练稳定性：

$$\boldsymbol{F}_{\mathrm{out}}=\mathrm{MLP}\left(\boldsymbol{F}_{\mathrm{att}}+\boldsymbol{F}_{\mathrm{dyn}}\right)+\boldsymbol{F} \tag{7}$$

式中，最后一项表示引入的残差连接，用于保留输入特征的原始语义信息，从而防止特征表示退化。

本文提出的DSGA模块通过动态卷积与全局自注意力的协同机制，在三维目标检测任务中实现了局部几何细节增强与全局上下文建模的统一。动态卷积通过自适应调整卷积核，增强了目标区域的细节表征能力，特别在处理小目标和遮挡目标时，有效缓解了点云投影过程中导致的几何信息丢失，显著提升了对小目标几何特征的保留能力；而全局自注意力机制则通过建立远距离像素间的语义关联，增强了对遮挡目标的上下文感知能力。这种针对三维检测难点的协同设计使DSGA模块在复杂三维场景中具备更强的特征判别与目标定位能力，为提升三维目标检测性能提供了有效的解决方案。

1.3 高效多尺度伪图像特征融合网络

伪图像特征提取通常采用基于卷积与下采样的网络架构，该方式虽能高效聚合局部空间结构信息，但在处理由点云投影生成的伪图像特征时仍面临固有局限。受限于卷积核感受野有限及层间特征交互不足，模型在长程依赖与全局上下文建模方面表现较弱；同时，多次下采样操作易导致伪图像中关键的几何细节信息丢失。这些问题在稀疏点云与小目标场景下尤为突出，易导致特征表达不充分与语义信息缺失等问题，进而影响检测精度与鲁棒性。

针对上述问题，本文受轻量化网络 EfficientFormerV2（Li 等，2023）的启发，结合伪图像的稀疏性与几何结构敏感性等特点，提出了一种高效多尺度伪图像特征融合网络（efficient multi-scale pseudo-image feature fusion network，EMPFN），整体结构如图3所示。该网络的设计核心在于通过精简的下采样路径与强化的多尺度融合，构建一个更适应伪图像特性的特征提取架构。具体而言，EfficientFormerV2原采用四阶段下采样策略（比例为 1/4、1/8、1/16 和 1/32），在自然图像任务中取得了良好效果。然而，该架构在处理伪图像特征时，其深层次下采样会加剧几何细节的流失。为此，本文采用三阶段下采样结构（尺度比例调整为1/2、1/4与1/8），旨在控制模型复杂度的同时，有效抑制了几何细节的丢失，保留了更丰富的空间信息，为后续的几何特征提取提供高分辨率的特征基础。进一步地，为充分利用不同层级特征的优势，本文借鉴特征金字塔网络的融合思想，通过自顶向下的上采样与横向连接，将不同层级的特征统一至相同分辨率并进行通道融合，以此构建一个能同时捕捉浅层精确空间细节与深层丰富语义的层次化特征表示，从而全面提升模型的多尺度目标感知与判别能力。

所提EMPFN网络主要由初始特征模块、三阶段层次化特征提取结构以及多尺度特征融合机制三部分组成。该架构通过自底向上的分层特征提取与自顶向下的多尺度融合路径，能有效地兼顾局部细节与全局语义。各组成部分介绍如下：

1）初始特征模块。作为网络的输入层，采用步长为 2 的标准卷积，并结合批归一化与GELU激活函数，对输入伪图像进行初步下采样，将其从原始空间映射至基础特征空间。该模块在压缩空间分辨率

的同时，尽可能保留边界与结构信息，为后续特征提取提供稳定的初始特征。

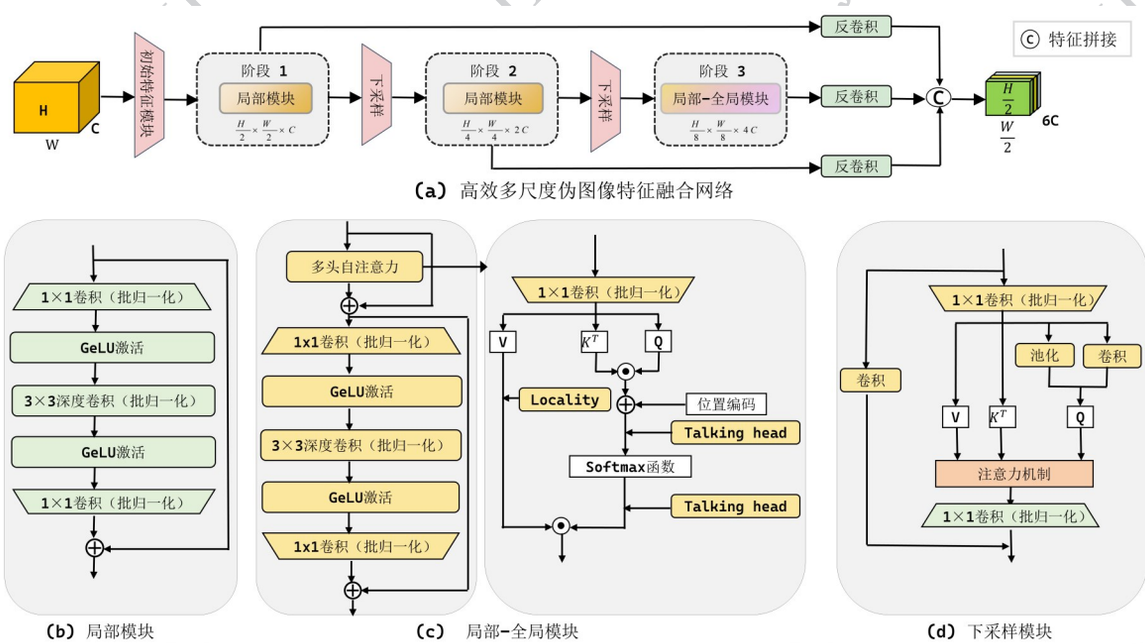
2）局部模块。第一与第二阶段均采用局部模块进行局部特征提取，其结构如图3（b）所示。该模块通过1×1卷积进行通道维度的扩展与压缩，并利用3×3深度可分离卷积在局部邻域内高效提取几何特征。模块内部引入的残差连接有助于稳定训练过程，其设计目标是在保持轻量化的前提下，充分捕捉伪图像中的边缘、角点等细粒度几何信息，强化模型对局部结构的感知能力。

3）局部-全局模块。第三阶段引入该模块，其结构如图3（c）所示。该模块在卷积结构基础上嵌入轻量化多头自注意力机制，并结合位置编码，旨在实现局部特征与全局语义的协同建模。此设计使模型在保持对局部几何结构敏感性的同时，能够捕捉远距离像素间的依赖关系，从而提升对复杂场景的特征建模能力。

4）下采样模块。在第一与第二阶段之后采用该模块，其结构如图3（d）所示。该模块在卷积与池化操作的基础上，嵌入轻量化注意力机制，使下采样过程能够自适应地保留信息更丰富的特征，增强特征图的语义连续性，避免常规下采样可能导致的信息均匀丢失问题。

5）多尺度特征融合机制。该机制将各阶段输出的特征图通过反卷积操作上采样至相同空间分辨率，随后在通道维度进行拼接，并利用1×1卷积完成特征融合与降维。通过该方式，融合后的特征图能够同时具备浅层特征的高空间分辨率与深层特征的强语义表达能力，为后续的检测任务提供信息更完备、判别力更强的特征表示。

综合以上模块设计，本文构建的EMPFN网络为伪图像特征提取提供了一种新的解决方案。该网络通过三阶段下采样与多尺度融合的有机结合，实现了几何细节保留与全局上下文建模的有效平衡。



（a）高效多尺度伪图像特征融合网络

（b）局部模块　（c）局部-全局模块　（d）下采样模块

（(a) Efficient multi-scale pseudo-image feature fusion network；(b) Local module；(c) Local-global module；(d)Subsample module）

图3　高效多尺度伪图像特征融合网络

Fig. 3　Efficient multi-scale pseudo-image feature fusion network

具体而言，网络以初始特征模块保证初始信息完整，通过局部与局部-全局模块分别强化局部几何感知与全局语义关联，并借助下采样模块减少关键语义信息损失，最终经多尺度融合形成判别力强的特征表示。因此，相较于传统特征提取网络，EMPFN更加契合伪图像的稀疏特性与三维检测任务对几何信息的高度依赖，为复杂场景下的精准、鲁棒三维检测提供了可靠的技术途径。

1.4　双重感知动态上采样

在多尺度伪图像特征融合网络处理后，增强后的伪图像特征已有效融合局部几何结构与全局语义信息。然而，受点云稀疏性的影响，小尺度目标在特

征图中仍存在边界模糊与特征响应不足的问题，这直接制约了三维目标检测模型在复杂场景下的定位与识别精度。为此，本文提出双重感知动态上采样模块（dual perception dynamic upsampling module，DuSample），其结构如图4所示。该模块在DySample（Liu 等，2023）基础上，针对三维点云中目标尺度差异大、边缘结构易丢失的特点，采用边缘感知分支与小目标显著性增强双分支策略，协同优化上采样过程中的空间细节恢复能力，从而显著提升模型对小尺度目标及复杂边界的建模性能。具体而言，设输入特征为 $\boldsymbol{X} \in \mathbf{R}^{C \times H \times W}$，首先通过线性映射生成初始偏移量：

$$\boldsymbol{O}_{\text{init}} = \text{f}_{\text{linear}}(\boldsymbol{X}), \quad \boldsymbol{O}_{\text{init}} \in \mathbf{R}^{2gs^2 \times H \times W} \tag{8}$$

式中，$\boldsymbol{X}$ 表示增强后的伪图像特征，其维度 C、H、W 分别为通道数、高度和宽度；$\text{f}_{\text{linear}}(\cdot)$ 表示线性映射操作；g 与 s 分别表示基准采样网格大小和上采样倍率，因此 gs^2 代表输出特征图上每个空间位置所对应的采样点总数；$\boldsymbol{O}_{\text{init}}$ 为初始偏移量，其维度 $2gs^2$ 表示每个位置对应的 gs^2 个采样点均包含二维 (x, y) 坐标偏移。

为增强模型对小尺度目标的感知能力，并缓解因稀疏性导致的特征弱响应问题，本文设计小目标显著性增强分支，通过卷积与Sigmoid激活函数生成显著性因子 α，引导模型关注小目标区域。表达式为：

$$\alpha = \sigma(\text{f}_{\text{salient}}(\boldsymbol{X})), \quad \alpha \in [0,1]^{H \times W} \tag{9}$$

式中，$\text{f}_{\text{salient}}(\cdot)$ 表示显著性增强卷积操作，$\sigma(\cdot)$ 为Sigmoid激活函数，α 为逐点显著性因子。

随后，利用显著性因子对偏移量进行调制，得到调制后的偏移量：

$$\boldsymbol{O} = \alpha \cdot \boldsymbol{O}_{\text{init}} \tag{10}$$

式中，$\boldsymbol{O}$ 为调制后的偏移量，该操作通过增强小目标区域的采样灵活性，使模型能够自适应聚焦于关键几何结构，从而缓解上采样过程中的细节损失。

在得到调制偏移量 $\boldsymbol{O}$ 后，将其与标准采样网格 $\boldsymbol{G}$ 相加，生成具有几何感知的采样点集合：

$$\boldsymbol{S} = \boldsymbol{G} + \boldsymbol{O} \tag{11}$$

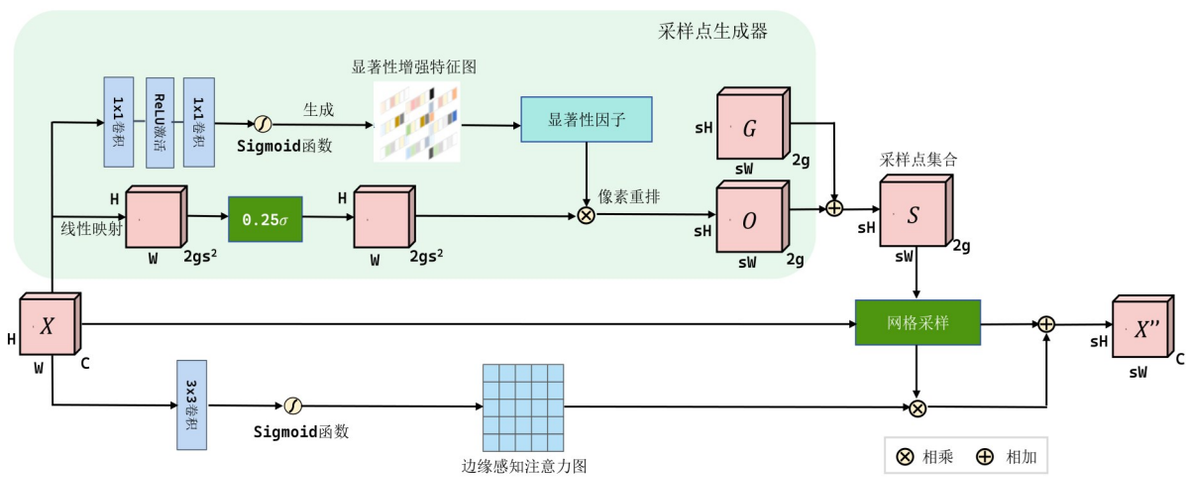


图4　双重感知动态上采样模块

Fig. 4　Dual perception dynamic upsampling module

式中，$\boldsymbol{G}$ 为标准采样网格，$\boldsymbol{S}$ 为最终采样点集合。接着，利用grid_sample完成动态上采样：

$$\boldsymbol{X}_{\text{mid}} = \text{grid\_sample}(\boldsymbol{X}, \boldsymbol{S}) \tag{12}$$

式中，$\boldsymbol{X}_{\text{mid}}$ 为上采样后的特征图，$\text{grid\_sample}(\cdot)$ 表示基于采样点集合的重采样操作。

为进一步增强边缘结构信息的保留，避免上采样过程中边缘细节的平滑效应，通过卷积与Sigmoid激活函数生成边缘感知注意力图 $\boldsymbol{M}_{\text{e}}$：

$$\boldsymbol{M}_{\text{e}} = \sigma(\text{f}_{\text{edge}}(\boldsymbol{X})) \tag{13}$$

式中，$\text{f}_{\text{edge}}(\cdot)$ 表示边缘特征提取卷积操作。利用边缘感知注意力图 $\boldsymbol{M}_{\text{e}}$ 对上采样后的特征图 $\boldsymbol{X}_{\text{mid}}$ 进行增强，得到最终输出特征：

$$\boldsymbol{X}' = \boldsymbol{X}_{\text{mid}} \odot (1 + \boldsymbol{M}_{\text{e}}) \tag{14}$$

式中，$\odot$ 表示逐元素相乘，该操作在不破坏原始特征分布的前提下，实现对边缘及轮廓区域的针对性强化。

通过上述双重感知机制，DuSample模块通过小目标显著性分支对偏移量进行自适应调制，增强对小尺度目标的空间敏感性；同时借助边缘感知分支对采样结果进行结构增强，提升几何细节的恢复质量。该设计实现了对上采样过程的多维度感知引导，有效提升了特征表征的判别能力，显著改善了稀疏点云场景下三维目标检测的性能。

2 实验和结果分析

2.1 实验设置

2.1.1 数据集

本研究选取KITTI（Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago）数据集（Geiger 等，2012）与DAIR-V2X-V数据集（Yu 等，2022）进行实验，以评估所提方法在不同自动驾驶场景下的三维目标检测性能。

KITTI数据集（Geiger 等，2012）由德国卡尔斯鲁厄理工学院和丰田北美技术研究院联合创建，涵盖城市、乡村和高速公路等多种真实驾驶场景。

该数据集提供7481个训练样本和7518个测试样本，每个样本均包含同步采集的图像与激光雷达点云数据及其标注。本文依据Chen等人（2017）的数据划分策略，将7481个训练样本进一步划分为3712个样本的训练集和3769个样本的验证集，所有模型的训练与评估均基于该划分完成。KITTI数据集主要包含汽车、行人与骑行者三个检测类别，并依据3D目标大小、遮挡与截断程度划分为简单（easy）、中等（moderate）与困难（hard）三个难度等级。

DAIR-V2X-V数据集（Yu 等，2022）由百度Apollo团队与清华大学联合发布，旨在支持车路协同环境下的三维感知研究。该数据集属于DAIR-V2X大规模自动驾驶数据集的车端数据部分，包含22325帧图像数据与22325帧点云数据，并提供2D与3D边界框标注信息。原始标注涵盖10个类别，为保持与KITTI数据集的类别一致性，本研究对原有类别进行了合并：将三轮车（tricyclist）、摩托车（motorcyclist）和手推车（barrowlist）统一并入骑行者（cyclist）类别，最终仅保留“汽车”、“行人”和“骑行者”三类作为检测目标。同时，将标注格式转换为KITTI数据集的标注形式，以便于数据读取和后续评价指标的计算。本文参照周昊等人（2024）的数据划分方法，将15627帧具有完整标注的图像数据和点云数据按照1∶1的比例划分为训练集与测试集，分别用于模型训练与性能评估。

2.1.2 评价指标

本文按照KITTI官方评估标准，以平均精度（average precision，AP）作为检测性能的评估指标（Cordts 等，2016）。具体而言，对于汽车类别，IOU（Intersection over Union）阈值设为0.7，对于行人和骑行者类别，IOU阈值设为0.5。AP的计算基于召回率区间（0，1］上均匀分布的40个采样点，

其计算表达式如下：

$$AP=\frac{1}{40}\sum_{r\in\{0,0.025,\cdots,1.0\}}P(r) \tag{15}$$

式中，r表示召回率采样位置{0，0.025，0.05，…，1.0}，$P(r)$表示召回率为r时的精度。

为全面评估模型的整体检测性能，本文进一步计算了各类别在“中等”难度等级下的平均精度均值（mean average precision，mAP），作为综合评估指标。所使用的KITTI和DAIR-V2X-V数据集均采用上述评价指标，对不同难度等级下的各类别计算AP，并以mAP衡量整体检测性能。

2.1.3 实验环境及参数配置

本实验基于OpenPCDet（Nikolovski 等，2021）3D目标检测框架，在Ubuntu 24.04操作系统下完成。硬件平台配置为2×NVIDIA GeForce RTX 4090 GPU（单卡显存24 GB）与Intel Xeon Silver 4310 CPU。软件环境包括Python 3.8、CUDA 11.8及PyTorch 2.0.0。网络结构基于PyTorch框架构建，并采用Adam优化器进行端到端训练。所有训练任务均在单张GPU上执行。主要训练参数如下：批量大小设为4，训练周期为80轮，权重衰减为0.01，初始学习率为0.003，动量为0.9。训练过程中，学习率根据样本规模进行自适应调整。

在点云预处理阶段，本文将点云范围沿x、y、z三个方向分别限制为[0，69.12]m、[-39.68，39.68]m、[-3.0，1.0]m。点云被划分为体素，体素大小设置为(0.16，0.16，4)m。单个体素内最多包含32个点，单帧点云样本的最大非空体素数在训练阶段设为16000。

2.2 实验结果与分析

为评估所提算法的性能，本文在KITTI与

DAIR-V2X-V 两个公开数据集上对比了多种主流三维目标检测方法。

2.2.1 KITTI数据集实验结果

在 KITTI 数据集上，本文将所提模型与多种具有代表性的三维目标检测方法进行对比，涵盖多模态以及基于激光雷达的检测方法。对比方法包括 AVOD（Ku 等，2018）、F-PointNet（Qi 等，2018）、VeloFCN（Li 等，2016）、SECOND（Yan 等，2018）、Associate-3Det（Du 等，2020）、VoxelNet（Zhou 和 Tuzel，2018）、PointRCNN（Shi 等，2019）、HS-Pillar（Fu 等，2021）、TANet（Liu 等，2020）、CenterPoint（Yin 等，2021）、VoxelNeXt（Chen 等，2023）、MAT-PointPillars（Yao 等，2025）、PointBi-FPN（A 等，2024）、XPillars（Zhang 等，2025）和 PointPillars（Lang 等，2019）等15种代表性模型，实验结果如表1所示。

可以看出，本文提出的动态空间全局注意力增强的伪图像特征3D目标检测方法相较于其他方法在汽车、行人和骑行者三个检测类别不同难度等级上均表现出明显的性能提升。与基线模型 PointPillars 相比，从3D视角进行评估，本文方法在简单、中等和困难三个难度等级下汽车类别 AP 分别提升 1.31%、2.51% 和 2.76%；在行人检测中分别提升 6.96%、6.07% 和 6.10%；同样地，在骑行者检测中分别提升 10.62%、10.33% 和 9.81%。总体而言，本文方法在三个目标类别上均表现出更高的检测精度，尤其在行人骑行者等小目标检测中效果更为显著。在中等难度等级下，本文所提算法的 mAP 达到 65.44%，较基线模型提升 6.3 个百分点。由上述分析可知，本文所提算法在多类别、多难度等级下均取得了优于主流对比方法的检测性能，尤其在小目标检测中表现更加突出，充分验证了所提算法的有效性。

2.2.2 DAIR-V2X-V数据集实验结果

在 DAIR-V2X-V 测试集上，本文将所提模型与多种具有代表性的三维目标检测方法进行对比，包括 Part-A2（Shi 等，2020）、SECOND、PointRCNN 和 PointPillars 等检测方法，实验结果如表 2 所示。

可以看出，本文提出的动态空间全局注意力增强的伪图像特征3D目标检测方法在 DAIR-V2X-V 数据集上同样表现出优异的检测性能。在汽车、行人和骑行者三个检测类别中，本文方法均取得了优于对比算法的检测精度。其中，在汽车类别上，简单、中等和困难三个难度等级下的 AP 分别提升 0.08%、0.07% 和 0.30%；在行人检测中分别提升 7.81%、6.48% 和 6.04%；在骑行者检测中分别提升 7.45%、8.70% 和 8.64%。总体来看，本文方法在 DAIR-V2X-V 数据集上的 mAP 达到 44.37%，较基线模型提升 5.09 个百分点。该实验结果表明，本文所提算法不仅在 KITTI 数据集上取得了较好效果，在 DAIR-V2X-V 数据集上同样展现出较强的适用性，进一步验证了所提算法的有效性。

2.2.3 模型计算开销与复杂度分析

为进一步验证所提模型在计算开销与复杂度方面的合理性，本文在 KITTI 数据集中等难度下，将其与 PointPillars、SECOND、PointRCNN 及 HybridPillars（Huang 等，2024）等代表性 3D 目标检测方法对比，从参数量、推理速度和检测精度三个维度展开综合分析，实验结果如表3所示。

由表 3 可知，所提模型通过引入动态空间全局注意力、高效多尺度伪图像特征融合网络和双重感知动态上采样模块，在检测精度上取得了一定提升。尽管模型参数量略有增加，推理速度因复杂度的提升而有所下降，但其帧率仍稳定满足自动驾驶场景的实时检测要求。总体而言，模型在实现精度优化的同时，计算开销与推理效率均维持在可控范围内，较好地兼顾了检测性能与实际部署的实时性需求。

2.3 消融实验

为深入分析动态空间全局注意力机制、高效多尺度伪图像特征融合网络、以及双重感知动态上采样模块对目标检测性能的作用，本文开展了一系列消融实验，以系统评估各模块对模型性能的贡献。实验以 PointPillars 为基线模型，对不同变体进行了标注，分别记为 Baseline、模型 A、模型 B、模型 C、模型 D 和 模型 E（本文方法）。在 KITTI 验证集上，从 3D、鸟瞰图（bird's eye view，BEV）和平均方向相似度（average orientation similarity，AOS）三个视角进行实验分析，实验结果如表 4、表 5 和表 6 所示。

观察表4—表6的实验结果可以发现，Baseline 模型的表现相对较弱，其 mAP 指标在 3D、BEV 和 AOS 三个视角下分别为 59.14%、66.88% 和 67.58%。相比之下，模型 A 使用多尺度伪图像特征融合网络替换骨干网络，在 3D 视角下使 mAP 提升 1.73%，其中骑行者检测精度提升尤为显著，验证了该多尺度特征融合网络在特征提取方面的优势。模

表1　KITTI数据集上不同算法的AP对比

Table 1　AP comparison of different algorithms on the KITTI dataset

方法	模态	AP_{car}(%)			$AP_{pedestrain}$(%)			$AP_{cyclist}$(%)			mAP(%)
		简单	中等	困难	简单	中等	困难	简单	中等	困难	
AVOD	图像+点云	83.07	71.76	65.73	36.10	27.86	25.76	57.19	42.08	38.29	47.23
F-PointNet	图像+点云	81.20	70.39	62.19	51.21	44.89	40.23	71.96	56.77	50.39	57.35
VeloFCN	点云	15.20	13.66	15.98	N/A	N/A	N/A	N/A	N/A	N/A	N/A
SECOND	点云	86.44	76.97	73.39	47.47	40.47	36.26	81.28	63.49	59.29	60.31
Associate-3Det	点云	85.99	77.40	70.53	N/A	N/A	N/A	N/A	N/A	N/A	N/A
VoxelNet	点云	77.47	65.11	57.73	39.48	33.69	31.50	61.22	48.36	44.37	49.05
PointRCNN	点云	85.13	74.59	68.79	53.46	47.86	43.30	74.73	62.45	56.31	61.63
HS-Pillar	点云	86.05	75.66	72.79	44.60	40.76	37.53	N/A	N/A	N/A	N/A
TANet	点云	85.39	76.94	69.82	54.87	48.53	44.61	80.48	61.40	58.89	62.29
CenterPoint	点云	86.93	77.04	75.50	49.30	47.02	44.40	80.42	61.54	58.85	61.87
VoxelNeXt	点云	83.88	75.58	70.77	47.46	39.97	37.43	78.18	61.74	54.68	59.10
MAT-PointPillars	点云	N/A	N/A	N/A	N/A	N/A	N/A	83.03	66.38	58.07	N/A
PointBi-FPN	点云	86.28	76.74	73.86	51.87	44.86	41.12	75.46	61.13	50.18	60.91
XPillars	点云	87.72	76.24	73.28	54.55	47.47	42.74	80.24	66.18	62.03	63.30
PointPillars	点云	86.52	75.75	72.74	50.07	43.91	39.13	74.30	57.76	53.84	59.14
Ours	点云	**87.83**	**78.26**	**75.50**	**57.03**	**49.98**	**45.23**	**84.92**	**68.09**	**63.65**	**65.44**

注：加粗字体表示各列最优检测结果，“N/A”表示没有相关数据。

表2　DAIR-V2X-V数据集上不同算法的AP对比

Table 2　AP comparison of different algorithms on the DAIR-V2X-V dataset

方法	模态	AP_{car}(%)			$AP_{pedestrain}$(%)			$AP_{cyclist}$(%)			mAP(%)
		简单	中等	困难	简单	中等	困难	简单	中等	困难	
Part-A2	点云	66.18	55.45	51.52	37.29	34.69	34.12	35.34	34.27	32.05	41.47
SECOND	点云	69.04	57.37	54.62	38.69	35.62	35.47	40.81	38.26	36.05	43.75
PointRCNN	点云	69.19	55.22	50.09	34.45	28.04	26.57	34.84	28.37	26.53	37.21
PointPillars	点云	69.36	57.60	54.68	32.67	30.38	30.41	33.86	29.87	27.88	39.28
Ours	点云	**69.44**	**57.67**	**54.98**	**40.48**	**36.86**	**36.45**	**41.31**	**38.57**	**36.52**	**44.37**

注：加粗字体表示各列最优检测结果，“N/A”表示没有相关数据。

型B加入动态空间全局注意力机制，通过动态卷积核生成与全局自注意力机制进一步提升了检测性能，表明该模块在空间特征建模方面的有效性。模型C采用双重感知动态上采样模块，在3D视角下带来4.37%的mAP提升，同时将AOS视角下困难难度行人的AP提升3.21%，证明该模块在优化困难样本方向感知与特征质量方面的有效性。模型A、B和C均有效提高了目标检测性能，充分验证了

表3　模型检测精度、参数量和推理速度对比

Table3　Model comparisons in terms of detection accuracy, parameter size, and inference speed

方法	参数量(M)	推理速度(FPS)	mAP(%)
PointPillars	4.8	42.06	59.14
SECOND	5.4	27.87	60.31
PointRCNN	4.0	10.00	61.83
HybridPillars	7.1	21.83	N/A
Ours	5.9	29.10	65.44

各模块的作用。模型D同时集成了模型B和模型C两种策略，在3D、BEV和AOS视角下的mAP分别较基线提升5.36%、4.70%和6.37%。模型E（本文方法）则将模型A、B和C三种策略联合应用，通过协同效应实现了更显著的性能提升。与基线模型相比，模型E在3D视角下对汽车、行人和骑行者在三个难度级别的AP均有显著提升，整体mAP达到65.44%，较基线提升6.30%；在BEV和AOS视角下，mAP分别提升至72.50%和74.43%，显示出最优性能。

消融实验从多维度证明了各模块及其组合的有效性。动态空间全局注意力机制、高效多尺度伪图像特征融合网络以及双重感知动态上采样模块分别从空间建模、特征提取和特征优化三个关键层面提升模型性能，其协同工作最终使检测精度在不同评估视角、难度及类别上均取得显著进步，充分验证了本文所提算法的有效性。

2.4　可视化结果与分析

为直观验证所提动态空间全局注意力增强的伪图像特征3D目标检测模型的有效性与优越性，本文在KITTI与DAIR-V2X-V数据集上分别选取四个典型场景，将本文方法与基准模型PointPillars的检测结果进行可视化对比，结果如图5与图6所示。图中采用统一标注方式，绿色边界框表示汽车，蓝色表示行人，黄色表示骑行者；圆圈标识误检，方框标识漏检，三角形标识目标朝向检测错误。

由图5和图6可视化结果可知，本文所提方法在多种挑战性场景下均表现出显著优势。在目标密集与存在遮挡的场景中，原始PointPillars算法因特征感知能力有限，出现显著漏检现象。由图5场景四中可见，指示牌遮挡处的行人在原始算法中未被检出；图6场景一与场景三也显示，多目标密集区域存在部分车辆漏检。相比之下，本文算法在上述场景中能够更完整地识别目标，有效提升了召回率。在复杂背景与远距离场景下，原始算法易受干扰而产生误检。如图5场景一所示，原始算法在远距离稀疏场景中出现错误检测；图5场景三中，其将路旁的树干树枝误识别为骑行者；图6场景二中，长条状矮灌木丛亦被误判为汽车。本文算法通过增强对细节特征的保留与上下文感知能力，抑制此类误检，提升检测结果的可靠性。在目标朝向估计方

表4　KITTI数据集上消融实验3D视角下的检测精度

Table 4　The detection accuracy of ablation experiments from a 3D perspective on the KITTI dataset

方法	多尺度伪图像特征融合网络	动态空间全局注意力	双重感知动态上采样	AP_{car}(%)			$AP_{pedestrain}$(%)			$AP_{cyclist}$(%)			mAP(%)
				简单	中等	困难	简单	中等	困难	简单	中等	困难	
Baseline	×	×	×	86.52	75.75	72.74	50.07	43.91	39.13	74.30	57.76	53.84	59.14
模型A	√	×	×	87.01	76.00	73.16	51.28	43.94	39.59	79.06	62.67	58.45	60.87
模型B	×	√	×	87.27	77.75	74.65	51.96	45.60	41.25	77.68	62.09	57.82	61.81
模型C	×	×	√	86.94	77.00	73.18	55.43	47.67	43.20	81.25	65.85	61.38	63.51
模型D	×	√	√	87.31	78.19	75.43	56.13	48.98	43.92	83.44	66.34	61.67	64.50
模型E(本文)	√	√	√	**87.83**	**78.26**	**75.50**	**57.03**	**49.98**	**45.23**	**84.92**	**68.09**	**63.65**	**65.44**

注："√"表示引入该模块，"×"表示不引入该模块。

表5　KITTI数据集上消融实验BEV视角下的检测精度

Table5　The detection accuracy of ablation experiments from a BEV perspective on the KITTI dataset

方法	多尺度伪图像特征融合网络	动态空间全局注意力	双重感知动态上采样	AP_{car}(%)			$AP_{pedestrain}$(%)			$AP_{cyclist}$(%)			mAP(%)
				简单	中等	困难	简单	中等	困难	简单	中等	困难	
Baseline	×	×	×	91.20	87.76	85.12	57.08	50.75	46.21	78.57	62.13	57.9	66.88
模型A	√	×	×	91.45	87.50	85.01	57.64	50.22	46.06	85.67	68.59	63.9	68.77
模型B	×	√	×	91.80	88.03	86.66	58.22	51.96	47.96	81.46	66.38	61.86	68.79
模型C	×	×	√	91.70	87.77	85.19	61.85	54.43	50.22	87.40	71.15	66.22	71.12
模型D	×	√	√	92.14	88.05	85.31	62.23	55.28	51.32	86.40	71.40	66.72	71.58
模型E(本文)	√	√	√	**92.28**	**88.37**	**85.50**	**63.40**	**56.15**	**51.71**	**88.93**	**72.97**	**68.13**	**72.50**

注：“√”表示引入该模块，“×”表示不引入该模块。

表6　KITTI数据集上消融实验AOS视角下的检测精度

Table6　The detection accuracy of ablation experiments from a AOS perspective on the KITTI dataset

方法	多尺度伪图像特征融合网络	动态空间全局注意力	双重感知动态上采样	AP_{car}(%)			$AP_{pedestrain}$(%)			$AP_{cyclist}$(%)			mAP(%)
				简单	中等	困难	简单	中等	困难	简单	中等	困难	
Baseline	×	×	×	95.23	91.31	88.45	47.38	42.65	39.67	83.65	68.78	65.06	67.58
模型A	√	×	×	95.09	91.27	88.58	51.10	45.83	42.88	88.19	73.23	68.84	70.11
模型B	×	√	×	95.36	**91.82**	90.33	56.81	52.29	49.03	85.14	72.17	68.48	72.09
模型C	×	×	√	95.11	91.22	89.63	51.02	45.64	42.88	87.32	74.17	69.46	70.34
模型D	×	√	√	95.28	91.71	90.28	60.34	53.09	50.16	89.88	77.06	72.39	73.95
模型E(本文)	√	√	√	**95.57**	**91.82**	**90.36**	**61.57**	**54.16**	**51.76**	**90.16**	**77.30**	**72.63**	**74.43**

注：“√”表示引入该模块，“×”表示不引入该模块。

面，原始算法的检测框回归能力存在不足。从图5场景二与图6场景四中可以看出，原始算法对车辆朝向的估计存在明显偏差。相比之下，本文算法展现出更优的方向感知能力。

综上所述，本文所提动态空间全局注意力增强的伪图像特征3D目标检测模型能够有效缓解原始

模型在复杂场景下的漏检、误检与朝向估计偏差问题，整体检测性能得到显著提升。

3　结　论

针对点云经柱体化映射为二维伪图像时易导致的几何细节丢失，以及伪图像特征提取阶段建模不足的问题，提出一种动态空间全局注意力增强的伪图像特征3D目标检测模型。首先，提出了一种动态空间全局注意力机制，结合动态卷积核生成与全局自注意力机制，有效提升伪图像特征的空间表达能力。其次，设计高效多尺度伪图像特征融合网络，在兼顾细粒度局部特征和全局语义的同时，增强模型对多尺度特征的提取能力。最后，提出了双重感知

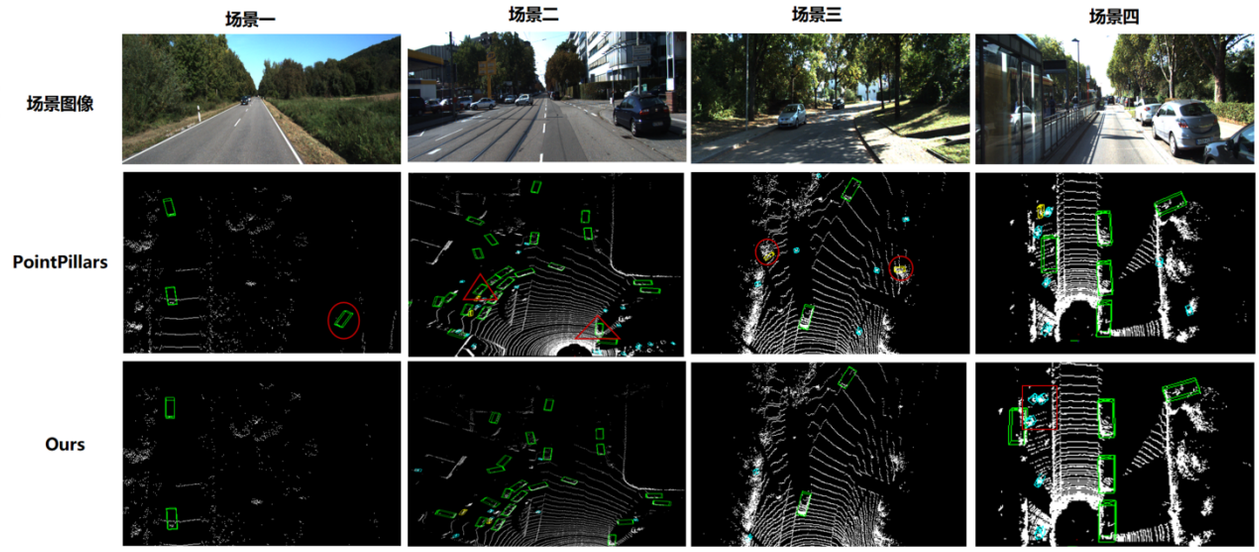


图5　KITTI数据集可视化结果对比

Fig. 5　Comparison of visualization results of the KITTI dataset

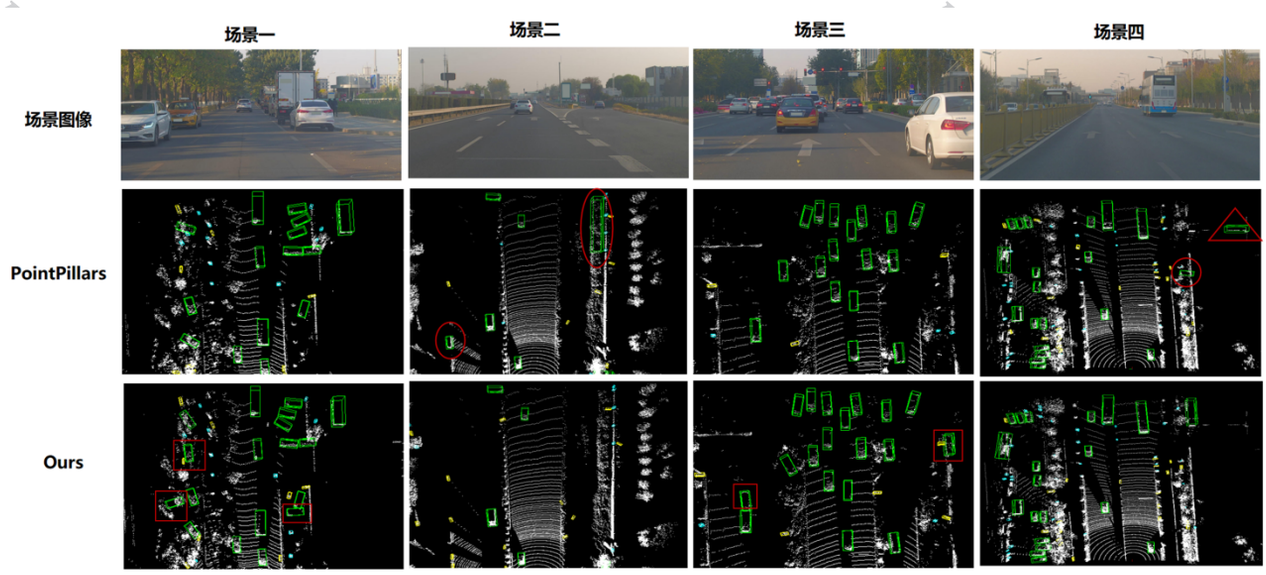


图6　DAIR-V2X-V数据集可视化结果对比

Fig. 6　Comparison of visualization results of the DAIR-V2X-V dataset

动态上采样，通过边缘感知与小目标显著性增强双通路设计，引导采样点分布，优化上采样过程并提升空间细节恢复与目标检测性能。在KITTI与DAIR-V2X-V数据集上的实验结果表明，本文算法均取得了较优的检测性能，并通过消融实验和可视化结果进一步验证了各模块的有效性。尽管显著提高了目标检测精度，但动态空间全局注意力与多尺度特征融合机制的引入，也相应带来了模型复杂度与参数量的提升。未来的研究工作将探索模型的轻量化设计，进一步提升算法的实时性与鲁棒性，并增强其在多样化复杂场景下的适用性；同时，进一步拓展真实开放场景下的测试与验证范围，系统评估模型在复杂环境中的鲁棒性与泛化能力，以增强其在实际自动驾驶场景中的应用适应性。

参考文献（References）

A S, C K M, Cenkeramaddi L R and Babu S. 2024. PointBi-FPN: An Extention to Pointpillars for LiDAR 3D Object Detection in Autonomous Vehicles Using Bi-Directional Feature Pyramid Network// Proceedings of the TENCON 2024-2024 IEEE Region 10 Confer-

ence. Singapore, Singapore: IEEE: 1950-1953 [DOI: 10.1109/TENCON61640.2024.10902951]

Charles R Q, Su H, Kaichun M and Guibas L J. 2017. PointNet: deep learning on point sets for 3D classification and segmentation//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 77-85 [DOI: 10.1109/CVPR.2017.16]

Chen L, Lin S B, Lu X K, Cao D P, Wu H B, Guo C, Liu C and Wang FY. 2021. Deep Neural Network Based Vehicle and Pedestrian Detection for Autonomous Driving: A Survey. IEEE Transactions on Intelligent Transportation Systems, 22(6): 3234-3246[DOI:10.1109/tits.2020.2993926]

Chen X Z, Kundu K, Zhang Z Y, Ma H M, Fidler S and Urtasun R. 2016. Monocular 3d object detection for autonomous driving//Proceedings of the 2016 IEEE conference on computer vision and pattern recognition. Las Vegas, USA: IEEE: 2147-2156 [DOI: 10.1109/CVPR.2016.236]

Chen X Z, Ma H M, Wan J, Li B and Xia T. 2017. Multi-view 3D object detection network for autonomous driving//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6526-6534 [DOI: 10.1109/CVPR.2017.691]

Chen Y K, Liu J H, Zhang X Y, Qi X J and Jia J Y. 2023.VoxelNeXt: Fully Sparse VoxelNet for 3D Object Detection and Tracking//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 21674-21683 [DOI:10.1109/CVPR52729.2023.02076]

Chen Y P, Dai X Y, Liu M C, Chen D D, Yuan L and Liu Z C. 2020. Dynamic Convolution: Attention Over Convolution Kernels//Proceedings of the 2020 IEEE/CVF conference on computer vision and pattern recognition. Seattle, USA: IEEE: 11027-11036 [DOI: 10.1109/CVPR42600.2020.01104]

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S and Schiele B. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA:IEEE:3213-3223[DOI:10.1109/CVPR.2016.350]

Du L, Ye X Q, Tan X, Feng J F, Xu Z B, Ding E and Wen S L. 2020. Associate-3Ddet: Perceptual-to-Conceptual Association for 3D Point Cloud Object Detection//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13326-13335 [DOI: 10.1109/CVPR42600.2020.01334]

Fu J H, Ren G H, Chen Y P and Liu S. 2021. Improved Pillar with Fine-grained Feature for 3D Object Detection [EB/OL]. [2025-11-05]. http://arxiv.org/pdf/2110.06049.pdf

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

Gong J Y, Lou Y J, Liu F Q, Zhang Z W, Chen H M, Zhang Z Z, Tan X, Xie Y and Ma L Z. 2023. Scene point cloud understanding and reconstruction technologies in 3D space. Journal of Image and Graphics, 28(6): 1741-1766 (龚靖渝, 楼雨京, 柳奉奇, 张志伟, 陈豪明, 张志忠, 谭鑫, 谢源, 马利庄. 2023. 三维场景点云理解与重建技术. 中国图象图形学报, 28(6): 1741-1766) [DOI: 10.11834/jig.230004]

Huang Z C, Huang Y X, Zheng Z J, Hu H F and Chen D H. 2024. HybridPillars: Hybrid Point-Pillar Network for Real-Time Two-Stage 3-D Object Detection.IEEE Sensors Journal ,24(12): 38318-38328[DOI:10.1109/JSEN.2024.3468646]

Ku J, Mozifian M, Lee J, Harakeh A and Waslander S L. 2018. Joint 3D proposal generation and object detection from view aggregation//Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems. Madrid, Spain: IEEE: 1-8[DOI: 10.1109/IROS.2018.8594049]

Lang A H, Vora S, Caesar H, Zhou L B, Yang J and Beijbom O. 2019. PointPillars: fast encoders for object detection from point clouds//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 12689-12697[DOI:10.1109/CVPR.2019.01298]

Li B, Zhang T L and Xia T. 2016.Vehicle detection from 3d lidar using fully convolutional network[EB/OL].[2025-11-05]. https://arxiv.org/pdf/1608.07916.pdf

Li J Y, Luo C X and Yang X D. 2023. PillarNeXt: Rethinking Network Designs for 3D Object Detection in LiDAR Point Clouds//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 17567-17576 [DOI:10.1109/CVPR52729.2023.01685]

Li P X, Zhao H C, Liu P F and Cao F D. 2020. RTM3D: Real-Time Monocular 3D Detection from Object Keypoints for Autonomous Driving//Proceedings of the 16th European Conference on Computer Vision.Glasgow, UK: Springer: 644-660[DOI: 10.1007/978-3-030-58580-8_38]

Li Y Y, Hu J, Wen Y, Evangelidis G, Salahi K, Wang Y Z, Tulyakov S and Ren J. 2023. Rethinking Vision Transformers for MobileNet Size and Speed//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 16843-16854 [DOI:10.1109/ICCV51070.2023.01549]

Liu W Z, Lu H, Fu H T and Cao Z G. 2023. Learning to Upsample by Learning to Sample//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 6004-6014[DOI:10.1109/ICCV51070.2023.00554]

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY and Berg A C. 2016. SSD: Single Shot MultiBox Detector//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-

0_2]

Liu Z, Zhao X, Huang T T, Hu R L, Zhou Y, Bai X. 2020. TANet: Robust 3D Object Detection from Point Clouds with Triple Attention//Proceedings of the AAAI Conference on Artificial Intelligence, 34(07):11677-11684[DOI:10.1609/aaai.v34i07.6837]

Nikolovski G, Reke M, Elsen I and Schiffer S. 2021. Machine learning based 3D object detection for navigation in unstructured environments// Proceedings of the 2021 IEEE Intelligent Vehicles Symposium Workshops. Nagoya, Japan: IEEE: 236-242[DOI: 10.1109/IVWorkshops54471.2021.9669218]

Qi C R, Liu W, Wu C X, Su H and Guibas L J. 2018. Frustum PointNets for 3D Object Detection from RGB-D Data//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 918-927[DOI: 10.1109/CVPR.2018.00102]

Reading C, Harakeh A, Chae J and Waslander S L. 2021.Categorical Depth Distribution Network for Monocular 3D Object Detection// Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition.Nashville, USA:IEEE:8551-8560[DOI:10.1109/CVPR46437.2021.00845]

Shi S S, Wang X G and Li H S. 2019. PointRCNN:3D object proposal generation and detection from point cloud//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach,USA:IEEE:770-779[DOI:10.1109/CVPR. 2019.00086]

Shi S S, Wang Z, Shi J P, Wang X G and Li H S. 2020. From Points to Parts: 3D Object Detection from Point Cloud with Part-aware and Part-aggregation Network. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(8): 2647-2664[DOI: 10.1109/TPAMI.2020.2977026]

Tao S B, Liang C, Jiang T P, Yang Y J and Wang Y J. 2021. Sparse voxel pyramid neighborhood construction and classification of LiDAR point cloud. Journal of Image and Graphics, 26(11):2703-2712(陶帅兵,梁冲,蒋腾平,杨玉娇,王永君. 2021. 激光点云的稀疏体素金字塔邻域构建与分类. 中国图象图形学报, 26(11):2703-2712)[DOI:10. 11834 / jig. 200262]

Wang Q L, Wu B G, Zhu P F, Li P H, Zuo W M and Hu Q H. 2020. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks//Proceedings of the 2020 IEEE/CVF conference on computer vision and pattern recognition. Seattle, USA: IEEE: 11531-11539[DOI: 10.1109/CVPR42600.2020.01155]

Wang Y J, Mao Q Y, Zhu H Q, Deng J J, Zhang Y, Ji J M, Li H Q and Zhang Y Y. 2023. Multi-Modal 3D Object Detection in Autonomous Driving: A Survey. International Journal of Computer Vision, 131(8): 2122-2152[DOI:10.1007/s11263-023-01784-z]

Xu S Q, Jiang S Y, Li F, Liu L, Song Z Y, Yang B and Yang ZX. 2024. SparseInteraction: Sparse Semantic Guidance for Radar and Camera 3D Object Detection//Proceedings of the 32nd ACM International Conference on Multimedia. New York, USA: Association for Computing Machinery: 9224-9233[DOI: 10.1145/3664647.3681565]

Yan Y, Mao Y X and Li B. 2018. Second: sparsely embedded convolutional detection. Sensors, 18(10):3337[DOI:10.3390/ s18103337]

Yao X P, Liu P Y, Zhou J M, Wang Z J, Fan S H and Wang Y C. 2025. MAT-PointPillars: Enhanced PointPillars algorithm based on multi-scale attention mechanisms and transformer. PLOS One, 20(6): e0325373[DOI:10.1371/journal.pone.0325373]

Yin T W, Zhou X Y and Krahenbuhl P. 2021. Center-based 3D Object Detection and Tracking//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 11779-11788[DOI: 10.1109/CVPR46437.2021.01161]

Yu H B, Luo Y Z, Shu M, Huo Y Y, Yang Z B, Shi Y F, Guo Z L, Li H Y, Hu X, Yuan J R and Nie Z Q. 2022. DAIR-V2X: a largescale dataset for vehicle-infrastructure cooperative 3d object detection//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 21329-21338[DOI:10.1109/CVPR52688.2022.02067]

Yu X M, Tang L L, Rao Y M, Huang T J, Zhou J and Lu J W. 2022. Point-BERT: Pre-Training 3D Point Cloud Transformers With Masked Point Modeling//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition.New Orleans, USA: IEEE: 19291-19300[DOI: 10.1109/CVPR52688.2022.01871]

Zhang L J, Fu Z H, Li Z Y and Li D M. 2025.XPillars: enhancing 3D object detection through cross-pillar feature fusion.The Visual Computer, 41(14):11977-11991[DOI:10.1007/s00371-025-04138-7]

Zhou H, Qi H G, Deng Y Q, Li J J, Liang H and Miao J. 2024. 3D object detection and classification combined with point cloud depth information. Journal of Image and Graphics, 29(08):2399-2412(周昊,齐洪钢,邓永强,李娟娟,梁浩,苗军. 2024. 融合点云深度信息的3D目标检测与分类. 中国图象图形学报, 29(08):2399-2412)[DOI:10. 11834/jig. 230568]

Zhou Y and Tuzel O. 2018. VoxelNet: end-to-end learning for point cloud based 3D object detection//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4490-4499[DOI: 10.1109/CVPR. 2018.00472]

作者简介

李幸,通信作者,女,硕士研究生,主要研究方向为计算机视觉、三维目标检测。 E-mail:2592449968@qq.com

郑钰辉,男,教授,主要研究方向为图像视频分析、场景理解、视觉跟踪和模式识别。 E-mail:zhengyh@vip.126.com