基于转移分块 Transformer 和特征金字塔的点-体素三维目标检测方法

刘良杰[1]，孙 凯[2]，任宏乾[3]，谢国涛[4*]

[1]株洲中车时代软件技术有限公司 湖南株洲
[2]国家能源集团陕西神延煤炭有限责任公司西湾露天煤矿 陕西榆林
[3]湖南大学机械与运载工程学院 湖南长沙
[4]湖南大学无锡智能控制研究院 江苏无锡

【摘要】随着环境感知技术的发展，激光雷达三维目标检测取得了显著进展。然而，基于体素的三维检测器在划分点云时，难以捕捉丰富的上下文信息和细节特征，尤其在处理遮挡和截断问题时，原始点云的细节信息常常丢失。为解决这些挑战，本文提出了一种新型数据增强策略，增强了模型对不完整点云的处理能力；并提出了基于转移分块 Transformer 和特征金字塔的点-体素三维目标检测模型 PV-FMRTNet，有效解决了点云转换为体素过程中位置信息丢失的问题。此外，设计了一种新的二维特征编码网络，提升了基于体素的三维目标检测系统的性能。评估结果显示，本文模型在检测汽车、行人和骑行者方面的准确度分别达到 84.30%、61.76%和 78.08%，相比主流算法 PointPillars 等基准模型平均提升 2.08%，展现出先进的准确性和鲁棒性。

【关键词】自动驾驶；深度学习；三维目标检测；特征金字塔；点体素

【基金项目】国家自然科学基金项目（青年科学基金项目，项目名称：融合意图推断与时空交互建模的行人轨迹预测方法，项目批准号：52102456）

【收稿日期】2024 年 12 月 16 日 **【出刊日期】**2025 年 1 月 19 日 **【DOI】**10.12208/j.jer.20250007

Point-voxel 3D object detection method based on transfer block Transformer and feature pyramid

Liangjie Liu[1], Kai Sun[2], Hongqian Ren[3], Guotao Xie[4]*

[1]Zhuzhou CRRC Times Software Technology Co., Ltd., Zhuzhou, Hunan
[2]Xiwan Open-pit Coal Mine of Shanxi Shenyan Coal Co., Ltd. of China National Energy Group, Yulin, Shaanxi
[3]HNU College of Mechanical and Vehicle Engineering, Hunan University, Changsha, Hunan
[4]Wuxi Intelligent Control Research Institute, HNU, Wuxi, Jiangsu

【Abstract】With the development of environmental sensing technology, lidar three-dimensional target detection has made significant progress. However, it is difficult for voxel-based 3D detectors to capture rich contextual information and detailed features when dividing point clouds. Especially when dealing with occlusion and truncation problems, the detailed information of the original point cloud is often lost. To address these challenges, this paper proposes a new data augmentation strategy to enhance the model's ability to handle incomplete point clouds; and proposes a point-to-voxel 3D target detection model PV-FMRTNet based on transfer block Transformer and feature pyramid, which effectively solves the problem of position information loss in the process of converting point clouds to voxels. In addition, a new 2D feature encoding network was designed to improve the performance of the voxel-based 3D object detection system. The evaluation results show that the accuracy of the proposed model in detecting cars, pedestrians and cyclists reached 84.30%, 61.76% and 78.08% respectively, which is an average improvement of 2.08% over the mainstream algorithm PointPillars and other benchmark models, showing advanced

作者简介：刘良杰（1982-）男，汉族，湖南省郴州人，正高级工程师，工学学士学位，主要研究方向为轨道交通传动系统、目标检测、智能化系统设计与研究；孙凯（1988-）男，汉族，陕西省子长市人，工程师，工学学士学位，主要研究方向为机电智能化研究与设计；任宏乾（1997-）男，汉族，河北省石家庄人，工程硕士学位，主要研究方向为激光雷达点云 3D 目标检测；
*通讯作者：谢国涛（1987-）男，土家族，湖北利川人，副研究员，硕导，博士，主要研究方向为智能网联技术、环境感知。

accuracy and robustness.

【**Keywords**】Autonomous driving; Deep learning; 3D object detection; Feature pyramid; Point voxel

引言

目前自动驾驶技术已经在全球范围内引起了广泛的关注，并且正逐步地重塑着当前汽车生态系统的构成[1-2]，在自动驾驶系统架构中，环境感知模块构成了自动驾驶技术实施的初始阶段[3]。目标检测是自动驾驶环境感知的关键步骤，涉及车辆行驶途中所有关键目标的具体位置、尺寸及类别[4]。激光雷达作为一种主动型传感器[5]，可以得到具有无序性、非结构化、稀疏性、不均匀等特点的点云[6]。采用激光雷达技术进行障碍物检测，能够有效克服光照条件变化带来的影响。因此，在自动驾驶上使用激光雷达进行三维目标检测已经成为了趋势[7,8]。

1 相关工作

1.1 基于点集的三维目标检测方法

基于点集的三维目标检测方法可以直接处理原始点云数据，避免了数据预处理的复杂性。为了高效地从三维点云中学习更有效的语义特征，斯坦福大学的 Charles 等人[9]提出了 PointNet 算法，以端到端的形式处理三维点云数据以及分类、分割任务。尽管该方法在全局特征的提取方面表现出色，但它在捕捉细粒度特征方面存在局限性，所以 Charles 等人[10]又进一步提出了改进的 PointNet++算法，其在局部区域内重复且迭代地应用 PointNet 框架，旨在通过在较小的空间区域内使用 PointNet，生成富含局部特征的新点集，从而促进多级特征学习的实现。在此基础上，Frustum PointNet[11]引入了一种创新的方法，通过结合 PointNet 作为核心的特征提取机制，实现了高效的目标检测。最近，DFAF3D[13]提出了一个创新的双空间-上下文特征提取模块，能够同时从点云中提取空间和上下文特征，优化三维目标检测过程。以上的方法逐个点的处理会使得计算和存储需求急剧增加，这限制了这些方法在实时应用或者处理大规模场景时的效率[14]。

1.2 基于体素的三维目标检测方法

基于体素的三维目标检测方法将点云转换为三维体素（Voxel）网格，并对每个体素进行特征提取和目标检测。Zhou 等人[15]提出 VoxelNet 算法，该模型首次将不规则的点云转换为规则的体素表示。考虑到计算量过大， Lang 等人[17]直接舍弃了三维稀疏卷积神经网络，提出了一种端到端的检测模型 PointPillars，极大地提升了点云数据处理地效率。近期，Zhou 等人[18]采用了创新方法来增强 pillar 的表示能力。他们对每个 pillar 进行最大池化处理，并引入了强大的注意力机制来自动识别关键的局部特征，适用于处理大规模的点云数据。然而，需要注意的是，体素化过程可能会导致一定程度的信息丢失，特别是当体素的尺寸较大时，可能会损失细节信息，从而影响最终目标检测的精度。

1.3 基于点与体素融合的三维目标检测方法

基于点与体素融合的三维目标检测方法目前已经成为一种三维目标检测的研究领域中一种有效的策略。这种混合方法旨在综合两种方法的优势：三维稀疏卷积的高效性与基于点的方法在获取精细局部信息上的灵活性。Shi 等人[19]提出了 PV-RCNN，首次将基于点和基于体素的三维目标检测方法相结合。随后 Deng 等人[20]提出了 Voxel-RCNN，更加高效地表示并集合为栅格特征。这种融合体素与点集优势的方法显著提升了点云数据分析的精度与效率，并在实际应用中展现了优异的性能。

尽管三维目标检测领域取得了快速进展，但现有模型在处理遮挡和截断问题时仍存在精度不足。为提升三维目标检测性能，本文提出了 PV-FMRTNet （Point-Voxel 3D Object Detection Using FPN-PAN with Moving Region Transformer）模型，该模型融合了点云与体素的三维特征编码网络，充分利用原始点云数据，并结合鸟瞰图特征，设计了融合转移分块 Transformer 与 FPN-PAN 特征金字塔的二维特征编码网络，有效提取多尺度局部信息，增强多尺度目标检测能力。此外，本文开发了点云随机区域消除技术，提升了模型对不完整点云的处理能力，显著增强了模型的鲁棒性和泛化性。

2 基于随机点云消除的数据增强方法

为解决不完整点云导致的性能下降问题，我们借鉴图像处理中应对遮挡和截断的数据增强方法，提出了一种创新的随机点云消除策略。该策略在训练阶段随机移除部分点云数据，模拟物体被遮挡或

截断的情形，旨在训练深度学习模型从部分或不完整的点云中推断出物体的完整形状信息。作为一个即插即用模块，该方法易于集成到现有的三维目标检测框架中，以提升模型在复杂场景中的处理能力。具体来说，基于激光雷达捕获的点云数据特性——点云主要分布于物体表面，我们对选定的目标真值边界框内的点云子集进行随机消除操作。此操作模拟物体表面点云的缺失现象，如遮挡和截断，从而增强深度学习模型的适应性和鲁棒性。基于此原理，方案以随机概率对选定的真值目标边界框部分进行如下操作：

（1）首先，将目标的真值边界框划分为 8 个区块，随机选择并删除其中一个区块的全部点云（如图 1 所示的红色区域），以模拟物体遭受部分遮挡的情况，以帮助网络从剩余的点推断出完整的形状。

（2）接着，将真值边界框随机划分为上下或左右三块，随机删除这些划分中的一块的所有点云（如图 1 所示的绿色区域），以模拟物体遭受截断的情况，以帮助网络从剩余的点推断出完整的形状。

（3）为了保持操作的现实性，遮挡和截断模拟不会同时进行。此外，考虑到点云数据的固有稀疏性，此操作仅应用于检测范围内前一半的物体，以确保增强操作的有效性和合理性。

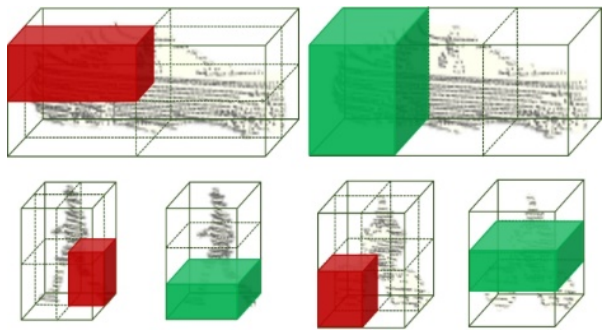

图 1　基于点云随机消除的数据增强示意图

3 融合转移分块 Transformer 的点-体素融合三维目标检测方法

3.1 PV-FMRTNet 三维目标检测模型

为了克服现有三维目标检测算法在处理大规模点云数据时的局限性，本文提出了一种新的基于点-体素融合的三维目标检测模型——PV-FMRTNet。该模型首先对体素化后的点云进行编码，结合残差网络和原始点云位置信息，聚合成体素特征。然后，通过三维稀疏卷积网络进行体素特征学习，有效利用点云数据中的原始信息，恢复三维结构的上下文，并提取局部点云信息，从而学习到更丰富的判别性特征和细粒度信息。此外，模型还有效利用鸟瞰图像的二维特征编码层，聚焦不同尺度的局部和上下文信息；最后，使用 FPN-PAN 特征金字塔将不同尺度的自注意力特征图联系起来，送入多头区域提议网络，以更好地提取不同尺度的目标。整体框架如图 2 所示。

3.2 基于点-体素融合的细粒度三维特征编码网络

在三维物体检测中，目前检测精度较高的框架会将无序点云转化为规则体素，并将其发送到三维稀疏卷积中，这将不可避免地丢失部分原始点云信息。基于上述问题，提出将基于点的三维特征编码网络和基于体素的三维特征编码网络相结合的方案，旨在将细粒度信息融入到三维稀疏卷积中。具体而言，为了避免点云远近稀疏性不同的缺点，采用多尺度分组策略，取不同大小的同心球，得到多个相同中心但规模不同的局部邻域特征后进行拼接，以此捕捉从局部到全局的空间信息。在此过程中，分别进行了四次迭代操作，接着再上采样给每个点赋予高层几何结构信息，通过反向插值以及和之前降采样操作中得到拼接后的特征进行跳跃连接，强化了特征的空间连续性和层级融合。得到最终的点云数据，为后续的目标检测任务提供了充分且丰富的特征表示。总体流程图和多尺度分组策略示意图分别如图 3、图 4 所示。

此外，将原始点云位置信息拼接到对应的点云高层几何结构信息中丰富特征信息，分别利用最大池化和平均池化处理点云高层几何结构信息和原始位置信息，原理如式（1~3）所示

$$F_{\text{geo}} = \text{Max}\left(P_{1\text{pos}}，P_{2\text{pos}}，\cdots P_{\text{npos}}\right) \tag{1}$$

$$F_{pos} = \text{Ave}\left(P_{1pos}，P_{2pos}，\cdots P_{\text{n}pos}\right) \tag{2}$$

$$F_{vox} = F_{geo} \oplus F_{\text{pos}} \tag{3}$$

其中，F_{vox}、F_{geo}、F_{pos} 分别表示体素的整体特征、体素高层的几何结构特征和体素的位置特征。然后将处理后的体素送入基于体素的三维特征编码网络中利用一系列三维稀疏卷积进行特征编码。

3.3 融合转移分块 Transformer 和特征金字塔的二维编码网络

传统的基于体素的三维目标检测方法在特征提

取上难以捕获足够的细节信息，限制了精确目标预测的性能。为充分利用鸟瞰图像的丰富信息，本文引入了一种结合转移分块 Transformer 和特征金字塔网络（FPN-PAN）的高级特征提取策略。首先，将下采样后的鸟瞰图像特征图划分为若干不重叠区域，并在每个区域内独立执行自注意力计算，以捕获局部信息。

通过融合 FPN-PAN 结构，本方法不仅增强了特征金字塔的表达能力，还提升了各尺度目标检测的准确性，丰富了模型的特征提取和增强能力。二维特征编码网络的整体结构如图 5 所示，第二列区域的蓝色部分代表区域 Transformer 块。

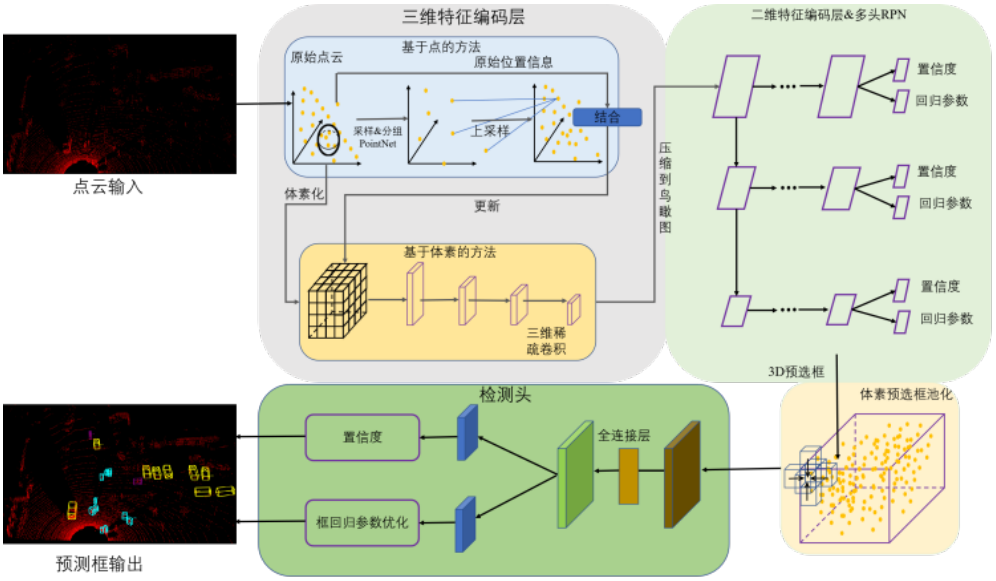

图 2　模型整体流程

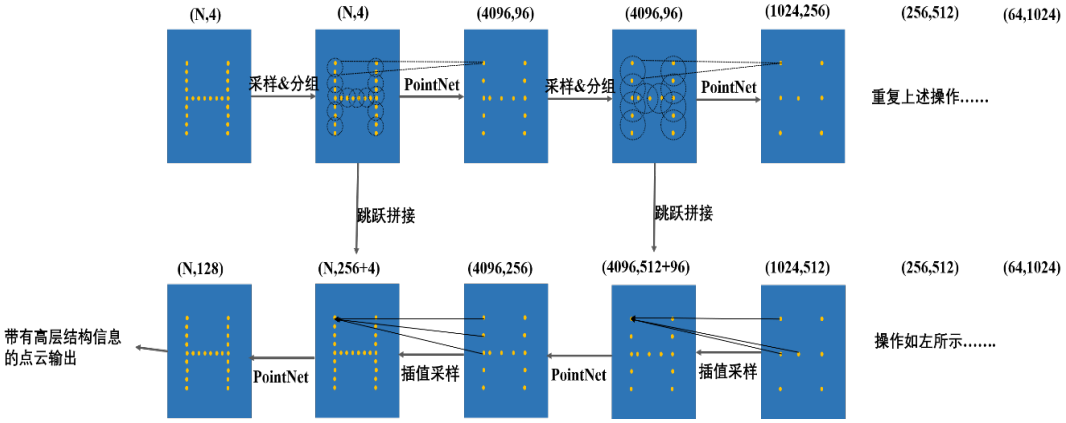

图 3　网络总体流程

连接(contact)

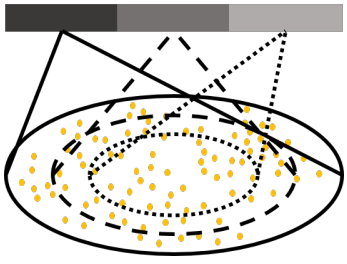

图 4　多尺度分组策略

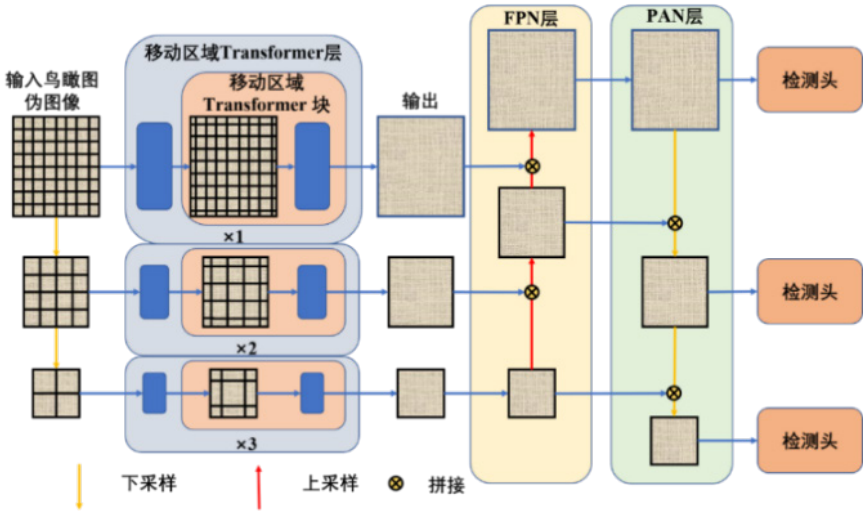

图 5　二维特征编码整体结构示意图

3.3.1 区域和转移分块 Transformer 块

在目标检测任务中，捕获局部关联更为关键。因此借鉴 Swin Transformer[22]的理念，提出区域 Transformer 块，以实现特征的高效变换和信息的全面整合。区域 Transformer 块结构如图 6 所示：

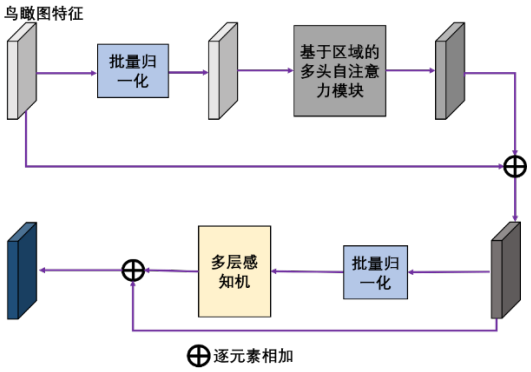

图 6　区域 Transformer 块

具体而言，该方法首先将输入特征图按预定尺寸细分为多个 6×6 的互不重叠小区域，以捕获细节丰富的局部特征。每个区域采用基于区域的多头自注意力机制（MSA）模块，随后是多层感知机（MLP）和 GELU 非线性激活函数，通过自注意力机制深入挖掘区域内部特征关系，并通过 MLP 进一步加工特征。

为了超越局部信息限制并融入更广阔的上下文视角，设计了转移分块 Transformer 块。该块通过将标准区域配置整体向右下方平移，创建新的空间配置，如图 7 所示。这一策略使模型不仅能捕捉每个区域内部的精细信息，还能通过重新计算移动后区域的自注意力，有效整合周围上下文信息。

通过上述设计，转移分块 Transformer 层得以实现，包括区域 Transformer 块和转移分块 Transformer 块的结合使用。三组特征图分别经过 1、2、3 层转移分块 Transformer 层进行特征编码。这不仅强化了模型对不同尺寸特征图的局部信息捕获能力，同时显著提升了全局上下文信息的融合效率。

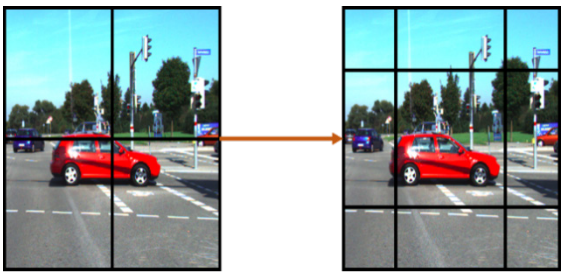

图 7　区域移动示意

3.3.2 相对位置编码

将传统的位置编码替换为了相对位置编码，旨在更精确地捕获图像内部的空间关系，巧妙地解决了传统 Transformer 在处理图像数据时忽视像素之间相对位置信息的问题，这对于理解图像的空间结构和上下文信息至关重要，总体公式如式（4）所示：

$$\mathrm{Attention}(Q,K,V)=\mathrm{Softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}+B\right)V \tag{4}$$

具体而言，本文引入了一个可学习的相对位置偏置，与区域内每对像素的相对位置相关，并将其加入自注意力计算中。这样调整后的自注意力矩阵

能够更准确地反映图像内部元素的空间关系。以 2×2 的区域为例，本方法首先选定自注意力计算中的一个原点（如区域内第一行第一列的元素），作为相对位置编码的参考点（0,0）。在此设置下，其他元素的相对位置索引定义为原点的绝对位置索引减去该元素的绝对位置索引，确保每个元素的位置信息相对于原点进行编码。如图 8 所示。

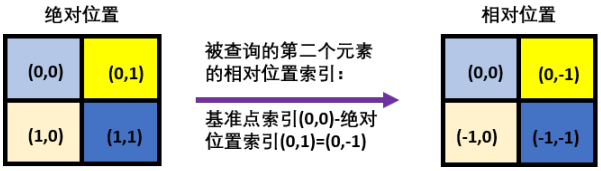

图 8　相对位置计算过程

3.3.3FPN-PAN 特征金字塔模块

为了有效处理多尺度物体检测问题，本文使用 FPN 特征金字塔（Feature Pyramid Networks）来捕获不同尺寸的物体。然而，FPN 结构只关注高层语义特征的传递，限制了低层细节特征向高层的有效传递，尤其在精细化定位和分割应用中。因此，引入 FPN-PAN 结构，通过增加自底向上的信息传递通路，实现特征层之间的双向信息交换。这不仅加强了高层语义特征对低层细节特征的指导，也使低层精细信息能更有效地被高层利用，从而提升了网络对复杂场景的表征能力。

为提升多尺度信息的利用效率，特别是在目标检测任务中对不同尺寸物体的识别能力，本文将增强后的特征图输入到多头区域提议网络。为确保信息一致性和模型可扩展性，所有特征图均统一至固定维度（通道数设为 128）。

这些特征图拼接后得到通道数为 256 的综合特征图。值得强调的是，融合转移分块 Transformer 与 FPN-PAN 的二维特征编码网络是一种即插即用的模块，可灵活应用于各种基于体素的三维目标检测模型中。

3.4 损失函数

模型总损失函数分为第一阶段和第二阶段两个部分。第一阶段为了预测三维物体的目标框，旨在准确预测三维物体的边界框，对三维物体标签以及锚框感兴趣区域建议都进行了如式（5-7）所示的参数化处理。

$$\Delta x = \frac{x^{gt} - x^{a}}{d^{a}}, \Delta y = \frac{y^{gt} - y^{a}}{d^{a}}, \Delta z = \frac{z^{gt} - z^{a}}{h^{a}} \tag{5}$$

$$\Delta w = \log\frac{w^{gt}}{w^{a}}, \Delta l = \log\frac{l^{gt}}{l^{a}}, \Delta h = \log\frac{h^{gt}}{h^{a}} \tag{6}$$

$$\Delta\theta = \sin(\theta^{gt} - \theta^{a}) \tag{7}$$

其中，$\mathrm{da} = \sqrt{(\mathrm{w}^{\mathrm{a}})^{2} + (\mathrm{l}^{\mathrm{a}})^{2}}$，$\mathrm{x}^{\mathrm{gt}}, \mathrm{y}^{\mathrm{gt}}, \mathrm{z}^{\mathrm{gt}}$代表的是真值目标框的坐标，$\mathrm{x}^{\mathrm{a}}, \mathrm{y}^{\mathrm{a}}, \mathrm{z}^{\mathrm{a}}$代表的是锚框的坐标；$\mathrm{l}^{\mathrm{gt}}, \mathrm{l}^{\mathrm{gt}}, \mathrm{l}^{\mathrm{gt}}$代表的是真值目标框的长宽高，$\mathrm{l}^{\mathrm{a}}, \mathrm{w}^{\mathrm{a}}, \mathrm{h}^{\mathrm{a}}$代表的是锚框的长宽高；$\theta^{\mathrm{gt}}$代表的是真值目标框的航向角，$\theta^{\mathrm{a}}$代表的是锚框的航向角。

为了衡量预测边界框与真实目标框之间的差异，选择使用了 SmoothL1 作为回归损失函数，其中目标值用标签数据和区域建议的残差表示。重要的是回归损失只对分配的正样本使用。回归损失函数如式（8）所示：

$$L_{loc} = \sum_{\gamma\in\{x,y,z,l,w,h,\theta\}} L_{\mathrm{smoothL1}}(\Delta\gamma_{p}, \Delta\gamma) \tag{8}$$

其中$\Delta\gamma_{p}$为预测框与预设锚框的残差值。

对于分类损失，由于第一阶段的分类损失是对所有锚框使用的，无论正样本还是负样本，所以会存在类别不平衡的问题。所以使用 Focal Loss[24]作为分类损失函数。该损失函数的核心思想是通过减少易分类样本的权重，从而使模型更加关注于难易分类的样本。具体来说，Focal Loss 函数定义如式（9）所示：

$$L_{cls} = -\alpha(1 - p_{i})^{\beta} log_{2} p_{i} \tag{9}$$

其中，p_{i}是模型对于每个样本分类正确的概率，α是用于平衡正负样本权重的系数，正样本设置为 0.25，负样本设置为 0.75，β是调节易分类样本权重的聚焦参数，设置为 2。

为了提高检测模型在物体朝向预测上的准确性，并防止将方向相反的两个目标误识别为同一物体，采用了交叉熵损失函数来定义方向分类损失如式（10）所示：

$$L_{total}^{1} = \beta_{1} L_{cls} + \beta_{2} L_{loc} + \beta_{3} L_{dir} \tag{10}$$

其中$\beta_{1}, \beta_{2}, \beta_{3}$分别设置为 1.0，2.0 和 0.2。

第二阶段为了进一步细化目标定位并优化模型性能，采用了特定的损失函数配置。在这一阶段中，回归损失函数和定位损失函数同第一阶段，但对残差表示进行了调整，现在残差代表了感兴趣区域建议与真值目标框的。分类损失函数改为了二分类交叉熵损失[25]，这是因为模型在此阶段的目的是区分感兴趣区域提案是属于前景还是背景，而不涉及具

体的类别判断。置信度真值改为了如式（11）所示：

$$L_i\begin{cases}0 & IoU_i<\theta_L\\ \frac{IoU_i-\theta_L}{\theta_H-\theta_L} & \theta_L\le IoU_i<\theta_H\\ 1 & IoU_i\ge\theta_H\end{cases} \tag{11}$$

其中IoU_i是第 i 个方案和对应真值框的 IoU。θ_H,θ_L是前景 IoU 和背景 IoU 的阈值。同第一阶段，分类损失是针对所有感兴趣区域建议的，而回归损失和方向回归损失仅针对那些被判定为前景的感兴趣区域提案。第二阶段的损失函数如式（12）所示：

$$L_{total}^2=\beta_1L_{cls}+\beta_2L_{loc}+\beta_3L_{dir} \tag{12}$$

其中β_1,β_2,β_3分别设置为 1.0,1.0,1.0。总损失为第一二阶段损失相加。

4 试验结果分析

为深入评估提出的模型及其框架在三维目标检测领域的有效性和精确度，本章致力于在广受认可且挑战性十足的 KITTI 数据集上进行细致的性能验证。

4.1 数据集与评价指标

KITTI 数据集（Geiger, Lenz, & Urtasun, 2012）在全球范围内被广泛认可为自动驾驶领域内计算机视觉算法评估的领先数据集之一[26]。本次试验为了促进深度学习模型在实际应用中的性能评估和比较，将 KITTI 数据集的训练样本被分为两部分：一部分包含 3712 个样本，用于模型的训练；另一部分则包含 3769 个样本，用作模型验证。

本文采用了平均精确度（mAP）来量化模型性能，其中针对汽车类别的交并比（IoU）阈值设定为 0.7，而对于行人和骑自行车者类别，则设定为 0.5。

4.2 试验环境及训练参数

开发和验证工作均在 Ubuntu 18.04 操作系统环境下进行，采用 Python 作为主要编程语言。在模型实现过程中，广泛应用了若干第三方库以支持相关的数据处理和模型训练任务。具体如表 1 所示。

表 1　模型所需系统以及开源软件配置表

类别	名称
操作系统	Ubuntu18.04
编程语言	python
数据转换	Numpy
深度学习框架	Pytorch
点云处理	Open3D
模型框架	OpenPCDet

在本研究中，构建的网络框架通过采用 Adam 优化算法进行训练和优化，网络配置为在 80 个训练周期（epochs）内，以每批（batch）12 个样本的大小进行迭代学习。学习率初始值设定为 0.015，且为了优化训练过程，采用了余弦退火策略对学习率进行动态调整。在优化器的配置中，用于控制梯度一阶矩估计和二阶矩估计的衰减率参数和分别设置为 0.9 和 0.999，以保证优化过程的稳定性和效率。

4.3 方法对比分析

在本研究中，我们在 KITTI 数据集上全面评估了所提出的模型框架。我们不仅报告了在 KITTI 验证集上的性能，还将结果提交至在线测试平台，与当前主流和先进方法进行比较。根据评估协议，在验证集上报告了 11 个不同召回率位置下的平均精度（AP），并在 KITTI 测试服务器上使用 40 个召回位置细致评估模型性能。

通过提交结果到 KITTI 在线测试服务器，我们将提出的模型与 KITTI 测试集上的几种主流方法进行比较。如表 2~4 所示，在 KITTI 测试集上对 3 类三维目标框检测（40 个召回位置度量）进行了综合比较。符号“L”和“R”分别代表激光雷达（LIDAR）和图像（RGB）模式。我们的模型在平衡三类目标检测性能方面表现卓越，特别是在汽车和自行车检测上优于大多数现有方法。通过融合点体素表示、移动区域 Transformer、FPN-PAN 特征金字塔以及基于随机点云消除的数据增强方法，模型在汽车、行人和骑行者的中等水平上分别达到了 81.39%、41.68%和 69.03%的平均精度（AP）。

此外，我们在 KITTI 验证集上通过 11 个召回位置计算 AP，以便进行细致比较。如表 5~7 所示，模型在验证集中表现优越，在 9 个项目中有 7 个领先，展现了稳健的性能。与先进的 PV-RCNN 模型相比，本文模型在汽车、行人和骑自行车者上的绝对改进率分别为 0.68%、3.79%和 4.73%。与 PV-CNN 相比，汽车和骑行者的绝对改进率分别为 11.18%和 12.21%。模型在各级别的性能均有显著提升。

为了直观理解，图 9 展示了来自 KITTI3D 数据集的六组数据的可视化结果。每组从上到下依次显示 RGB 图像、本文模型的可视化输出和 PV-RCNN 模型的可视化输出。白色圆圈表示漏检，红色圆圈表示误检。可视化结果清晰展示了本文模型在不同场景下具有更高的召回率和精度，表现出更高的鲁棒性。

表 2　模型在 KITTI 汽车在线测试集的表现（单位%）

方法	模式	简单	中等	困难	mAP	对比
MV3D[31]	L+R	74.97	63.63	54.00	64.20	+18.79
F-PointNet[11]	L+R	82.19	69.79	60.59	70.86	+12.13
AVOD-FPN[27]	L+R	83.07	71.76	65.73	73.52	+9.47
UberATG-MMF[32]	L+R	88.40	77.43	70.22	78.68	+4.31
EPNet[30]	L+R	89.81	79.28	74.59	81.23	+1.76
3D-CVF[29]	L+R	89.20	80.05	73.11	80.79	+2.20
CLOCs[33]	L+R	88.94	80.67	77.15	82.25	+0.74
PointRCNN[12]	L	86.96	75.64	70.70	77.77	+5.22
SA-SSD[34]	L	88.75	79.79	74.16	80.90	+2.09
PV-RCNN[19]	L	90.25	81.43	73.11	82.83	+0.16
Voxel-RCNN[20]	L	90.90	81.62	77.06	83.19	
PDV[35]	L	90.43	81.86	77.36	83.20	
M3DeTR[36]	L	90.28	81.73	76.93	82.97	+0.02
DFAF3D[13]	L	88.59	79.37	72.21	80.06	+2.93
GD-MAE[37]	L	88.14	79.03	73.55	80.24	+2.75
PV-FMRTNet （ours）	L	90.66	81.39	76.93	82.99	

表 3　模型在 KITTI 行人在线测试集的表现（单位%）

方法	模式	简单	中等	困难	mAP	对比
MV3D[31]	L+R	–	–	–	–	
F-PointNet[11]	L+R	50.53	42.15	38.08	43.59	
AVOD-FPN[27]	L+R	50.46	42.27	39.04	43.92	
UberATG-MMF[32]	L+R	–	–	–	–	
EPNet[30]	L+R	–	–	–	–	
3D-CVF[29]	L+R	–	–	–	–	
CLOCs[33]	L+R	–	–	–	–	
PointRCNN[12]	L	47.98	39.37	36.01	41.12	+1.21
SA-SSD[34]	L	–	–	–	–	
PV-RCNN[19]	L	52.17	43.29	40.29	45.25	
Voxel-RCNN[20]	L	–	–	–	–	
PDV[35]	L	47.8	40.56	38.46	42.27	+0.06
M3DeTR[36]	L	45.7	39.94	37.66	41.1	+1.23
DFAF3D[13]	L	47.58	40.99	37.65	42.07	+0.26
GD-MAE[37]	L	–	–	–	–	
PV-FMRTNet （ours）	L	46.32	41.68	38.98	42.33	

表 4　模型在 KITTI 骑行者在线测试集的表现（单位%）

方法	模式	简单	中等	困难	mAP	对比
MV3D[31]	L+R	–	–	–	–	
F-PointNet[11]	L+R	72.27	56.12	49.01	59.13	+12.05
AVOD-FPN[27]	L+R	63.76	50.55	44.93	53.08	+18.10
UberATG-MMF[32]	L+R	–	–	–	–	
EPNet[30]	L+R	–	–	–	–	
3D-CVF[29]	L+R	–	–	–	–	
CLOCs[33]	L+R	–	–	–	–	
PointRCNN[12]	L	74.96	58.82	52.53	62.1	+9.08
SA-SSD[34]	L	–	–	–	–	
PV-RCNN[19]	L	78.6	63.71	57.65	66.65	+4.53
Voxel-RCNN[20]	L	–	–	–	–	
PDV[35]	L	83.04	67.81	60.46	70.43	+0.75
M3DeTR[36]	L	83.83	66.74	59.03	69.86	+1.32
DFAF3D[13]	L	82.09	65.86	59.02	68.99	+2.19
GD-MAE[37]	L	–	–	–	–	
PV-FMRTNet （ours）	L	83.08	69.03	61.47	71.18	

表 5　模型在 KITTI 汽车验证集的表现（单位%）

方法	模式	简单	中等	困难	mAP	对比
F-PointNet[11]	L+R	83.76	70.92	63.65	72.78	+11.52
PV-CNN[21]	L+R	84.02	71.54	63.81	73.12	+11.18
AVOD-FPN[27]	L+R	84.41	74.44	68.65	75.83	+8.67
SIFRNet[28]	L+R	85.62	72.05	64.19	73.95	+10.35
3D-CVF[29]	L+R	89.67	79.88	78.47	82.67	+1.63
SECOND[16]	L	88.61	78.62	77.22	81.48	+2.82
PointPillars[17]	L	86.46	77.28	74.65	79.46	+4.84
PointRCNN [12]	L	88.72	78.61	77.82	81.72	+2.58
PV-RCNN[19]	L	89.03	83.24	78.59	83.62	+0.68
Voxel-RCNN[20]	L	89.23	84.23	78.92	84.13	+0.17
PV-FMRTNet （ours）	L	89.45	84.60	78.84	84.30	

表 6　模型在 KITTI 行人验证集的表现（单位%）

方法	模式	简单	中等	困难	mAP	对比
F-PointNet[11]	L+R	70.00	61.32	53.59	61.64	+0.12
PV-CNN[21]	L+R	73.20	64.71	56.78	64.90	
AVOD-FPN[27]	L+R	–	58.80	–	–	
SIFRNet[28]	L+R	69.35	60.85	52.95	61.05	+0.71
EPNet[30]	L+R	66.74	59.29	54.82	60.28	+1.48
SECOND[16]	L	56.55	52.98	47.73	52.42	+9.34
PointPillars[17]	L	57.75	52.29	47.90	52.65	+9.11
PointRCNN [12]	L	62.72	53.85	50.24	55.60	+6.16
PV-RCNN[19]	L	63.71	57.37	52.84	57.97	+3.79
Voxel-RCNN[20]	L	64.20	58.54	52.93	58.56	+3.20
PV-FMRTNet （ours）	L	66.52	61.93	56.83	61.76	

表 7　模型在 KITTI 骑行者验证集的表现（单位%）

方法	模式	简单	中等	困难	mAP	对比
F-PointNet[11]	L+R	77.15	56.49	53.37	62.34	+15.74
PV-CNN[21]	L+R	81.40	59.97	56.24	65.87	+12.21
AVOD-FPN[27]	L+R	-	49.70	-	-	
SIFRNet[28]	L+R	80.87	60.34	56.69	65.97	+12.11
EPNet[30]	L+R	83.88	65.50	62.70	70.69	+7.39
SECOND[16]	L	80.58	67.15	63.10	70.28	+7.80
PointPillars[17]	L	80.05	62.68	59.70	67.48	+10.60
PointRCNN [12]	L	86.84	71.62	65.59	74.68	+3.46
PV-RCNN[19]	L	86.06	69.48	64.50	73.35	+4.73
Voxel-RCNN[20]	L	84.77	71.96	68.90	75.21	+2.87
PV-FMRTNet （ours）	L	86.63	74.68	72.93	78.08	

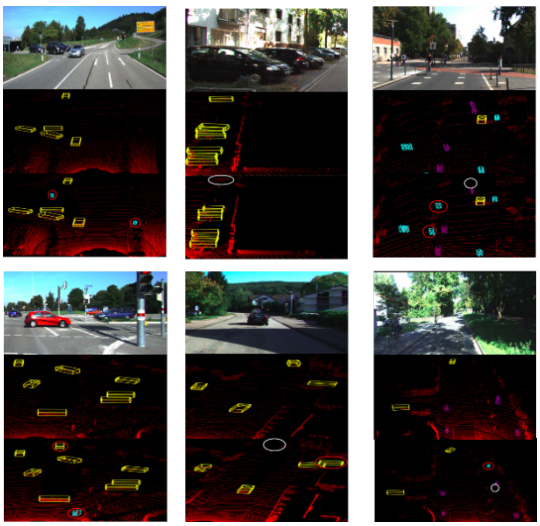

图 9　可视化评估结果对比

4.4 消融试验

为了深入理解提出模型各组成部分的贡献及其对整体性能的影响，我们设计了一系列消融实验。这些实验通过移除或修改特定模块，揭示各组件对最终检测性能的具体作用。实验在 KITTI 数据集上进行，所有模型均在独立的训练集上训练，并在验证集上评估。

作为基线设置，我们选取了 SECOND 算法中的三维和二维特征编码层，以及 Voxel-RCNN 中的感兴趣区域体素特征提取网络（Voxel ROI Pooling）。这一基准模型构成了对比分析的出发点，以量化所提出模型中各模块对整体性能的贡献。通过对比实验结果，可以明确各模块在模型中的作用及其对提升目标检测精度的重要性，从而全面评价所提模型的有效性与改进点。

我们评估了每个模块对 KITTI 数据集上整体模型的贡献，使用 11 个和 40 个召回位置算法，并在中等难度级别上评估三个类别。FGFC 代表基于点-体素融合的细粒度三维特征编码网络，FMRT 代表融合移动区域 Transformer 和 FPN-PAN 特征金字塔的二维特征编码网络，REPC 代表基于随机点云消除的数据增强方法。

实验结果表明（见表 8），数据增强方法显著提升了模型在复杂场景下的识别和定位能力，特别是在物体部分遮挡或截断的情况下。细粒度特征的加入增强了快速编码分支的特征提取能力，对整体性

能提升起到了关键作用。融合移动区域 Transformer 和 FPN-PAN 特征金字塔的方法在上下文信息提取和特征增强方面取得了显著进步，使模型能够更精准地理解和表征三维空间中的目标对象，从而大幅提高了三维目标检测的准确率和效率。

表 8　消融实验结果（单位%）

FG FC	FMRT	REPC	Car		Pedestrian		Cyclist.	
			$AP11_{3D}$	$AP40_{3D}$	$AP11_{3D}$	$AP40_{3D}$	$AP11_{3D}$	$AP40_{3D}$
			81.23	81.76	58.53	59.01	71.66	72.85
√			82.58	82.89	59.79	60.13	72.68	74.29
√	√		84.17	84.49	61.31	61.63	74.16	75.98
√	√	√	84.56	85.24	61.93	62.23	74.68	76.89

5 结语

本文针对点云三维目标检测模型在利用原始信息不足及对目标遮挡和截断鲁棒性低的问题，提出了随机点云消除的数据增强方法、基于点-体素融合的细粒度三维特征编码网络，以及融合转移分块 Transformer 和特征金字塔的二维特征编码网络。实验结果表明，这些改进显著提升了汽车、行人和骑行者的检测准确度。未来研究将继续探索多模态任务的扩展性，特别是结合纯视觉信息和点云数据的鸟瞰图，以进一步提升目标检测精度。

参考文献

[1] Xie G, Zhang X, Gao H, et al. Situational assessments based on uncertainty-risk awareness in complex traffic scenarios[J]. Sustainability, 2017, 9(9): 1582.

[2] 刘秀红,姜圣.基于自动驾驶的车辆安全技术应用探讨[J].时代汽车, 2023(6):175-177.

[3] 李克强,戴一凡,李升波等.智能网联汽车(ICV)技术的发展现状及趋势[J].汽车安全与节能学报,2017,8(01):1-14.

[4] 吕璐,程虎,朱鸿泰等.基于深度学习的目标检测研究与应用综述[J].电子与封装,2022,22(01):72-80.

[5] 郑少武,李巍华,胡坚耀.基于激光点云与图像信息融合的交通环境车辆检测[J].仪器仪表学报,2019,40(12):143-151.

[6] 张银,任国全,程子阳,孔国杰.三维激光雷达在无人车环境感知中的应用研究[J].激光与光电子学进展,2019,56(13):1-11.

[7] 叶语同,李必军,付黎明.智能驾驶中点云目标快速检测与跟踪[J].武汉大学(信息科学版),2019,44(01):139-144+152.

[8] 王刚,王沛.基于深度学习的三维目标检测方法研究[J].计算机应用与软件,2020,37(12):164-168.

[9] Qi C R, Su H, Mo K, et al. Pointnet: Deep learning on point sets for 3d classification and segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 652-660.

[10] Qi C R, Yi L, Su H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[J]. Advances in neural information processing systems, 2017, 30.

[11] Qi C R, Liu W, Wu C, et al. Frustum pointnets for 3d object detection from rgb-d data[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 918-927.

[12] Shi S, Wang X, Li H. Pointrcnn: 3d object proposal generation and detection from point cloud[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 770-779.

[13] Tang Q, Bai X, Guo J, et al. DFAF3D: A dual-feature-aware anchor-free single-stage 3D detector for point clouds[J]. Image and Vision Computing, 2023, 129: 104594.

[14] 周燕,蒲磊,林良熙,et al.激光点云的三维目标检测研究进展[J].计算机科学与探索, 2022, 16(12):23.

[15] Zhou Y, Tuzel O. Voxelnet: End-to-end learning for point cloud based 3d object detection[C]//Proceedings of the IEEE conference on computer vision and pattern

recognition. 2018: 4490-4499.

[16] Yan Y, Mao Y, Li B. Second: Sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.

[17] Lang A H, Vora S, Caesar H, et al. Pointpillars: Fast encoders for object detection from point clouds[C] //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 12697-12705.

[18] Zhou S, Tian Z, Chu X, et al. FastPillars: a deployment-friendly pillar-based 3D detector[J]. arXiv preprint arXiv:2302.02367, 2023.

[19] 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. 2020: 10529-10538.

[20] Deng J, Shi S, Li P, et al. Voxel r-cnn: Towards high performance voxel-based 3d object detection[C]// Proceedings of the AAAI conference on artificial intelligence. 2021, 35(2): 1201-1209.

[21] Liu Z, Tang H, Lin Y, et al. Point-voxel CNN for efficient 3D deep learning[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 965-975.

[22] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[23] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[24] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[25] Shannon C E. A mathematical theory of communication[J]. The Bell system technical journal, 1948, 27(3): 379-423.

[26] 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.

[27] Ku J, Mozifian M, Lee J, et al. Joint 3d proposal generation and object detection from view aggregation[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018: 1-8.

[28] Zhao X, Liu Z, Hu R, et al. 3D object detection using scale invariant and feature reweighting networks[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 9267-9274.

[29] Yoo J H, Kim Y, Kim J, et al. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16. Springer International Publishing, 2020: 720-736.

[30] Huang T, Liu Z, Chen X, et al. Epnet: Enhancing point features with image semantics for 3d object detection[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. Springer International Publishing, 2020: 35-52.

[31] Chen X, Ma H, Wan J, et al. Multi-view 3d object detection network for autonomous driving[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2017: 1907-1915.

[32] Liang M, Yang B, Wang S, et al. Deep continuous fusion for multi-sensor 3d object detection[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 641-656.

[33] Pang S, Morris D, Radha H. CLOCs: Camera-LiDAR object candidates fusion for 3D object detection[C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020: 10386-10393.

[34] He C, Zeng H, Huang J, et al. Structure aware single-stage 3d object detection from point cloud[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11873-11882.

[35] Hu J S K, Kuai T, Waslander S L. Point density-aware voxels for lidar 3d object detection[C]//Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition. 2022: 8469-8478.

[36] Guan T, Wang J, Lan S, et al. M3detr: Multi-representation, multi-scale, mutual-relation 3d object detection with transformers[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022: 772-782.

[37] Yang H, He T, Liu J, et al. GD-MAE: generative decoder for MAE pre-training on lidar point clouds[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 9403-9414.