DOI:10.19651/j.cnki.emt.2518691

基于伪点云融合的多模态三维目标检测方法*

李 旭[1] 张永宏[1,2] 朱灵龙[2] 阚 希[2]
(1.南京信息工程大学自动化院 南京 210044; 2.无锡学院物联网工程学院 无锡 214105)

摘 要: 针对目前的纯激光雷达三维检测方法不可避免地受到点云稀疏性的影响,且激光雷达扫描得到的点云数据在远距离表现比近距离更加稀疏导致模型训练过程中正负样本不均衡的问题,提出一种新的基于伪点云融合的多模态框架 MCA-VoxelNet,它由两个关键设计组成:利用深度补全产生的伪点云来解决点云稀疏性问题,并且通过距离感知采样模块丢弃大量附近的冗余体素来提高计算效率;利用多阶段级联注意力检测结构聚合多个检测阶段的目标特征,平衡正负样本数量并逐步完善 RPN 网络输出的区域建议。在权威的 KITTI 自动驾驶数据集上的实验结果表明,MCA-VoxelNet 以 17.54 的 FPS 在简单、中等和困难三个难度类别上的汽车精度分别达到 94.19%、85.93%和 86.17%,比次优的方法分别高出 2.64%、1.16%和 1.91%。

关键词: 机器视觉;自动驾驶;三维目标检测;伪点云;注意力机制
中图分类号: TN958.98 **文献标识码**: A **国家标准学科分类代码**: 510.4

Multi-modal 3D object detection based on pseudo point cloud fusion

Li Xu[1] Zhang Yonghong[1,2] Zhu Linglong[2] Kan Xi[2]
(1. School of Automation, Nanjing University of Information Science and Technology, Nanjing 210044, China;
2. School of Internet Engineering, Wuxi University, Wuxi 214105, China)

Abstract: To address the inevitable limitations of current LiDAR-only 3D detection methods, which are affected by point cloud sparsity—where LiDAR-scanned point clouds exhibit significantly higher sparsity at long range compared to short range, leading to imbalanced positive and negative samples during model training—we propose a novel multi-modal framework named MCA-VoxelNet, based on pseudo-point-cloud fusion. It consists of two key designs: the pseudo-point clouds generated by depth completion are utilized to solve the problem of point cloud sparsity, and a large number of nearby redundant voxels are discarded through the distance-aware sampling module to enhance computational efficiency; a multi-stage cascaded attention detection structure is employed to aggregate the target features of multiple detection stages, balance the number of positive and negative samples, and gradually improve the region proposals output by the Region Proposal Network. Experiments on the authoritative KITTI autonomous driving dataset demonstrate that MCA-VoxelNet achieves an inference speed of 17.54 FPS and attains car detection accuracies of 94.19%, 85.93%, and 86.17% on the easy, moderate, and hard difficulty levels, respectively. These results outperform the second-best method by 2.64%, 1.16%, and 1.91%.

Keywords: machine vision;autonomous driving;3D object detection;virtual point cloud;attention mechanism

0 引 言

三维目标检测在自动驾驶中具有重要意义[1-2]。激光雷达(light detection and ranging,LiDAR)传感器通过点云的形式获取场景的深度信息,即使在复杂照明条件下依然能实现稳定的目标定位[3-5]。近年来,基于 LiDAR 的三维检测方法受到广泛关注,早期方法中,Chen 等[6]提出将点云投影为鸟瞰图(bird's eye view,BEV)或深度图,用于目标检测。随后,基于点的方法逐渐成为主流。SA-SSD[7]通过一个辅助网络,将骨干网络中的卷积特征还原为点级表示,利用点云的结构信息提升单阶段探测器的定位精度。PointPillars[8]采用 PointNet[9-10]提取柱状体中点云的特征,用于后续的目标检测。PointRCNN[11]提出了经典的两阶段检测框架:第一阶段自下而上生成三维建议,第二阶段在

收稿日期:2025-04-26
* 基金项目:国家自然科学基金(42175157,42475151,42305158)、无锡市"太湖之光"科技攻关计划(基础研究)项目(K20231021)资助

规范坐标系中对建议进行精细回归，获得最终检测结果。3DSSD[12]则通过融合采样策略，进一步降低计算开销并提升效率。同时，基于体素的稀疏卷积方法也逐渐兴起。Voxel-R-CNN[13]结合两阶段框架与体素特征，兼顾精度与效率；SECOND[14]引入稀疏卷积加快推理速度；而 PV-RCNN[15]综合利用体素卷积和 PointNet 的灵活感受野来获得更多的目标判别特征。

尽管基于 LiDAR 的三维目标检测技术取得了显著进展[16-18]，但对于远距离目标上的检测性能仍存在明显不足，主要原因在于 LiDAR 在远距离场景中采样密度较低，导致点云数据过于稀疏[19-20]。相比之下，彩色图像传感器具备高分辨率和丰富的语义信息[21]，近年来，利用二维图像生成伪点云以补充稀疏点云成为研究热点，此类方法通过在原始 LiDAR 点周围生成虚拟点，从而增强点云密度和目标几何信息。例如，MVP[22]提出了一种将 RGB 图像无缝融合到基于 LiDAR 的三维识别流程中的方法，利用 2D 图像生成密集的三维虚拟点，以增强原始稀疏点云的表达能力；SFD[23]基于深度补全网络构建虚拟点，有效重建远处物体的几何结构，显著提升了三维检测性能。

然而，由图像生成的虚拟点云在提升点云密度的同时，也引入了过度稠密的问题。例如，在 KITTI[24]数据集中，一张分辨率为 1 242×375 的图像可生成约 40 万～50 万个虚拟点，造成了巨大的计算开销，影响模型的训练与推理效率。为此，部分工作尝试使用下采样策略进行密度控制，如 RandLA-Net[25]采用高效的随机抽样方法显著减少点云数量，但由于 LiDAR 在近距离能够获取较为完整的物体形状，而远距离点云则更为稀疏，盲目地对虚拟点进行下采样可能会舍弃关键的远距离几何信息，进而影响检测精度。因此，如何在提升点云密度的同时兼顾信息保留与计算效率，仍是一个亟待解决的问题。

针对上述问题，本文提出了一种基于伪点云融合的多模态三维目标检测框架（multi-stage cascaded attention-VoxelNet，MCA-VoxelNet）。该框架利用深度补全生成的伪点云，有效缓解点云稀疏问题。同时，设计距离感知采样模块，用于剔除近距离的冗余体素，从而提升计算效率。同时，设计多阶段级联注意力检测结构，用于融合不同检测阶段的特征表示，逐步完善区域候选，并缓解正负样本不均衡问题。工作的主要贡献如下：

1）基于深度补全网络构建伪点云，并与原始点云融合生成多模态点云；

2）设计了一种距离感知采样模块，用于调控点云密度。在激光雷达扫描中，近距离物体的几何形状通常较为完整，近距离生成的大量虚拟点对性能提升有限，却显著增加计算开销。为此，该模块优先采样关键的远距离虚拟点，舍弃大量近距离点云，从而显著提高网络计算效率；

3）为解决远近距离正负样本不均衡问题，设计多阶段级联注意力检测网络，在检测过程中增强远距离目标的响应，缓解正负样本分布不均，确保后续阶段能够获得充足特征，有效恢复被忽略的远处目标。

1　算法框架

基于 Voxel-RCNN 网络框架，在数据输入层，首先通过深度补全网络处理 RGB 图像，利用图像语义信息生成的虚拟点云数据；然后将虚拟点云与激光雷达采集的原始点云进行空间对齐，构建多模态点云数据。预处理阶段，数据经过距离感知采样策略进行降采样，并通过体素特征编码层将不规则点云转化为规则化的体素表示。特征提取阶段，采用 3D 稀疏卷积主干网络对点云进行三维特征提取，网络最后一层在 BEV 视角下沿高度维度将三维特征压缩为二维特征图。二维特征提取网络对特征图进行不同尺度的下采样后送入区域建议网络（region proposal network，RPN），该网络通过预设的 3D 锚框生成初步的三维目标预测框和类别概率分布。最后，输入多阶段级联注意力检测结构对建议框进行多阶段级联细化，输出最终的物体类别信息、三维尺寸和检测置信度。图 1 展示了该方法的总体框架。

2　基于伪点云融合的多模态三维目标检测

2.1　伪点云生成

深度补全网络旨在在彩色图像的辅助下，从稀疏的深度图中预测出稠密的深度图。该任务主要服务于下游应用，如语义分割、目标检测和三维重建等。然而，目前在三维目标检测领域中，深度补全方法的应用仍较为有限。最近，在基于图像的三维物体检测中，有一些工作[26-27]使用深度补全网络来生成伪点云。图 2 展示了深度补全网络结构图，首先基于图像的深度估计开始，将一对水平偏移的摄像机拍摄的左右图像 $\boldsymbol{I}_l$ 和 $\boldsymbol{I}_r$ 作为输入，假设该算法将左图像 $\boldsymbol{I}_l$ 视为参考，并输出记录每个像素 (u,v) 的水平视差到右图像 $\boldsymbol{I}_r$ 的视差图 $\boldsymbol{D}$，理想情况下，$\boldsymbol{I}_l(u,v)$ 和 $\boldsymbol{I}_r(u,v+\boldsymbol{D}(u,v))$ 将描绘相同的 3D 位置。因此，可以通过下式导出深度图 $\boldsymbol{Z}$：

$$\boldsymbol{Z}(u,v)=\frac{\boldsymbol{f}_U\times b}{\boldsymbol{D}(u,v)} \tag{1}$$

式中：$\boldsymbol{Z}(u,v)$ 是深度图中像素 (u,v) 的深度值，f_U 是左相机的水平焦距，b 是左右相机之间的基线距离，$\boldsymbol{D}(u,v)$ 是视差图中像素 (u,v) 的视差值。

将所得到的深度图 $\boldsymbol{Z}$ 反投影到 3D 点云中，通过下式将像素 (u,v) 变换为 3D 中的 (x,y,z)：

$$z=\boldsymbol{Z}(u,v) \tag{2}$$

$$x=\frac{(u-\boldsymbol{c}_u)\times z}{\boldsymbol{f}_U} \tag{3}$$

$$y=\frac{(u-\boldsymbol{c}_u)\times z}{\boldsymbol{f}_V} \tag{4}$$

式中：(c_u,c_v) 是相机中心的像素位置，f_V 是垂直焦距。

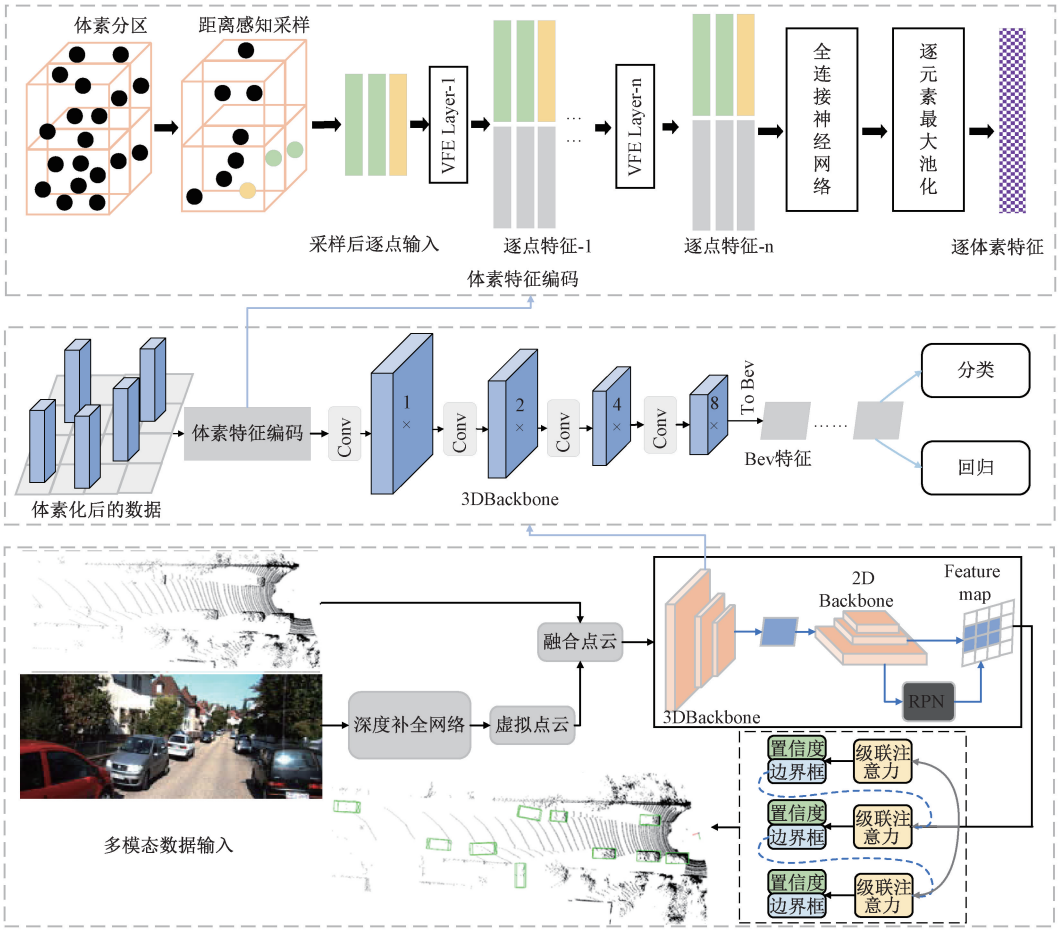


图 1　总体框架图

Fig. 1　Overall framework diagram

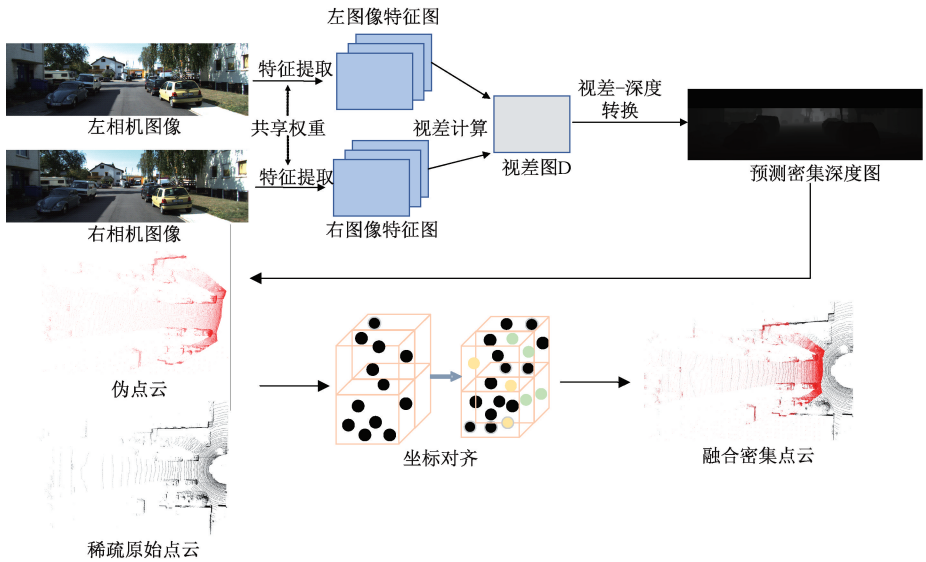


图 2　深度补全网络结构图

Fig. 2　Depth completion network structure diagram

通过将所有像素反向投影到 3D 坐标中,得到 3D 点云 $\{x^{(n)},y^{(n)},z^{(n)},r\}_{n=1}^{N}$,其中 N 是像素数,r 为反射强度。这样的点云可以被变换到给定参考视点和观察方向的任何坐标系中,将得到的点云称为伪点云。由于原始点云和

伪点云可能来自不同的坐标系，需要用以下变换将它们对齐到同一个坐标系中：

$$\boldsymbol{P}'_v(x',y',z',r') = \boldsymbol{R} \cdot \boldsymbol{P}_v(x,y,z,r) + \boldsymbol{t} \tag{5}$$

式中：$\boldsymbol{P}_v$ 是虚拟点云中的点，$\boldsymbol{P}'_v$ 是对齐后的点，$\boldsymbol{R}$ 为旋转矩阵，$\boldsymbol{t}$ 为平移向量。

将原始点云 $\boldsymbol{P}_o$ 和伪点云 $\boldsymbol{P}_v$ 合并为融合点云 $\boldsymbol{P}_{fus}$：

$$\boldsymbol{P}_{fus} = \{(\boldsymbol{P}_o, r_o), (\boldsymbol{P}'_v, r_v)\} \tag{6}$$

图 3 展示了融合点云的可视化效果，其中黑色点云代表原始点云，红色点云则是由 RGB 图像生成的虚拟点云。可以观察到，在近距离范围内，原始点云和虚拟点云较为密集，冗余信息可能干扰物体特征的提取；而在远距离区域，虚拟点云有效弥补了原始点云的稀疏性，使远处目标（如车辆）的几何特征更加完整，从而有助于提升目标检测的精度。

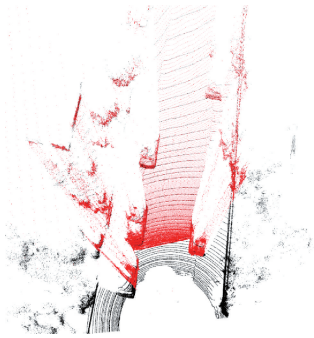


图 3　融合点云

Fig. 3　Fused point cloud

2.2　距离感知采样模块

为缓解虚拟点带来的计算负担并提升检测的鲁棒性，引入了一种距离感知采样模块，它通过在训练和推理过程中丢弃输入的虚拟点来加快网络的速度。现在有两种常见的采样方法可以减少点云数据的输入量：随机采样和最远点采样。但是随机采样是在整体体素数据上随机抽取一定量的数据来减少数据的输入进而降低计算成本，它并不关注数据的重要程度。基于对点云数据的观察，激光雷达扫描的物体往往在近距离的特征表现的比较完整，而在远处物体特征会变的稀疏，因此随机采样会使不同距离的特征保持不平衡，会不可避免地牺牲一些有用的形状线索。最远点采样的核心思路是致力于让各个采样点之间的距离达到尽可能远的程度，从本质上来说，也就是促使数据能够尽可能地实现离散且均匀分布，具体实施办法是：假设输入点云集合为 P，其中包含 n 个点，首先从中随机选取一个初始点 P_0，构成采样点集合 $S = \{P_0\}$。然后，计算集合内所有点到采样点的最小欧式距离，选取距离最远的点作为下一个采样点，加入 S。重复此过程，每次更新每个点到当前采样点集合 S 的最小距离，并选取最大值对应的点加入集合 P，直到采样得到一定数量的点为止。在处理大量点云时，由于频繁计算欧式距离，最远点采样计算开销较大。

为了解决这个问题，引入了一种距离感知采样模块来执行高效和平衡的采样，该模块主要包括两个步骤：首先，根据点云中点到传感器的距离，将输入点云划分为若干个距离区间（共 10 个），并设定最大采样距离为 60 m，以确定各区间的边界。其中，距离小于 30 m 的区域被定义为近距离区间。在这些近距离区间内，点云密度较高，为避免过多冗余信息，采用随机下采样的方式，控制每个体素内的点数在 0～1 000，通过设定整体的丢弃率，只需很小的计算成本就能得出需要丢弃的体素数量。对于距离超过 30 m 的远距离区间，由于点云数据较为稀疏，保留所有体素信息，以确保远距离目标的检测精度。

2.3　用于建议优化的级联注意力网络

级联检测网络在二维图像目标检测任务中取得了显著的效果。Cai 等[28]提出的 Cascade-R-CNN 采用了一种多阶段的级联目标检测结构，在检测头部分采用了一系列经过增加交并比（intersection over union，IoU）阈值训练的检测器来优化感兴趣区域建议。Cascade R-CNN 包含 N 个级联细化模块，第 j 个细化模块将前一阶段的区域建议 $\boldsymbol{B}^{j-1}$ 作为输入，并使用特征提取器提取目标特征 $\boldsymbol{F}^j$，目标特征 $\boldsymbol{F}^j$ 再通过一个置信度预测分支和一个边界框回归分支分别输出一个新的目标 $\boldsymbol{C}^j$ 和边界框 $\boldsymbol{B}^j$，这个迭代细化过程可以简化为：

$$\boldsymbol{F}^j = \boldsymbol{\varphi}^j(\boldsymbol{B}^{j-1}) \tag{7}$$

$$\boldsymbol{C}^j = \boldsymbol{S}^j(\boldsymbol{F}^j) \tag{8}$$

$$\boldsymbol{B}^j = \boldsymbol{R}^j(\boldsymbol{F}^j) \tag{9}$$

式中：$j = 1, 2, \cdots, N$。

然而，普通的级联结构没法在 3D 目标检测器上取得性能优化，原因为：忽略了远处的物体特征，在多阶段方法中，由于缺乏负训练样本，在后期训练阶段会存在训练过度拟合的情况，在二维目标检测方法中，通过级联结构来建立逐步提高 IoU 阈值的检测头，用上一个阶段训练好的结果作为输入输送到下一个检测阶段来重新采样平衡训练样本。然而，在 3D 点云数据中，由于点云通常是非均匀分布的，近距离和远距离物体之间样本数量不平衡，在近距离处具有密集点的物体可以产生高质量的区域建议，被选为正样本，而远处的物体由于更加稀疏往往被判定为负样本。

在这种不平衡的训练下，多阶段的训练方法可以准确预测 LiDAR 附近的物体，而忽略远处的物体。为了解决该问题，引入了一种多阶段级联注意力检测头，通过在前期阶段增加更多目标的出现，确保后续阶段能够获得足够的物体特征，从而有效恢复被忽略的远处目标。简而言之，需要在每个检测阶段之间建立有效的连接，引用跨阶段聚合特征的方式来增加目标的特征，从而能够更加准确，更容易地检测远处和近处难以识别的物体。

现有级联结构只关注当前阶段的候选特征，忽略了前序阶段所提供的潜在信息，导致多阶段特征未能充分利

用,为解决这一问题,需要进一步对各阶段的特征信息进行融合,一种简单的实现方式是将各阶段输出的特征直接拼接,但是不同检测阶段输出得到的特征重要性很难学习,会导致性能提升有限。鉴于近年来注意力机制在特征选择方面的有效性,本文设计了一种基于多阶段级联注意力的特征聚合策略,能够自适应地捕捉各阶段特征的重要性,从而实现更有效的信息整合与目标检测性能的提升。结构如图 4 所示。

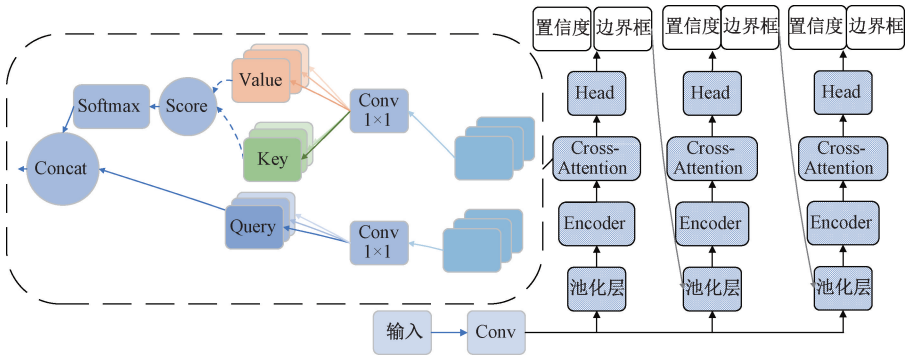


图 4　多阶段级联注意力检测结构

Fig. 4　Multi-stage cascaded attention detection structure

对于每个由 RPN 网络输出的编码特征 $\boldsymbol{F}_j$,首先通过卷积核对特征进行处理并映射到高维空间,对于第 j 个检测阶段,收集之前所有阶段和当前阶段构成特征 $\boldsymbol{F}_j=(\boldsymbol{F}_0,\boldsymbol{F}_1\cdots)$,然后根据式(10)~(13)分别计算查询矩阵向量 $\boldsymbol{Q}^j$,键向量矩阵 $\boldsymbol{K}^j$,值向量矩阵 $\boldsymbol{V}^j$,和注意力权重矩阵 $\boldsymbol{H}^j$。

$$\boldsymbol{Q}^j=\boldsymbol{F}_j\boldsymbol{W}_q^j \tag{10}$$

$$\boldsymbol{K}^j=\boldsymbol{F}_j\boldsymbol{W}_k^j \tag{11}$$

$$\boldsymbol{V}^j=\boldsymbol{F}_j\boldsymbol{W}_v^j \tag{12}$$

$$\boldsymbol{H}^j=softmax\left(\frac{\boldsymbol{Q}^j(\boldsymbol{K}^j)^{\mathrm{T}}}{\sqrt{c'}}\right)\boldsymbol{V}^j \tag{13}$$

式中:$\boldsymbol{C}'$ 为多头注意力的特征维度。

通过执行交叉注意力的操作,聚合来自不同阶段的特征。通过采用这种级联关注设计,模型可以更好地估计建议在各个阶段的质量,有助于提高建议的细化精度。表 1 为多头注意力参数表。

表 1　多头注意力参数

Table 1　Multi-head attention parameters

参数名	数值
注意力头数	4
特征维度	256

对于预测边界框的回归,遵循 Voxel-RCNN 的方法,对边界框尺寸、位置以及方向残差进行回归。在训练阶段,与前文提到的级联检测网络类似,设置三维 IoU 阈值以此来定义不同检测阶段中的负样本和正样本。其中,在第一个阶段,如果一个建议框和地面真值框具有至少 0.5 个 3D IoU,则该建议框被视为建议框细化分支的正样本,否则将被视为负样本,随后每个阶段的 IoU 阈值增加 0.1。在测试阶段,对所有检测阶段的边界框和分数取平均值,从而生成最终的检测结果。

2.4　特征提取主干网络

最近的许多网络使用三维稀疏卷积作为主干网络,以提高准确性和效率,MCA-VoxelNet 也采用这种设置。三维稀疏卷积主干提取特征的步骤如下:首先将原始点分割成小体素,对于每个体素,使用所有内部点的原始特征的平均值来计算原始特征,采用多个三维稀疏卷积块将三维点云编码特征向量,卷积块由一系列 3×3×3 三维稀疏卷积核组成,这些卷积核将空间特征下采样到 1×、2×、4×,最终下采样到 8×的张量。最后一层的三维特征沿着 Z 轴维度被压缩为 BEV 特征,用于感兴趣区域目标建议生成。

2.5　区域建议网络 RPN

区域建议网络 RPN 通过在 2D 骨干网络输出的二维图像的卷积特征图上应用一系列卷积窗口来生成物体的候选区域。具体过程是:首先通过 3 个卷积层提取图像的高层次特征图,然后在该特征图上使用一个小的卷积核滑动,针对每个位置生成多个锚点,每个锚点会被分类为前景或背景,并通过回归预测调整该锚点的位置、大小和方向,以尽量接近真实物体的边界框,RPN 通过基于 IoU 的匹配将真实边界框分配给锚点,对于第 i 个锚点,分别用 α_i、α_i'、δ_i、δ_i' 来表示预测的得分、目标实际得分、预测的残差和目标实际残差。区域建议网络 RPN 的损失定义为:

$$L_{\mathrm{RPN}}=\sum_i L_{\mathrm{score}}(\alpha_i,\alpha_i')+I(IoU_i>u)\sum L_{\mathrm{reg}}(\delta_i,\delta_i') \tag{14}$$

$$L_{\mathrm{score}}=\begin{cases}\dfrac{1}{2}x^2, & |x|<1\\ |x|-\dfrac{1}{2}, & \text{其他}\end{cases} \tag{15}$$

式中:$\mathrm{I}(IoU_i>u)$ 表示只有当检测物体的目标建议 $IoU_i>u$ 才会产生回归损失,$x=\alpha_i-\alpha_i'$。

在 KITTI 数据集中,IoU 阈值 u 设置为 0.5 和 0.7 两

个档位。L_{score} 和 L_{reg} 分别表示平滑 L1 损失和边界框回归二元交叉熵损失。

3　损失函数

网络的总体损失由区域建议网络 RPN 的损失 L_{RPN} 和多阶段级联注意力检测头的损失 L_{MCA} 组成，总体损失函数可以表示为：

$$L_{Total} = L_{RPN} + L_{MCA} \tag{16}$$

RPN 的损失 L_{RPN} 如上节所述，多阶段级联注意力检测头的损失 L_{MCA} 是多个阶段中的多个细化损失的总和，在每一个建议细化阶段，L_{MCA} 采用边界框回归损失 L_{reg} 和加权得分损失 L_{score}，对于第 j 个细化阶段的第 i 个区域建议，多阶段级联注意力检测头的损失 L_{MCA} 可以表示为：

$$L_{MCA} = \sum_i \sum_j [L_{MCA}(\alpha_i, \alpha_i') + I(IoU_i^j > u_i^{j'}) L_{reg}(\delta_i, \delta_i')] \tag{17}$$

式中：α_i，α_i'，δ_i，δ_i' 分别表示预测分数、目标实际分数、预测残差和目标实际残差。

4　实验及分析

4.1　数据集及评估指标

为实现与其他方法进行公平对比，选用在自动驾驶领域被公认为极具权威性的数据集——KITTI 数据集来开展相应实验。KITTI 数据集是自动驾驶三维目标检测最流行的数据集之一，包含了道路场景的激光雷达点云和配套的图片数据，其中有 7 481 个训练样本和 7 518 个测试样本，并细分为简单、中等和困难 3 个层级。这 3 个等级的依据图片中二维包围框的像素高度、遮挡程度和截断比例 3 个指标进行划分，表 2 展示了 3 个难度层级的划分依据。本文主要使用 KITTI 的两个指标进行检测结果的评估：三维目标检测性能（3D mean average precision，3D mAP，3DIoU＝0.7）和鸟瞰图检测性能（BEV mean average precision，BEV mAP，2DIoU＝0.7）。

表 2　KITTI 数据集中难度层级的划分依据

Table 2　Basis for difficulty levels in KITTI dataset

指标	困难	中等	简单
样本数量	28 742	18 971	12 611
像素高度	＞25	＞25	＞40
遮挡程度	难以察觉	部分可见	全部可见
最大截断	＜50％	＜30％	＜15％

由于汽车类目标在自动驾驶场景中种类最多、数量最多，且检测结果较为稳定，因此其检测性能能够较好地反映网络的整体表现。基于此，主要以汽车作为实验评估的目标。按照 KITTI 数据集的官方要求，主要使用 40 个召回位置的平均精度对检测结果进行评估，该指标的 IoU 阈值分别为汽车、行人和骑自行车者的 0.7、0.5 和 0.5。平均精准度为对 Precision-Recall 曲线上的 Precision 值求均值，定义如下：

$$\boldsymbol{AP}_R = \frac{1}{R} \sum\nolimits_{r \in R} \boldsymbol{P}_{interp}(r) \tag{18}$$

$$\boldsymbol{P}_{interp}(r) = \max_{\tilde{r} > r} \boldsymbol{p}(\boldsymbol{r}) \tag{19}$$

式中：AP 为平均精度，$R = \left\{0, \frac{1}{40}, \frac{2}{40}, \cdots, 1\right\}$，表示 40 点插值，$\boldsymbol{P}_{interp}(\boldsymbol{r})$ 为插值函数，r 为设定召回值，$\tilde{r}$ 为实际召回值。

4.2　参数设置

区域建议网络 RPN 采用锚框机制，表 3 为三类检测类别的锚框尺寸。

表 3　锚框尺寸

Table 3　Anchor frame size

类别	锚框尺寸		
	长	宽	高
汽车	3.92	1.62	1.58
行人	0.81	0.59	1.75
骑行者	1.78	0.62	1.71

所有的传感器都使用相同的标准检测范围，KITTI 数据集坐标系方向为传感器朝向为 x 轴正方向，x 轴向左为 y 轴正方向，向上为 z 轴正方向，x 轴上的范围为[0，70.4]m；y 轴上的范围为[－40，40]m；z 轴上的范围为[－3，1]m，体素分辨率为[0.05，0.05，0.05]m。还使用了常规的数据增强方法如地面真值采样，局部旋转平移，全局旋转尺度变换增强，随机旋转角度和尺度缩放比例，均服从正态分布，角度分布区间为[－0.78°，0.78°]，尺度缩放比例区间为[0.9，1.1]。

在训练过程中，使用自适应矩估计（adaptive moment estimation，Adam）优化器和余弦退火策略调整学习率；Batchsize 为 4；初始学习率为 0.001 25；训练周期为 60；学习率衰减系数为 0.01。实验设备使用一台配备 RTX4090GPU 和 Intel 14900KF CPU 的计算机。

4.3　定量结果与分析

为确保与其他方法进行公平对比，首先在相同的实验环境中复现了几种最先进的三维目标检测方法，其中单模态方法有：PV-RCNN、Voxel-RCNN 和 SE-SSD；激光点云和图像融合的多模态方法有：SFD、VPFNet 和 TED-M。所有比较方法的源代码都是公开的，所有参数都按照原文建议配置，且都是基于 OpenPCDet 工具箱实现。

表 4 展示了本方法与其他先进方法在 KITTI 测试集上的对比实验结果，最优结果用加粗字体表示。其中，“LiDAR”代表纯点云三维目标检测方法，而“LiDAR＋RGB”代表融合点云和图像的多模态三维目标检测方法。如表 4 所示，引入额外的图像数据使得多模态三维目标检测方法

在整体精度上优于纯激光雷达点云三维目标检测方法。本文方法在纯点云和多模态三维目标检测方法中均实现了最高的三维检测精度,以 Car 3D AP(R40)指标为例,在 KITTI 测试集的简单、中等和困难难度下,本方法分别比次优的方法高出 2.64%、1.16%和 1.91%,展示了其在汽车三维目标检测中的卓越表现。特别需要关注的是,最近提出的 SFD 方法,也是采用在稀疏的点云上生成额外的 RGB 虚拟点以补充物体特征来提升 3D 目标检测精度,本方法对比其的 Car 3D AP (R40)在简单、中等和困难难度上分别高出 2.46%、1.16%和 8.25%,正是距离感知体素丢弃模块在训练期间丢弃近距离冗余的体素不仅加快了网络的速度,同时通过模拟更稀疏的训练样本,提高了检测的鲁棒性。

表 4　KITTI 测试集上的对比实验结果

Table 4　The results of comparative experiments on the KITTI test set　%

算法	模态	Car 3D AP(R40)			Car BEV AP (R40)		
		简单	中等	困难	简单	中等	困难
PV-RCNN	LiDAR	90.25	81.43	76.82	94.98	86.14	90.65
Voxel-RCNN	LiDAR	90.90	81.62	77.06	94.85	88.83	86.13
SE-SSD	LiDAR	91.49	82.54	77.15	**95.68**	91.84	86.72
SFD	LiDAR+RGB	91.73	84.76	77.92	95.64	91.85	86.83
VPFNet[29]	LiDAR+RGB	91.02	83.21	78.20	93.94	90.52	86.25
TED-M[30]	LiDAR+RGB	91.55	84.48	84.26	95.42	**91.93**	88.11
Ours	LiDAR+RGB	**94.19**	**85.92**	**86.17**	95.41	91.72	**91.92**

图 5 展示了 MCA-VoxelNet 与基线模型 Voxel-RCNN 在车辆类别上的 3D 目标检测可视化对比结果,第 1 列为 RGB 图像,第 2 与第 3 列为本文模型与基线模型的点云检测可视化对比图。从图中可以看出,本文模型在不同距离范围内均展现出出色的检测能力,尤其在远距离场景下,能够成功识别出更多低密度点云区域中的车辆目标(红圈中为基线模型漏检的车辆)。相比之下,基线模型在同一场景下未能检测到部分远距离车辆,说明所采用的点云融合策略有效缓解了原始点云在远距离下稀疏带来的信息缺失问题。整体来看,所提方法在保证检测精度的同时,进一步增强了对复杂交通场景中目标的感知能力。

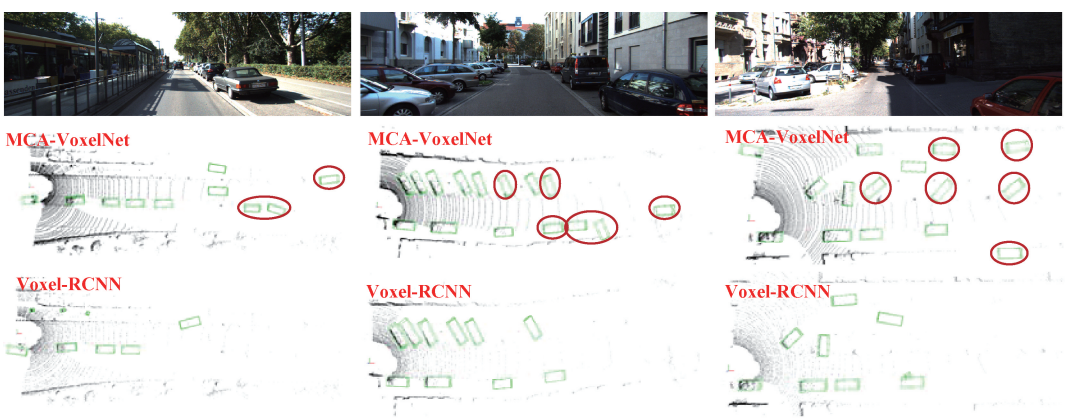


图 5　可视化结果对比

Fig. 5　Visualization result comparison

4.4　消融实验结果

在 KITTI 验证集上的消融实验结果如表 5 所示,表内结果为三次重复实验的均值+标准差。首先使用融和虚拟点云后的数据在 Voxel-RCNN 的框架上经过 60 轮训练后在验证集上进行推理,三个难度类别的精度分别提高了 2.38%、3.69%、5.54%;随后添加了距离感知采样方法,每秒帧数(frames per second,FPS)的提升对比未使用采样达到 3.25;单独添加了 MCA 方法的模型对比 Voxel-RCNN 三个难度类别分别提升了 2.15%、4.67%、4.78%,性能的提高主要来自多阶段级联注意力检测头的设计,聚合了来自多个检测阶段的建议特征,从而实现更有效和更全面的检测物体特征细化;最终,添加了多模态数据、体素采样方法和 MCA 的模型分别在简单、中等和困难难度汽车类别上分别比基线模型 Voxel-RCNN 高出 3.30%、4.46%和 8.92%。

进行一系列消融实验来验证提出的 MCA,首先使用

表 5　消融实验结果

Table 5　Results of ablation experiments

虚拟点	距离感知采样	MCA	Car 3D AP(R40)/%			FPS
			简单	中等	困难	
×	×	×	90.83±0.14	81.55±0.17	77.13±0.20	25.12±0.11
√	×	×	93.21±0.12	85.24±0.13	82.67±0.15	21.87±0.09
√	√	×	93.11±0.11	85.37±0.15	82.65±0.18	24.05±0.13
×	×	√	92.98±0.10	**86.22**±0.14	81.91±0.16	18.70±0.08
√	√	√	**94.13**±0.14	86.01±0.09	**86.05**±0.16	**17.42**±0.10

Voxel-RCNN 作为基线模型，构建了一个基于自注意力的检测头模型，比较结果如表 6 所示，可以看到简单的注意力结构并不能带来性能提升，随后，通过加入的多阶段级联注意力检测头，检测性能在简单、中等和困难难度上分别提升到了 93.85%、86.83%和 86.13%，这是因为 MCA 融合了来自不同检测阶段的目标特征，从而在点云稀疏的场景中实现了更具鲁棒性的检测效果。

表 6　不同注意力方法对比结果

Table 6　Comparison results of different attention methods

%

方法	Car 3D AP(R40)			
	简单	中等	困难	平均
自注意力机制	93.03	86.31	81.85	87.06
多阶段级联注意力	**93.85**	**86.83**	**86.13**	**89.94**

为了研究多阶段级联注意力检测结构中的级联阶段数量对本文所提检测方法性能的影响，在 KITTI 验证集进行了相关实验，结果如表 7 所示，MCA-VoxelNet 使用 3 个级联阶段在中等和困难的 Car 3D AP 中达到最佳结果，使用 5 个级联阶段在简单难度中达到最佳结果，可以观察到使用 3、4、5 三个阶段的检测性能彼此接近，但运行时间却逐步上升，因此采用 3 个级联阶段可以达到最佳的精度效率权衡。

表 7　级联细化阶段数量的消融实验结果

Table 7　Ablation experiment results using different numbers of cascade refinement stages

级联阶段数	Car 3D AP(R40)/%			运行时间/ms
	简单	中等	困难	
1	93.28	85.14	84.97	**60.45**
2	93.79	85.60	85.52	68.92
3	94.10	**85.93**	**86.06**	72.89
4	93.94	85.74	85.89	88.61
5	**94.19**	85.88	85.06	113.27

随后进行实验证明距离感知采样模块的先进性，使用 Voxel-RCNN 作为基线模型，在添加了融和数据后，在输入数据时分别采用随机均匀降采样和距离感知采样模块，下采样率都设置为 80%，比较结果如表 8 所示，在困难难度中，随机均匀降采样因为在采样过程中丢失了大量目标特征而检测精度下滑了高达 25.65%，主要是因为没有关注远处体素和近距离体素的重要程度而丢失了更多有用特征但保留更多无用特征而带来精度的下降，而距离感知体素丢弃模块在训练期间丢弃近距离冗余的体素不仅加快了网络的速度，还有效降低了融和点云的冗余问题带来的精度下降。

表 8　不同采样方法对检测性能的影响

Table 8　Influence of different sampling methods on detection performance

%

方法	Car 3D AP(R40)			
	简单	中等	困难	平均
随机均匀采样	80.95	62.47	57.87	67.10
距离感知采样	**92.99**	**85.87**	**83.52**	**87.46**

图 6 为输入体素采样率梯度对检测精度和 FPS 的影响曲线，采样率梯度差为 2%，观察图 6 可以得出，随着采样率逐渐上升，检测精度变化较为平缓，保持稳定状态，但当采样率超过 80%后，精度会突然下降，表现出明显的性能退化。与此同时，FPS 随着采样率的提高而逐步增加。将输入体素的采样率设置为 80%，可以在精度与效率之间取得最佳平衡。

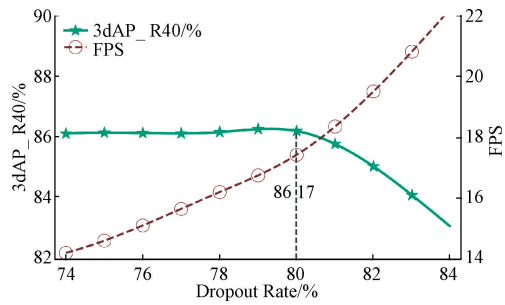
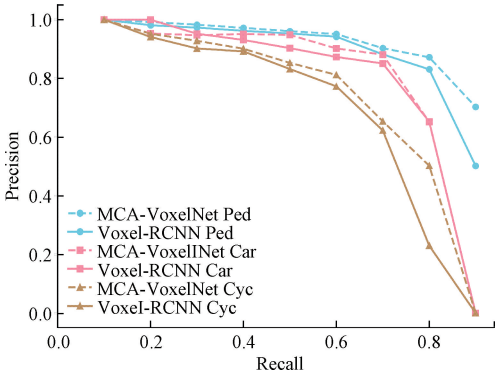


图 6　输入体素采样率梯度对检测精度和 FPS 的影响曲线图

Fig. 6　Input voxel sampling rate gradient effect curve on detection accuracy and FPS

图 7 展示了 MCA-VoxelNet 与基线模型在车辆、行人和骑行者三类目标上的 3D 检测任务中对应的 Precision-Recall 曲线对比结果。可观察到,所提模型在 3 个类别上均呈现出更为优越的曲线走势,整体更接近右上角,显示出在召回率提升的同时能够保持较高的精度。此外,MCA-VoxelNet 在各类别曲线下的面积显著大于基线模型,表明其在不同检测置信度阈值下均具有更稳定和准确的检测性能。该结果验证了模型在多类 3D 目标检测任务中的有效性和泛化能力。



注:虚线为本文方法,实线为基线方法

图 7　KITTI 验证集上的召回-精度曲线,MCA-VoxelNet 与 Voxel-RCNN 的比较

Fig. 7　Recall-precision curves of MCA-VoxelNet and Voxel-RCNN on the KITTI validation set

为了研究模型在哪些地方对基线模型的改善最大,基于不同的距离在 KITTI 验证集上评估了检测性能结果,检测指标为 Car 3D AP(R40)和 Car BEV AP(R40),结果如图 8、9 所示,折线图为模型性能对比,柱状图为检测精度提升。MCA-VoxelNet 对于范围在 20～40 m 远距离物体有显著的改进,改进主要来源于模型利用虚拟点融合的策略改进了原始点云在远距离物体的几何特征稀疏的缺点,更好地补充了远距离稀疏物体的几何特征,其次,引入多阶段级联注意力的检测头,在每个检测阶段之间建立有效的连接,通过级联注意力进行特征聚合,引用跨阶段聚合特征的方式可以增加目标的特征。

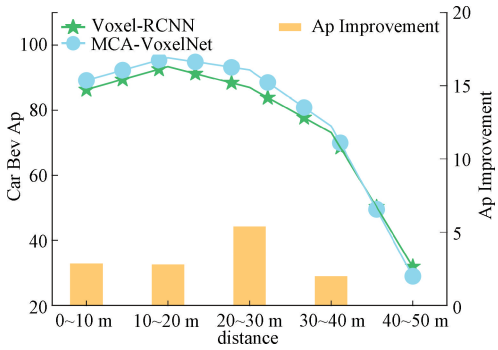


图 8　Car BEV AP(R40)在不同距离上的检测性能对比

Fig. 8　Comparison of detection performance of Car BEV AP (R40) at different distances

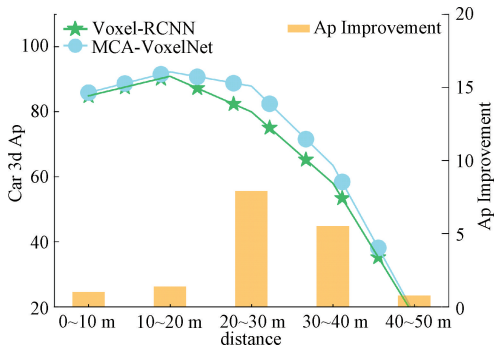


图 9　Car 3D AP(R40)在不同距离上的检测性能对比

Fig. 9　Comparison of detection performance of Car 3D AP (R40) at different distance

为评估模型在驾驶环境中对其他物体检测性能的提升,本文基于 3D AP(40)指标,对场景中的行人和骑行者检测结果进行了对比实验,结果如表 9 所示。与基线模型相比,MCA-VoxelNet 在所有检测类别上均表现出显著的性能提升。实验结果表明,该方法具有良好的通用性,能够有效推广至多类目标检测任务,从而进一步提升整体检测性能。

表 9　行人和骑行者检测精度对比结果

Table 9　Comparison of detection accuracy for pedestrians and cyclists　%

方法	Pedestrian 3D AP(R40)				Cyclist 3D AP(R40)			
	简单	中等	困难	平均	简单	中等	困难	平均
Voxel-RCNN	69.23	64.51	55.78	63.17	88.42	72.77	65.29	75.49
MCA-VoxelNet	**73.88**	**66.57**	**59.92**	**66.79**	**92.25**	**72.70**	**68.13**	**77.69**

为了突出 MCA-VoxelNet 在计算效率方面的优势,与现有方法的运行时间进行了详细分析,结果如表 10 所示。单模态(LiDAR)方法的平均运行时间为 66.06 ms,显著低于多模态(LiDAR+RGB)方法的 95.10 ms,这主要由于多模态方法需要处理更复杂的数据融合任务。MCA-VoxelNet 在多模态方法中以 79.34 ms 的运行时间表现出显著优势,不仅远低于多模态方法的平均值,还接近部分单模态方法的性能(PV-RCNN 的 72.89 ms)。通过高效的模态融合和计算优化,MCA-VoxelNet 在保证多模态方法高精度的同时,显著降低了计算负担,验证了其在计算效

表 10　不同方法运行时间对比结果

Table 10　Comparison of running time results for different methods

方法	模态	运行时间/ms
PV-RCNN	LiDAR	72.89
Voxel-RCNN	LiDAR	39.67
SE-SSD	LiDAR	29.91
SFD	LiDAR+RGB	121.74
TED-M	LiDAR+RGB	90.12
VPFNet	LiDAR+RGB	81.83
Ours	LiDAR+RGB	79.34

率上的优越性。

为验证模型在不同天气条件下的鲁棒性，在加拿大恶劣驾驶条件数据集(Canadian adverse driving conditions，CADC)[31]上进行了测试。该数据集包含雪天、低光照等复杂驾驶场景的激光雷达点云与图像数据。实验结果表明，模型在雪天环境下仍能保持稳定的检测性能，mAP 达到 65.06%，性能下降主要源于远距离目标被雪雨遮挡。尽管极端天气条件(如降雪、低能见度)对检测任务带来一定挑战，但模型整体表现可靠，验证了其在复杂场景下的鲁棒性。图 10 为在 CADC 数据集上的检测结果可视化图。

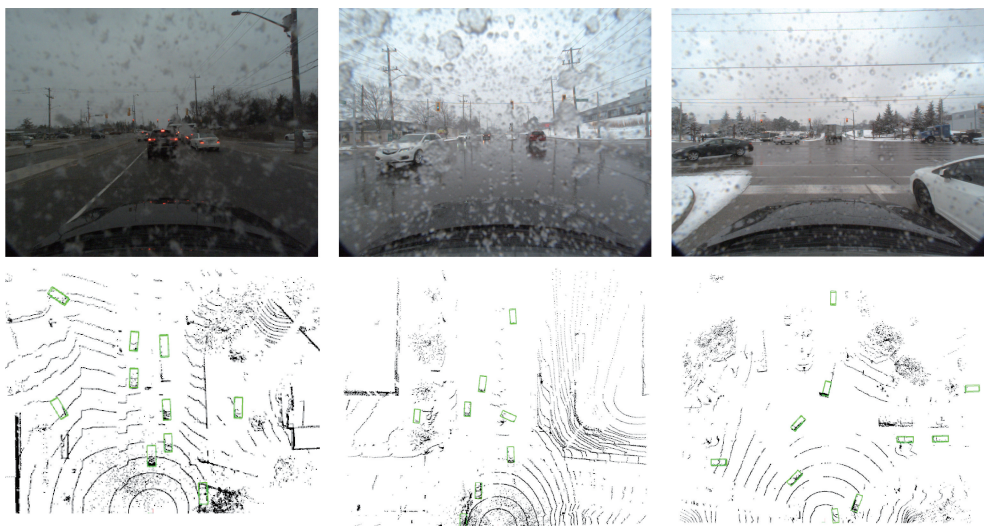


图 10　CADC 数据集检测结果可视化

Fig. 10　Visualization of CADC dataset inspection results

5　结　　论

文章提出了一种新的基于虚拟点的多模态三维目标检测框架，通过在原始点云中增加图像生成的虚拟点云以实现更准确的定位和回归，同时通过新设计的距离感知体素丢弃来解决虚拟点的密度问题，最后通过引入多阶段级联注意力检测结构来聚合多阶段的目标特征，解决了在多阶段三维目标检测中忽略远距离目标和误差传播的问题。

在权威的 KITTI 自动驾驶数据集上，MCA-VoxelNet 以 17.54 的 FPS 在难度分数阈值为 0.7、0.7、0.7 的简单、中等和困难难度的汽车检测上达到 94.19%、85.92%和 86.17%，对比基线模型 Voxel-RCNN 分别提高了 2.23%、3.57%和 5.68%，同时兼顾了检测精度和运行速度。此外，进行了一系列消融实验验证了 MCA 的有效性，对比自注意力机制分别提高了 0.82%、0.52%和 4.46%并最终确定 3 个检测阶段的最佳精度效率权衡；进行的一系列补充实验验证了距离感知体素丢弃模块参数设置在 80%能够达到最佳的精度效率权衡。

参考文献

[1]　陈慧娴，吴一全，张耀. 基于深度学习的三维点云分析方法研究进展[J]. 仪器仪表学报，2023，44(11)：130-158.
CHEN H X，WU Y Q，ZHANG Y. Research progress of 3D point cloud analysis method based on deep learning [J]. Chinese Journal of Scientific Instrument，2023，44(11)：130-158.

[2]　王庆林，李辉，谢礼志，等. 基于激光雷达点云的车辆目标检测算法改进研究[J]. 电子测量技术，2023，46(1)：120-126.
WANG Q L，LI H，XIE L ZH，et al. Research on improving vehicle target detection algorithm based on lidar point cloud [J]. Electronic Measurement Technology，2023，46(1)：120-126.

[3]　郑美琳，高建瓴. 融合多注意力机制与 PointRCNN 的三维点云目标检测[J]. 电子测量技术，2022，45(9)：127-132.

ZHENG M L, GAO J L. 3D point cloud object detection by integrating multiple attention mechanisms and PointRCNN [J]. Electronic Measurement Technology, 2022, 45(9): 127-132.

[4] 郑自立,徐健,刘秀平,等.联合多注意力和 C-ASPP 的单目 3D 目标检测[J].电子测量与仪器学报,2023,37(8):241-248.
ZHENG Z L, XU J, LIU X P, et al. Combinedmulti-attention and C-ASPP network for monocular 3D object detection[J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(8):241-248.

[5] AYALA R, MOHD T K. Sensors in autonomous vehicles:A survey[J]. Journal of Autonomous Vehicles and Systems, 2021, 1(3): 031003.

[6] CHEN X Z, MA H M, WAN J, et al. Multi-view 3D object detection network for autonomous driving[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017:1907-1915.

[7] HE CH H, ZENG H, HUANG J Q, et al. Structure aware single-stage 3D object detection from point cloud[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2023:11873-11882.

[8] LANG A H, VORA S, CAESAR H, et al. Pointpillars: Fast encoders for object detection from point clouds [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2019:12697-12705.

[9] CHARLES R Q, SU H, KAICHUN M, et al. PointNet: Deep learning on point sets for 3D classification and segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 77-85.

[10] QI C R, YI L, SU H, et al. PointNet++: Deep hierarchical feature learning on point sets in a metric space[C]. 31st International Conference on Neural Information Processing Systems, 2017:5105-5114.

[11] SHI S S, WANG X G, Li H S. Pointrcnn: 3D object proposal generation and detection from point cloud[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2019:770-779.

[12] YANG Z T, SUN Y N, LIU S, et al. 3DSSD: Point-based 3D single stage object detector [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2023:11040-1048.

[13] DENG J J, SHI S S, LI P W, et al. Voxel R-CNN: Towards high performance voxel-based 3D object detections [C]. AAAI Conference on Artificial Intelligence, 2021:1201-1209.

[14] YAN Y, MAO Y, LI B. Second: Sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.

[15] SHI S S, GUO C X, JIANG L, et al. PV-RCNN: Point voxel feature set abstraction for 3D object detection[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2020:10526-10535.

[16] YIN J, SHEN J, CHEN R N, et al. IS-FUSION: Instance-scene collaborative fusion for multimodal 3D object detection [C]. IEEE/CVF International Conference on Computer Vision, 2024: 14905-14915.

[17] HU Z F, ZHOU S F, ZHAO S B, et al. MVCTrack: Boosting 3D point cloud tracking via multimodal-guided virtual cues[J]. ArXiv preprint arXiv: 2412.02734, 2024.

[18] ZHANG G X, SONG Z Y, LIN L, et al. FGU3R: Fine-grained fusion via unified 3D representation for multimodal 3D object detection[J]. ArXiv preprint arXiv:2501.04373, 2025.

[19] 陈熙源,戈明明,姚志婷,等. 雨雪天气下的激光雷达滤波算法研究[J].仪器仪表学报, 2023, 44(7): 172-181.
CHEN X Y, GE M M, YAO ZH T, et al. Research on LiDAR filtering algorithm for rainy and snowy weather[J]. Chinese Journal of Scientific Instrument, 2023, 44(7):172-181.

[20] WANG L, SONG Z Y, ZHANG X Y, et al. SAT-GCN: Self-attention graph convolutional network-based 3D object detection for autonomous driving[J]. Knowledge Based Systems, 2023, 259: 110080.

[21] QI C R, LIU W, WU C, et al. Frustum pointnets for 3D object detection from RGB-D data [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018:918-927.

[22] YIN T W, ZHOU X Y, KRÄHENBÜHL P. Multimodal virtual point 3D detection [C]. Neural Information Processing Systems, 2021: 16494-16507.

[23] WU X P, PENG L, YANG H H, et al. Sparse fuse dense: Towards high quality 3D detection with depth completion[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2022: 5418-5427.

[24] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2012:3354-3361.

[25] HU Q Y, YANG B, XIU L H, et al. RandLA-Net: Efficient semantic segmentation of large-scale point clouds[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020: 11105-11114.

[26] WANG Y, CHAO W L, GARG D, et al. Pseudo-LiDAR from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2019: 8445-8453.

[27] YOU Y R, WANG Y, CHAO W L, et al. Pseudo-LiDAR++: Accurate depth for 3D object detection in autonomous driving [J]. ArXiv preprint arXiv: 1906.06310, 2019.

[28] CAI ZH W, VASCONCELOS N. Cascade R-CNN: Delving into high quality object detection[C]. IEEE/CVF International Conference on Computer Vision, 2018: 6154-6162.

[29] ZHU H Q, DENG J J, ZHANG Y, et al. VPFNet: Improving 3D object detection with virtual point based lidar and stereo data fusion[J]. ArXiv preprint arXiv: 2111.14382, 2019.

[30] WU H, WEN C L, LI W, et al. Transformation-equivariant 3D object detection for autonomous driving[C]. AAAI Conference on Artificial Intelligence, 2023: 2795-2802.

[31] PITROPOV M, GARCIA D, REBELLO J, et al. Canadian adverse driving conditions dataset[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2020: 681-690.

作者简介

李旭，硕士研究生，主要研究方向为无人驾驶目标检测，无人驾驶环境感知等。

E-mail：202312490469@nuist.edu.cn

张永宏（通信作者），博士，教授，主要研究方向为人工智能、遥感大数据分析与深度学习。

E-mail：zyh@nuist.edu.cn

朱灵龙，博士，讲师，主要研究方向为博士，讲师，主要研究方向为遥感大数据分析与深度学习应用。

E-mail：llzhu@cwxu.edu.cn

阚希，博士，副教授，主要研究方向为遥感图像处理与深度学习。

E-mail：kanxi@cwxu.edu.cn