•机械工程•
DOI:10.12454/j.jsuese.202500476
本刊网刊

基于多尺度融合的车载激光点云小目标检测方法

王寅泰[1,2],潮　群[1,2*],张腾飞[3],吴江东[1,2],刘成良[1,2]

(1.上海交通大学 机械与动力工程学院,上海 200240;2.上海交通大学 机械系统与振动全国重点实验室,上海 200240;
3.江苏徐工国重实验室科技有限公司, 江苏 徐州 221116)

摘　要:本文提出了一种基于多尺度融合的车载激光雷达点云小目标检测方法,旨在解决车载算力受限的消防车无人驾驶场景中小目标检测精度低的问题。首先,对原始激光点云进行体素化处理以减少计算量;其次,设计点云–体素注意力机制模块,利用残差连接和多头注意力机制聚合体素内点云特征;然后,将聚合后的特征输入二维卷积神经网络提取特征,并通过高效多尺度注意力模块实现多通道特征融合,增强对小目标的感知能力;最后,使用单次多框检测器完成目标分类与检测回归。针对体素化易造成局部几何信息损失及多尺度目标特征差异显著的问题,所设计的模块分别增强了体素内特征聚合与跨尺度通道表达,消融实验验证了二者的有效性。在消防车无人驾驶数据集上的实验结果表明所提方法的目标检测平均精度均值达78.1%,且小目标检测精度显著优于其他方法;本文方法的单帧推理时间为14.39 ms;在公开数据集KITTI上进一步验证了小目标检测的优势,证实本文方法能兼顾消防车无人驾驶场景中小目标检测精度和实时性需求。

关键词:无人驾驶;激光雷达点云;小目标检测;注意力机制;多尺度融合
中图分类号:TP391.4　　**文献标志码:**A　　**文章编号:**2096-3246(2026)06-0328-12

随着人工智能、传感器技术和大数据计算能力的飞速发展,消防车无人驾驶技术逐渐受到学术界和工业界的关注。传统的消防作业依赖人工操作,在火灾、烟雾等极端条件下,消防员存在较高的伤亡风险。而无人驾驶消防车通过自主导航和智能决策,能够在危险区域代替人工完成侦查、灭火等任务,从而显著提高灭火救援行动的安全性。无人驾驶技术在消防场景下的应用研究,为现代消防体系的智能化转型提供了新的思路。

在消防车无人驾驶系统中,环境感知是实现自主导航的核心环节,而三维目标检测作为环境感知的核心模块,发挥着不可替代的作用。激光雷达点云数据能够精确捕捉周围三维环境的几何信息,成为三维目标检测的重要数据源之一[1–2]。然而,传统三维点云目标检测方法在计算资源受限条件下难以同时满足对实时性和准确性的需求[3],尤其是在检测小目标时(如行人、路障等)表现不佳[4]。消防场景下小目标检测对自动驾驶安全至关重要,而小目标由于点云数据稀疏,特征表达不明显,容易被忽略;若设计复杂神经网络结构,则会导致推理时间过长,无法适应消防应用场景。因此,设计一种轻量化且准确性高的三维目标检测器,成为一项亟待解决的问题。

在激光雷达三维目标检测领域,点云的处理策略主要分为3类:基于点的处理策略、基于体素的处理策略和混合处理策略[5]。基于点的处理策略直接对原始点云数据进行操作,能够完整保留点云的几何信息。例如,PointNet模型通过直接提取点云特征,捕获物体三维结构[6]。PointNet++模型在PointNet基础上进一步引入多层次特征聚合,增强了局部和全局特征的捕捉能力[7]。IA–SSD模型通过改进投票机制得到更好的种子点特征,从而提高了目标中心点预测、边界预测和目标分类精度[8]。Shreelakshmi等[9]引入图神经网络处理点云数据,进一步提升了目标检测精度。基于点的处理策略可以充分利用原始点云

收稿日期:2025 – 06 – 24　　**修回日期:**2025 – 10 – 29　　**网络出版日期:**2025 – 11 – 13
基金项目:国家重点研发计划青年科学家项目(2022YFD2000400)
作者简介:王寅泰(2001—),男,硕士生. 研究方向:自动驾驶与感知识别. E-mail: 8972www@sjtu.edu.cn
* **通信作者:**潮　群,副教授, E-mail: chaoqun@sjtu.edu.cn

几何信息，但由于重复采样和聚合，存在计算量偏大的问题。相比之下，基于体素的处理策略将点云划分为规则的体素网格，利用卷积神经网络（CNN）提取特征，从而显著提高计算速率。例如，PointPillar将点云体素化后使用CNN进行处理，在实时性上取得了较大突破[10]。Voxel R–CNN通过在生成置信框时引入格点，再以这些格点为中心提取三维卷积模块中生成的体素特征，进一步提高了检测精度[11]。ConQueR模型[12]基于检测自注意力网络[13]的三维检测器构建了一种新颖的对比策略，可以有效消除密集重叠误报，实现更高精度的预测。然而，体素化过程中可能会丢失点云部分几何信息，导致小目标检测效果不佳。混合处理策略试图结合基于点和基于体素处理策略的优势。例如，Fast Point R–CNN[14]检测模型分为两阶段：第一阶段先将原始点云划分为体素作为输入，再通过神经网络生成候选框；第二阶段利用PointNet[6]对候选框及其周围点进行特征聚合，从而预测出目标种类与边界框大小。PV–RCNN[15]通过融合点与体素特征，在检测精度上取得了突破。Shi 等[16]进一步提出了一种新的三维检测器 PV–RCNN++，该检测器通过矢量池聚合，以更少计算量实现更好的局部点特征聚合。然而，混合处理策略一般具有较复杂的网络结构，难以在算力受限的场景下实车部署。考虑到消防车车载算力有限，本文选择体素处理策略以满足轻量化需求。

近年来，注意力机制在计算机视觉领域展示出了强大潜力，它通过模拟人脑的注意力分配方式，能够自动捕捉输入张量中的关键部分[17–18]，从而提升模型识别能力。在点云处理领域，注意力机制被广泛应用于各类目标检测与实例分割任务中。例如：Point Transformer[19]通过自注意力机制对点云局部特征进行聚合，大大提升了点云分类与分割效果；PDV[20]基于点密度感知体素网络，利用自注意力机制和密度编码提升了目标检测的鲁棒性；RAANet[21]基于范围感知注意力机制来提取鸟瞰图特征，改善了近距和远距物体的特征提取效果；SA–SSD[22]、MA–MFFC[23]、STA–GCN[24]等方法均从不同角度引入注意力机制，从而有针对性地处理稀疏点云。小体型目标因点云较为稀疏导致特征不明显，传统方法难以准确捕捉其边界和类别信息，注意力机制通过增强模型对稀疏点云的特征感知能力，为小目标点云检测提供了新的研究思路。

在消防车无人驾驶场景下，三维目标检测面临两大核心挑战。其一，在检测精度层面，针对行人等小尺寸目标，点云固有的稀疏性导致特征表达不充分，进而影响检测算法精度。其二，在实时部署层面，车载平台有限的计算资源约束了检测模型的复杂度，形成了算法性能与运行效率之间的矛盾。针对以上挑战，本文在保持轻量化的前提下，引入点云–体素注意力模块与高效多尺度注意力模块，提出了一种基于多尺度融合的车载激光点云小目标检测方法。首先，对原始点云进行体素化处理，降低计算复杂度并保留点云基本几何信息。其次，设计点云–体素（point to voxel，PTV）注意力机制模块，通过残差查询的方式构造查询向量，并利用多头注意力机制聚合体素内的点云特征，提升对小目标个体的感知能力。然后，将聚合后的特征张量输入CNN主干网络，使用高效多尺度注意力（efficient multi-scale attention，EMA）模块捕捉多通道特征，提升模型对不同检测种类的表达能力。最后，使用单次多框检测器（single shot multi-box detector，SSD）完成目标分类与检测框回归，生成最终预测结果。整个网络设计在体素化的基础上融合了多通道特征，兼顾了轻量化与准确性，适合算力受限的消防车无人驾驶场景。通过激光点云数据集和实车测试，验证所提方法在小目标检测任务中的有效性和优越性。

本文的主要贡献如下：1）针对小目标点云稀疏的问题，设计了PTV模块，通过残差查询与多头注意力机制，动态聚合点云特征；2）设计了EMA模块，通过捕捉多尺度多通道特征提升对消防车、行人等不同大小种类的感知能力；3）构建了消防车无人驾驶场景下的三维目标检测数据集，并通过实车部署验证了所提方法的有效性，为后续研究提供了参考。

1　小目标检测方法

1.1　方法概述

针对车载激光点云小目标检测点云数据稀疏、特征表达不明显和算力有限的特点，本文提出一种轻量化、高精度的检测架构。通过点云体素化降低整体的计算量，融合PTV与EMA两个注意力模块以增强对小目标的特征提取能力，最终通过SSD实现快速准确的目标分类与回归。本文提出的小目标检测算法整体结构如图1所示，包括激光雷达点云获取、点云数据处理、体素特征提取网络、二维点云主干网络4部分。图1中，P为非空体素个数，C为体素特征维数，W和H分别为映射至二维后的宽与高，结果输出的(w,l,h)表示物体的宽度、长度和高度。

首先，使用车载激光雷达采集周围环境点云，分割采集到的点云数据，裁剪出感兴趣区域，并进行翻转、平移等数据增强操作，划分训练集、验证集和测试

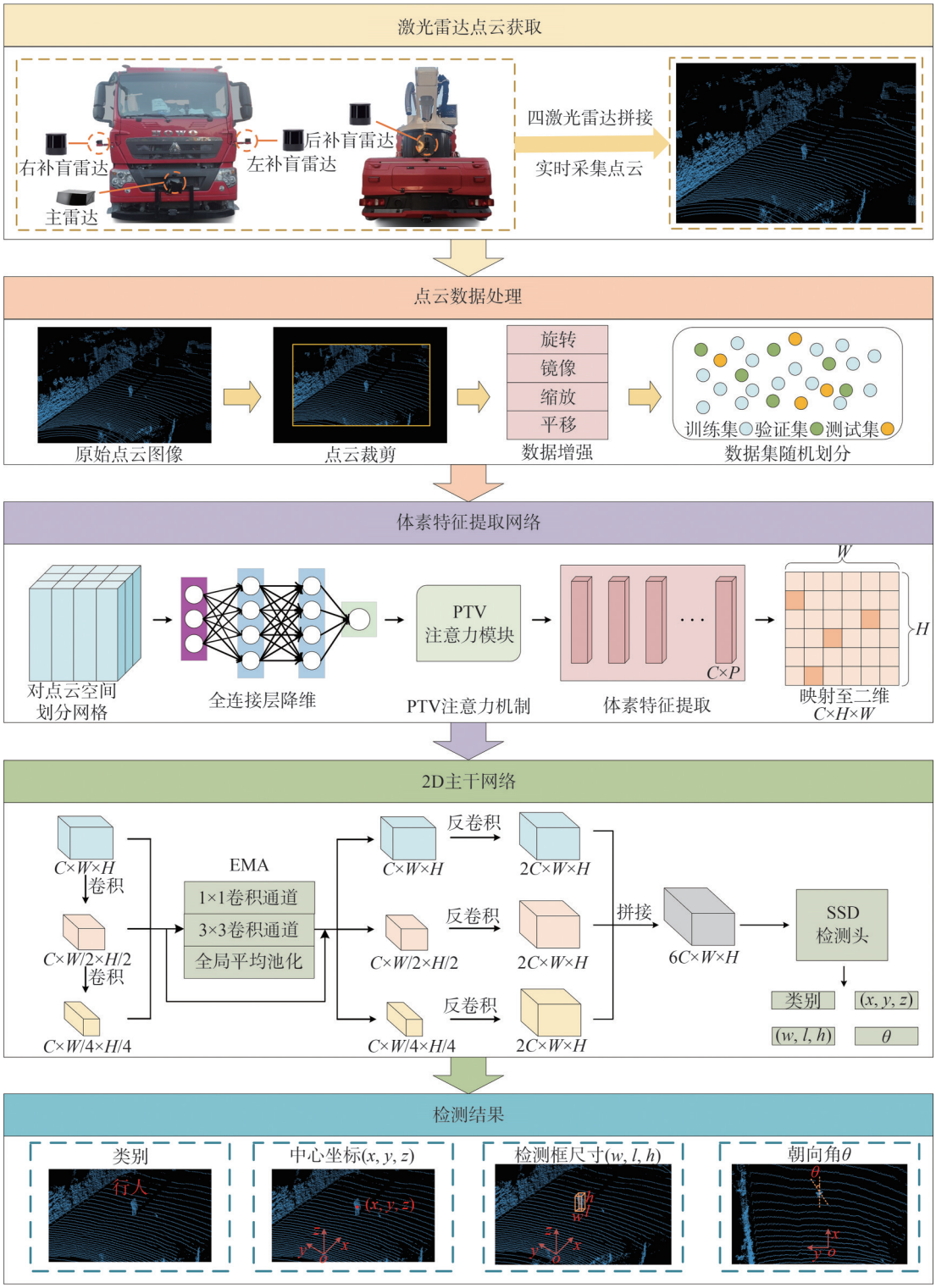


图 1　目标检测网络的整体架构

Fig. 1　Overall architecture of the object detection network

集，准备输入体素特征提取网络进行处理。其次，对输入的单帧点云进行体素划分，在 x–y 平面上划分细小立方体，每个立方体为一个体素，将所有点分配至均等大小的体素中。该设计的核心目的是提升计算效率。相比于直接处理几十万个离散点，基于体素的方法将无序的点云划分至规则的体素中，大幅减少了需要处理的数据数量。随后，通过对体素内的点进行特征编码并将三维体素空间压缩为二维伪图像，使得后续可以用计算效率更高的二维 CNN 完成特征提取。这一系列操作避免了计算成本高昂的三维卷积与复杂的点集查询，有效地降低了模型的推理时间。体素划分后，对其中的点云进行数据增强，原点云数据由

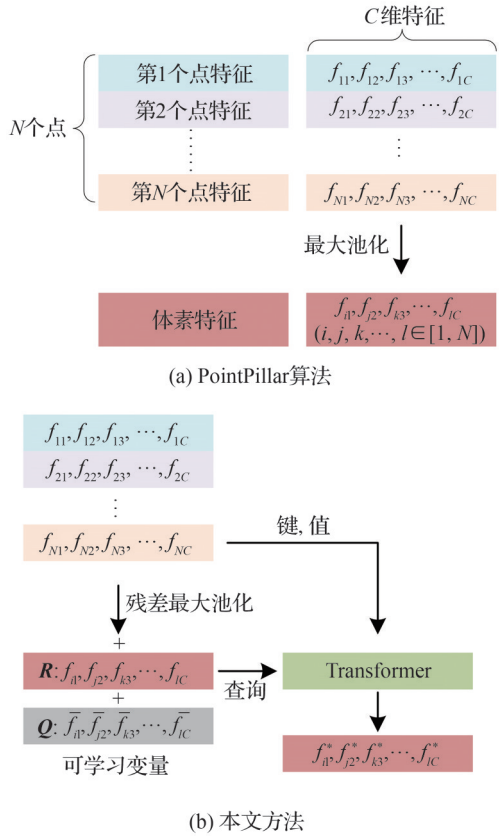
四维(x,y,z,r)扩充至十维($x,y,z,r,x_p,y_p,z_p,x_c,y_c,z_c$)，其中，($x,y,z$)表示点云三维坐标，$r$表示反射强度，下标$c$表示每个点到对应体素内所有点平均值的偏移量，下标$p$表示该点距离所在体素中心点的偏移量。单帧点云中包含P个非空体素，每个体素中过滤至N个点，每个点由十维向量$\boldsymbol{D}$组成，因此该帧点云可以由张量($\boldsymbol{D},N,P$)表示。采取轻量化的全连接层处理体素内的点，将向量$\boldsymbol{D}$降维至向量$\boldsymbol{C}$，构成新的张量($\boldsymbol{C},N,P$)。通过降维扩展维度特征，捕捉体素内点云的局部几何信息与上下文信息。为进一步增强特征表达质量，在特征编码阶段引入了PTV注意力机制模块，通过聚合体素内N个点的特征，提升小目标感知能力。该模块的详细设计将在第1.2节详细阐述。经过PTV模块后，生成维度为($\boldsymbol{C},\boldsymbol{P}$)的平面特征图，将$\boldsymbol{P}$拆分为$H\times W$两个维度，构成伪图像的高与宽，将每个体素数据重新放回原来的坐标分布中来组成伪图像数据，三维点云图像压缩至二维可以大幅提高模型的推理速率。然后，在二维点云主干网络阶段，输入$C\times H\times W$大小的二维伪图像张量，使用卷积操作对其特征进行多尺度提取与拼接融合。主干网络部分先采用下采样不断缩小平面特征图的分辨率并提升其维度，然后集成多尺度注意力EMA模块，利用跨通道和空间注意力机制捕获多尺度上下文信息，增强网络在不同目标尺度上的适应性，EMA模块的详细设计将在第1.3节描述。对EMA增强后的张量再进行反卷积操作，生成相同维数的特征图，并对其进行全连接操作，生成全连接层。最终，将全连接层张量输入SSD检测头，利用多尺度特征图检测不同尺寸目标，生成检测结果，包括目标类别和边界框中心坐标、几何尺寸和朝向角。

1.2　PTV模块设计

激光雷达目标检测中，点云的稀疏性导致部分体素内点数偏少，传统的最大池化方法容易丢失局部几何信息。针对这一问题，引入了基于Transformer的注意力机制模块[25]。PointPillar算法与本文所提方法对比，如图2所示。PointPillar直接使用最大池化提取体素内N个点的特征，而本文算法通过多头注意力机制动态学习体素内点特征的加权聚合方式，从而保留更多局部信息，并增强小目标个体的特征表达能力。

注意力机制通过查询(query)、键(key)和值(value)向量计算注意力权重，从而实现输入数据的动态加权聚合。对于每个体素内的点特征张量$\boldsymbol{L}\in\mathbb{R}^{N\times C}$，通过最大池化提取全局特征$\mathrm{maxpool}(\boldsymbol{L})\in\mathbb{R}^{1\times C}$，之后与可学习变量$\boldsymbol{Q}\in\mathbb{R}^{1\times C}$相加，得到残差查询向量：

$$\boldsymbol{R}_{\mathrm{Q}}=\mathrm{maxpool}(\boldsymbol{L})+\boldsymbol{Q} \tag{1}$$



注：f_{NC}表示体素内第N个点的第C维特征，f_{lC}表示随训练而更新的第C维特征，f^*_{lC}表示最终输出的体素特征。

图 2　PointPillar与本文所提方法对比

Fig. 2　Comparison between PointPillar and the proposed method

利用多头注意力机制，以$\boldsymbol{R}_{\mathrm{Q}}$作为查询，$\boldsymbol{L}$作为键和值，计算注意力输出：

$$\boldsymbol{V}=\mathrm{softmax}\left(\frac{\boldsymbol{R}_{\mathrm{Q}}\cdot\boldsymbol{L}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{L} \tag{2}$$

式中：$\boldsymbol{V}$为聚合后的体素特征，$\boldsymbol{V}\in\mathbb{R}^{1\times C}$；$d$为缩放因子，用于稳定梯度下降。

1.3　EMA模块设计

在消防场景中，小目标与大目标特征尺度差异显著，传统的卷积网络难以有效利用多通道特征，导致小目标检测效果不佳。针对该问题，本文在处理稀疏点云与多类别多尺度目标环节引入EMA模块[26]。EMA模块通过多尺度和跨空间学习，增强网络对不同尺度和空间位置特征的学习能力，从而提升特征表示质量。EMA模块架构如图3所示，由两个并行分支构成：一个1×1的卷积分支与一个3×3的卷积分支。1×1卷积分支通过通道分组捕捉通道间的特征关系，而3×3卷积分支通过局部空间卷积提取多尺度空间特征。此外，EMA模块还利用全局平均池化和Sigmoid激活

函数生成空间注意力图，从而实现跨空间的特征聚合。

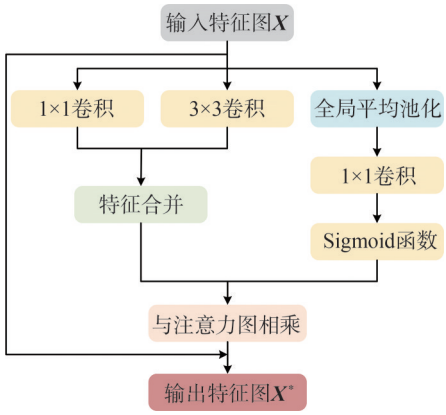


图 3 EMA 模块架构图

Fig. 3 Architecture diagram of the EMA module

在 EMA 模块输入张量 $\boldsymbol{X}\in\mathbb{R}^{C\times H\times W}$，经过 1×1 卷积操作 $\mathrm{Conv}_{1\times1}$ 后变为：

$$\boldsymbol{X}_1=\mathrm{Conv}_{1\times1}(\boldsymbol{X},\theta_1) \tag{3}$$

式中，θ_1 为 1×1 卷积网络参数。

经过 3×3 卷积操作 $\mathrm{Conv}_{3\times3}$ 后变为：

$$\boldsymbol{X}_2=\mathrm{Conv}_{3\times3}(\boldsymbol{X},\theta_2) \tag{4}$$

式中，θ_2 为 3×3 卷积网络参数。

通过全局平均池化生成通道注意力为：

$$z=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{X}(i,j) \tag{5}$$

通过 1×1 卷积和 Sigmoid 激活函数生成注意力图为：

$$\boldsymbol{a}=\sigma(\mathrm{Conv}_{1\times1}(z,\theta_{\mathrm{a}})) \tag{6}$$

式中，σ 为 Sigmoid 函数，θ_{a} 为注意力卷积参数。

将 1×1 与 3×3 分支的输出相加，并通过空间注意力图 $\boldsymbol{a}$ 进行加权得到增强特征。最终通过残差将增强特征与原始输入 $\boldsymbol{X}$ 相加，得到最终输出为：

$$\boldsymbol{X}^{*}=\boldsymbol{X}+(\boldsymbol{X}_1+\boldsymbol{X}_2)\odot\boldsymbol{a} \tag{7}$$

式中，⊙表示逐元素相乘。

通过将 EMA 模块放置在二维主干网络的下采样层与上采样层之间，确保 EMA 模块能够增强每个尺度的特征，在融合前能够捕捉到多尺度和跨空间的特征依赖关系。

1.4 损失函数

每个待检测物体的三维框可以表示为一个七维向量 (x,y,z,w,h,l,θ)。定义目标检测任务中向量 (x,y,z,w,h,l,θ) 中各变量的回归残差如式（8）～（14）所示：

$$\Delta x=\frac{x^{\mathrm{gt}}-x^{\mathrm{a}}}{d^{\mathrm{a}}} \tag{8}$$

$$\Delta y=\frac{y^{\mathrm{gt}}-y^{\mathrm{a}}}{d^{\mathrm{a}}} \tag{9}$$

$$\Delta z=\frac{z^{\mathrm{gt}}-z^{\mathrm{a}}}{h^{\mathrm{a}}} \tag{10}$$

$$\Delta w=\lg\frac{w^{\mathrm{gt}}}{w^{\mathrm{a}}} \tag{11}$$

$$\Delta l=\lg\frac{l^{\mathrm{gt}}}{l^{\mathrm{a}}} \tag{12}$$

$$\Delta h=\lg\frac{h^{\mathrm{gt}}}{h^{\mathrm{a}}} \tag{13}$$

$$\Delta\theta=\sin(\theta^{\mathrm{gt}}-\theta^{\mathrm{a}}) \tag{14}$$

式（8）～（14）中：$(x^{\mathrm{gt}},y^{\mathrm{gt}},z^{\mathrm{gt}},w^{\mathrm{gt}},h^{\mathrm{gt}},l^{\mathrm{gt}},\theta^{\mathrm{gt}})$ 为每个待检测物体三维框的真实值；$(x^{\mathrm{a}},y^{\mathrm{a}},z^{\mathrm{a}},w^{\mathrm{a}},h^{\mathrm{a}},l^{\mathrm{a}},\theta^{\mathrm{a}})$ 为每个待检测物体三维框的预测值；$(\Delta x,\Delta y,\Delta z,\Delta w,\Delta h,\Delta l,\Delta\theta)$ 为每个物体三维框真实值与预测值之间的回归残差；d^{a} 为锚框底面对角线长度，$d^{\mathrm{a}}=\sqrt{(w^{\mathrm{a}})^2+(l^{\mathrm{a}})^2}$。

基于上述真实值与预测值之间的偏移量以及 SmoothL1 损失函数[27]，定义定位损失函数 $\mathcal{L}_{\mathrm{loc}}$ 为：

$$\mathcal{L}_{\mathrm{loc}}=\sum_{b\in(x,y,z,w,l,h,\theta)}\mathrm{SmoothL1}(\Delta b) \tag{15}$$

式中，b 为 (x,y,z,w,h,l,θ) 中的变量，Δb 为 b 的回归残差。

为避免物体检测方向错误，引入 Softmax 损失函数学习物体检测方向[27]，待检测物体三维框的方向损失函数记为 $\mathcal{L}_{\mathrm{dir}}$。对于物体检测的类别分类，采用 Focal Loss 函数作为损失函数来调节正负样本均衡[10]，待检测物体三维框的分类损失函数为：

$$\mathcal{L}_{\mathrm{cls}}=-\alpha_{\mathrm{a}}(1-p^{\mathrm{a}})^{\gamma}\lg p^{\mathrm{a}} \tag{16}$$

式中，p^{a} 为待检测物体三维框的类别概率，α_{a} 和 γ 为上述 Focal Loss 函数的参数。

定义目标检测网络总损失函数为：

$$\mathcal{L}=\frac{1}{N_{\mathrm{pos}}}(\beta_{\mathrm{loc}}\mathcal{L}_{\mathrm{loc}}+\beta_{\mathrm{cls}}\mathcal{L}_{\mathrm{cls}}+\beta_{\mathrm{dir}}\mathcal{L}_{\mathrm{dir}}) \tag{17}$$

式中，N_{pos} 为正样本物体锚框数量，β_{loc}、β_{cls} 和 β_{dir} 为上述各损失函数的权重系数。

2 实验验证及结果讨论

2.1 数据集制作

实验用无人驾驶消防车共配备 4 个激光雷达，包含 1 个主激光雷达 AT128 和 3 个补盲雷达 QT128，消防车传感器安装位置如图 4 所示。主激光雷达 AT128 安装在消防车正前方固定支架上，左右补盲雷达 QT128 安装在消防车两侧，后补盲雷达安装在车尾。AT128 的水平视场角为 120°，垂直视场角为 25.4°，最大探测距离为 210 m；QT128 的水平视场角为 360°，垂直视场角为 105°，最大探测距离为 20 m。通过对 4 个激光雷达点云进行校正和拼接，将输入的 4 组点云组合为一组，实现对车辆周围环境 360°无死角检测。



图 4　无人驾驶消防车传感器安装位置分布

Fig. 4　Sensor installation distribution of the unmanned fire truck

采集消防车不同场景、角度和位置下的激光雷达点云数据，经过处理后得到消防车、私家车、行人、消防机器人和路障5类障碍物的点云数据共2 558帧，格式为pcd文件。

参照公开数据集Kitti[28]与Nuscenes[29]制作消防车无人驾驶数据集。对采集到的点云pcd文件，使用SUSTechPoints点云标注工具[30]手工标注所有检测目标，选用最小外接长方体框选出点云中的检测目标，分别用“Firetruck”“Firefightingrobot”“Car”“Person”“Roadblock”作为消防车、消防机器人、私家车、行人与路障5类检测目标的标签。最后，将标注完成的2 558帧点云pcd文件，按照7∶2∶1的比例划分为训练集、验证集和测试集。图5为5种不同检测目标的相机拍摄图、激光雷达点云图和标注图，后续检测时仅使用点云输入。

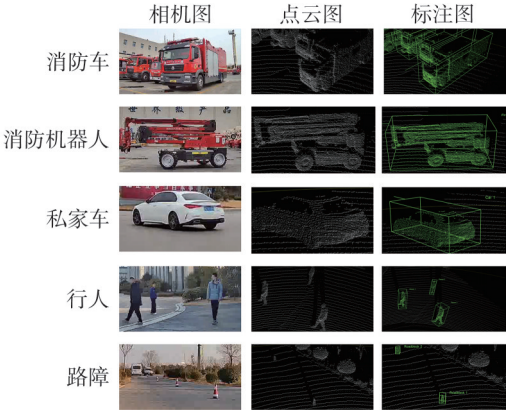


图 5　数据集中不同检测目标

Fig. 5　Different detection objects in the dataset

2.2　模型训练

实验中使用的服务器配置与系统环境如下：操作系统为Ubuntu-18.04，中央处理器CPU选用Intel(R) Core(TM) i9-14900K，图形处理器GPU选用NVIDIA RTX A6000，软件环境采用Python 3.8.10编程语言，基于PyTorch 1.10深度学习框架搭建，配套CUDA 11.3与cuDNN 8.2.0实现GPU加速计算。优化后的模型训练参数如表1所示。

表 1　模型训练参数

Tab. 1　Model training parameters

参数	数值
点云范围/m	[−23.04, −24.80, −1.00, 69.12, 24.80, 2.00]
训练批量大小	16
训练总轮数	160
初始学习率	0.003
权重衰减正则项	0.01
动量	0.9

2.3　评价指标

本文算法评估指标包括模型推理实时性与检测精度两方面。模型推理实时性通过每帧推理时间来衡量，检测精度通过精确率P'、召回率R、平均精度A_{P}、平均精度均值m_{AP}来衡量，计算公式为：

$$P'=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}} \tag{18}$$

$$R=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}} \tag{19}$$

$$A_{\mathrm{P}}=\int_0^1 P'\cdot R\,\mathrm{d}R \tag{20}$$

$$m_{\mathrm{AP}}=\frac{1}{n}\sum_{i=1}^{n}A_{\mathrm{P}i} \tag{21}$$

式(18)～(21)中，T_{P}为预测为正实际也为正例的数目，F_{P}为预测为正实际为负例的数目，F_{N}为预测为负实际为正例的数目，$A_{\mathrm{P}i}$为第i种的平均精度(i取1～n)。预测结果的正负通过交并比(intersection over union，IoU)来衡量，预测边界框与真实边界框的3D IoU大于0.5时，认为预测正确；小于0.35时，认为预测错误。该IoU阈值的设定主要参考了主流三维目标检测公开数据集Kitti的官方评价标准。在Kitti基准中，对于行人与自行车等小目标类别，普遍采用0.50与0.35的阈值进行评估。考虑到本文的数据集中也包含大量行人、路障等小目标，且本文方法主要针对小目标识别，因此该阈值能够更好地反映不同模型在本文数据集上的检测能力。

然而，对于行人等典型的小目标，3D IoU作为评价指标具有较高的敏感性。一个在绝对距离上偏差很小的预测框，其IoU值也可能低于0.35，从而被计算为假阴性。这一定程度上解释了大部分方法中行人预测的精度要低于消防车与消防机器人这些大目标的原因。因此，可以考虑引入基于中心点距离的评价指标或在多个IoU阈值下进行综合评估，以更全面地衡量模型的预测精度。

2.4 实验结果与分析

表2对比了本文方法与Second[27]、CenterPoint[31]、PV−RCNN、PointPillar、TED−S[32]5种算法的平均精度。表3对比了上述方法的推理时间。表4对比了上述方法在公开数据集KITTI上的表现。

由表2可知：在检测精度方面，本文方法在行人、私家车、路障等小目标类别上得到了显著提升，在行人目标上的平均精度达81.7%，远高于第二名(PointPillar，67.4%)；经计算，本文方法对各类障碍物检测平均精度均值为78.1%，表明本文方法在小目标检测场景有优势；在消防车与消防机器人这两类大尺寸目标上，本文方法平均精度略低于其他方法。对于大尺寸目标，其完整的几何结构分布在大量的连续体素中，而本文方法基于PointPillar架构，侧重于体素内的特征提取，在跨体素的全局上下文关联上，可能不如PV−RCNN等同时结合点级与体素级特征的更复杂网络。由表3可知，在实时性方面，本文方法的推理时间约14.39 ms，仅次于PointPillar的11.50 ms；参数量为5.05×10^6、模型大小为60.6 MB，接近第一名的结果。综合表2、3可知，本文方法在保障实时性的同时，显著提高了对小目标个体的检测效果。由表4可知，本文方法在KITTI数据集上对行人这类小目标上仍有较好的表现，平均精度高于其他方法，证明了本文方法在小目标上的有效性。

表 2 各类算法目标检测平均精度对比

Tab. 2 Comparison of A_{P} among different object detection algorithms

%

方法	消防车	行人	消防机器人	私家车	路障
Second	78.9	45.5	90.3	40.9	35.8
CenterPoint	78.4	44.2	90.5	31.9	59.8
PV−RCNN	71.9	41.0	90.8	45.5	34.0
PointPillar	62.1	67.4	88.9	76.9	64.8
TED−S	30.5	34.4	81.6	65.6	51.7
本文方法	63.6	81.7	89.7	81.2	74.3

表 3 各类算法模型复杂度与推理时间对比

Tab. 3 Comparison of model complexity and inference time among different algorithms

方法	时间/ms	参数量/10^6	模型大小/MB
Second	18.68	5.40	64.9
CenterPoint	29.09	6.92	83.2
PV−RCNN	41.03	13.19	158.5
PointPillar	11.50	4.91	59.0
TED−S	87.62	111.8	93.1
本文方法	14.39	5.05	60.6

表 4 各类算法在KITTI数据集上平均精度对比

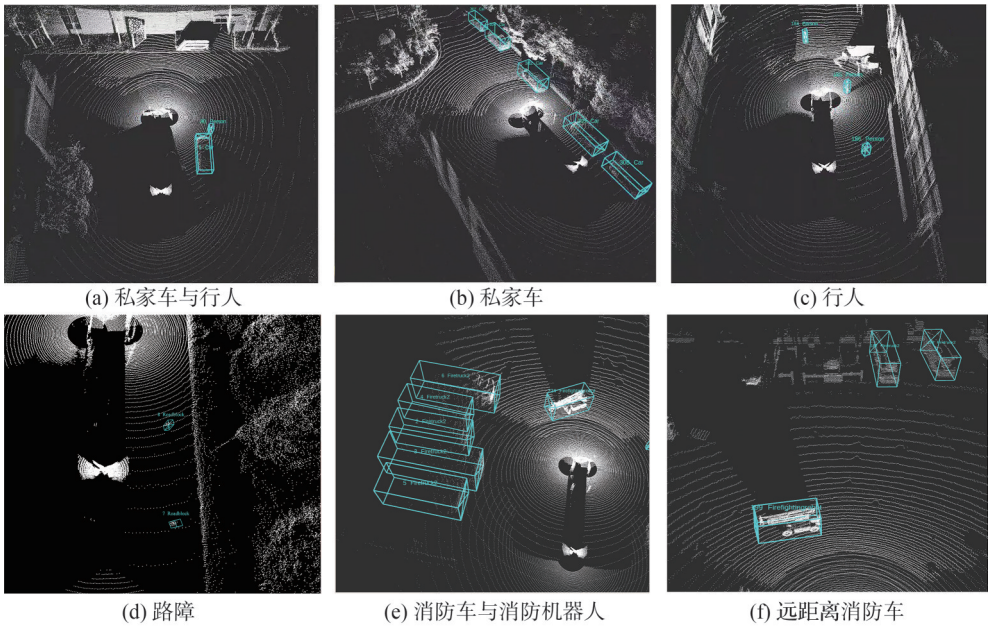
Tab. 4 Comparison of A_{P} of different algorithms on the Kitti dataset

%

方法	车辆			行人			自行车		
	简单	中等	困难	简单	中等	困难	简单	中等	困难
Second	89.9	81.6	78.9	55.3	49.6	45.4	83.9	67.0	63.2
PointPillar	87.5	76.0	72.7	58.1	51.1	46.2	82.2	61.8	57.8
PV−RCNN	89.0	80.4	77.5	66.5	59.5	53.9	87.0	69.7	65.7
CenterPoint	80.9	67.1	60.6	48.7	43.9	40.6	69.5	49.5	47.2
TED−S	92.7	83.2	78.5	55.5	49.6	43.4	86.6	68.9	64.5
本文方法	90.3	81.2	77.8	67.4	61.2	56.5	84.7	69.9	63.5

图6为本文方法在消防车无人驾驶数据集上的目标检测结果。从图6可视化结果可以看出，本文方法能够准确检测不同尺度的目标，尤其是行人这类小目标检测任务，检测框定位精确，类别识别准确。图7对比了PointPillar与本文方法的检测效果，可以发现PointPillar算法存在误检和漏检问题。PointPillar算法误检目标在图7(a)中以红框标出，其漏检目标在图7(c)中以绿框标出，局部放大情况如图7(b)所示。

对于具有明显尺度变化的小目标检测，本文引入的EMA模块发挥了重要作用。EMA模块通过并行的多卷积分支，结合跨空间学习机制，能够有效捕捉多尺度的上下文信息。在网络深层，较大的感受野有助于发现

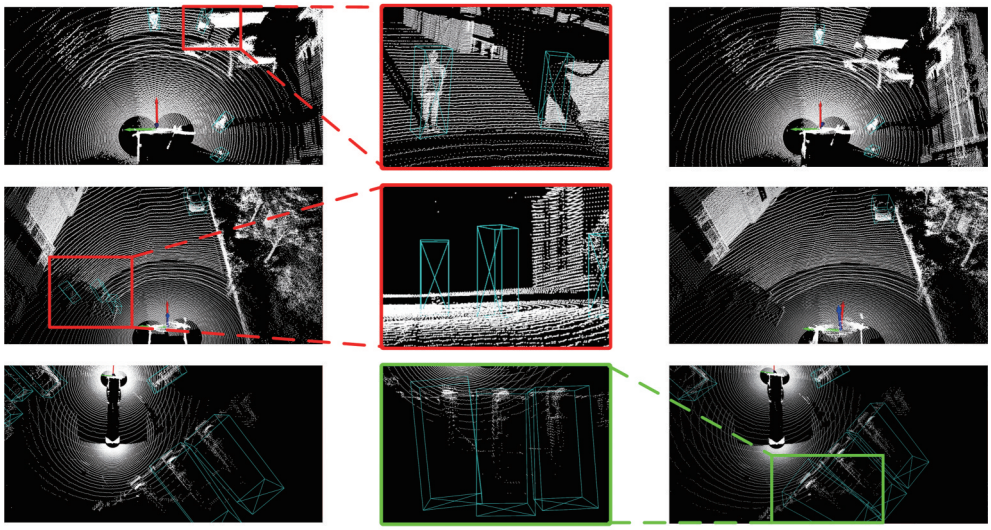


(a) 私家车与行人　(b) 私家车　(c) 行人

(d) 路障　(e) 消防车与消防机器人　(f) 远距离消防车

图 6　目标检测结果可视化

Fig. 6　Visualization of object detection results



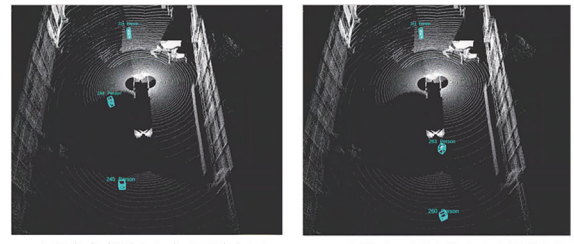
(a) PointPillar　(b) 局部放大　(c) 本文算法

图 7　PointPillar 与本文方法目标检测效果对比

Fig. 7　Comparison of object detection results between PointPillar and the proposed method

远处的微小物体，而EMA模块的多尺度融合能够保障网络在融合深层特征时，不会丢失远距离尺度变小的目标。不同尺度目标检测结果可视化如图8所示。由图8可知，行人贴近车辆时点云较为密集，远离车辆时点云稀疏，目标尺度也快速缩小，但本文方法对不同距离下不同尺度的行人均具有较好的检测效果。

实验结果表明，本文方法在消防车无人驾驶数据集上能够兼顾实时性和准确性需求，在小目标检测精度方面表现更为突出。本文方法的多尺度特征融合有效提升了小目标（如行人）的检测精度，且在不同类别目标上均取得了较好性能，推理时间接近最快的PointPillar算法，适用于算力资源受限的消防场景。



(a) 近端车侧行人与远端行人检测　(b) 近端车后行人与远端行人检测

图 8　不同尺度目标检测结果可视化

Fig. 8　Visualization of object detection results at different scales

2.5　消融实验

为评估PTV与EMA模块对三维目标检测模型性

能提升的贡献，开展了消融实验。在训练数据、实验参数均相同的情况下，通过逐一添加或移除 PTV 与 EMA 模块，检验它们对目标检测精度和推理时间的影响，结果见表 5，其中，√和×符号分别表示有无使用相应模块。由表 5 可见，基线模型 PointPillar 均不使用 PTV 和 EMA 模块，其平均精度均值 m_{AP} 为 72.0%，推理时间为 11.50 ms。基线模型仅添加 PTV 模块后，其 m_{AP} 提升至 76.8%，推理时间增加至 13.87 ms；进一步引入 EMA 模块，其 m_{AP} 提升至 78.1%，推理时间为 14.39 ms。

表 5 消融实验结果

Tab. 5 Results of ablation experiments

PTV	EMA	m_{AP}/%	推理时间/ms
×	×	72.0	11.50
√	×	76.8	13.87
√	√	78.1	14.39

实验结果表明，PTV 模块通过多头注意力机制，增强了体素特征的提取能力，使 m_{AP} 提升 4.8%；在 PTV 模块基础上加入 EMA 模块后，EMA 模块通过高效多尺度注意力机制优化特征融合，提升模型对不同尺度目标的检测能力，m_{AP} 进一步提升 1.3%。该实验验证了 PTV 与 EMA 模块对提升三维目标检测模型性能的有效性，通过 PTV 增强对体素特征的提取，EMA 进一步优化多尺度特征融合，两者共同作用使 m_{AP} 提升 6.1%。尽管本文方法推理时间相比基线模型 PointPillar 延长 2.89 ms，但在实际消防场景应用中仍能满足实时性要求。

3 结 论

1）针对消防车车载算力受限的问题，本文采用基于体素的方法，降低了目标检测任务计算复杂度，单帧推理时间为 14.39 ms，较好地满足了实时性要求。

2）针对小目标检测精度低的问题，本文通过引入注意力模块 PTV 与 EMA，优化了体素特征提取部分和主干网络特征融合拼接部分。与其他主流检测方法相比，本文方法取得了 78.1% 的最高平均精度均值，在目标检测实时性和精度指标之间实现了较优的平衡。

3）本文方法已在无人驾驶消防车上完成实车部署，能够准确、高效地检测多个目标。

参考文献：

[1] Wu Danni, Liang Zichen, Chen Guang.Deep learning for LiDAR-only and LiDAR-fusion 3D perception: A survey[J]. Intelligence & Robotics,2022,2(2):105–129.

[2] Zimmer W,Ercelik E,Zhou Xingcheng,et al.A survey of robust 3D object detection methods in point clouds[PP/OL]. arXiv(2022－03－31)[2025－06－24]. https://arxiv. org/abs/2204.00106. https://doi.org/10.48550/arXiv.2204.00106

[3] Song Ziying, Liu Lin, Jia Feiyang, et al. Robustness-aware 3D object detection in autonomous driving: A review and outlook[J]. IEEE Transactions on Intelligent Transportation Systems,2024,25(11):15407–15436.

[4] Wang Juncheng,Kong Xiangbo,Nishikawa H,et al.Dynamic point-pixel feature alignment for multimodal 3－D object detection[J]. IEEE Internet of Things Journal,2024, 11(7):11327–11340.

[5] Ren Keyan,Gu Meiying,Yuan Zhengqian,et al.3D object detection algorithms in autonomous driving: A review[J]. Control and Decision,2023,38(4):865–889.[任柯燕,谷美颖,袁正谦,等. 自动驾驶 3D 目标检测研究综述[J]. 控制与决策,2023,38(4):865–889.]

[6] Charles R Q,Hao Su,Mo Kaichun,et al.PointNet:Deep learning on point sets for 3D classification and segmentation[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Honolulu:IEEE,2017:77–85.

[7] Qi C R, Yi Li, Su Hao, et al. PointNet++: Deep hierarchical feature learning on point sets in a metric space[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems(NIPS 2017). Long Beach: Curran Associates Inc.,2017:5105–5114.

[8] Zhang Yifan,Hu Qingyong,Xu Guoquan,et al.Not all points are equal: Learning highly efficient point-based detectors for 3D LiDAR point clouds[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR).New Orleans:IEEE,2022:18931–18940.

[9] Shreelakshmi C R, Durbha S S, Singh G. 3D Object Detection in LiDAR Point Clouds using Graph Neural Networks [PP/OL]. arXiv(2023－01－29)[2025－06－24].https://doi.org/10.48550/arXiv.2301.12519.

[10] Lang A H,Vora S,Caesar H,et al.PointPillars:Fast encoders for object detection from point clouds[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR).Long Beach:IEEE,2019:12689–12697.

[11] Deng Jiajun,Shi Shaoshuai,Li Peiwei,et al.Voxel R－CNN: Towards high performance voxel-based 3D object detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence.Virtual: AAAI,2021,35(4):2074–2082.

[12] Zhu Benjin, Wang Zhe, Shi Shaoshuai, et al. ConQueR: Query contrast voxel-DETR for 3D object detection[C]//

Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Vancouver: IEEE,2023:9296–9305.

[13] Carion N,Massa F,Synnaeve G,et al.End-to-end object detection with transformers[C]//Computer Vision – ECCV 2020. Cham:Springer International Publishing,2020:213–229.

[14] Chen Yilun,Liu Shu,Shen Xiaoyong,et al.Fast point R-CNN[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision(ICCV). Seoul: IEEE, 2019:9774–9783.

[15] Shi Shaoshuai, Guo Chaoxu, Jiang Li, et al. PV – RCNN: Point-voxel feature set abstraction for 3D object detection [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle: IEEE,2020:10526–10535.

[16] Shi Shaoshuai, Jiang Li, Deng Jiajun, et al. PV – RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection[J]. International Journal of Computer Vision,2023,131(2):531–551.

[17] Tao Zhiyong, Guo Xijun, Ren Xiaokui, et al. Human action recognition method based on multi-channel fusion[J]. Advanced Engineering Sciences,2025,57(1):68–79.[陶志勇，郭希俊,任晓奎,等 . 基于多通道特征融合的人体动作识别方法[J]. 工程科学与技术,2025,57(1):68–79.]

[18] Li Lingli, Meng Lingbing, Li Jinbao.Salient object detection based on multi-scale feature extraction and multi-level feature fusion[J]. Advanced Engineering Sciences, 2021,53(1):170–177.[黎玲利,孟令兵,李金宝 . 多尺度特征提取和多级别特征融合的显著性目标检测方法[J]. 工程科学与技术,2021,53(1):170–177.]

[19] Zhao Hengshuang, Jiang Li, Jia Jiaya, et al. Point transformer[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision(ICCV). Montreal: IEEE,2021:16239–16248.

[20] K Hu J S,Kuai Tianshu,Waslander S L.Point density-aware voxels for LiDAR 3D object detection[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). New Orleans: IEEE, 2022:8459–8468.

[21] Lu Yantao, Hao Xuetao, Li Yilan, et al. Range-aware attention network for LiDAR-based 3D object detection with auxiliary point density level estimation[J]. IEEE Transactions on Vehicular Technology,2025,74(1):292–305.

[22] He Chenhang,Zeng Hui, Huang Jianqiang, et al.Structure aware single-stage 3D object detection from point cloud [C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle: IEEE,2020:11870–11879.

[23] Liu Minghui, Ma Jinming, Zheng Qiuping, et al. 3D object detection based on attention and multi-scale feature fusion [J].Sensors,2022,22(10):3935.

[24] Wang Li, Song Ziying, Zhang Xinyu, et al. SAT–GCN: Self-attention graph convolutional network-based 3D object detection for autonomous driving[J]. Knowledge-Based Systems,2023,259:110080.

[25] Leng Zhaoqi,Sun Pei,He Tong,et al.PVTransformer: Point-to-voxel transformer for scalable 3D object detection[C]// Proceedings of the 2024 IEEE International Conference on Robotics and Automation(ICRA). Yokohama:IEEE, 2024: 4238–4244.

[26] Ouyang Daliang, He Su, Zhang Guozhong, et al. Efficient multi-scale attention module with cross-spatial learning [C]//Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). Rhodes Island:IEEE,2023:1–5.

[27] Yan Yan,Mao Yuxing,Li Bo.SECOND:Sparsely embedded convolutional detection[J].Sensors,2018,18(10):3337.

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

[29] Caesar H,Bankiti V,Lang A H,et al.nuScenes:A multimodal dataset for autonomous driving[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle: IEEE, 2020:11618 – 11628.

[30] Li E,Wang Shuaijun,Li Chengyang,et al.SUSTech POINTS: A portable 3D point cloud interactive annotation platform system[C]//Proceedings of the 2020 IEEE Intelligent Vehicles Symposium(Ⅳ). Las Vegas:IEEE,2020:1108–1115.

[31] Yin Tianwei, Zhou Xingyi, Krahenbuhl P. Center-based 3D object detection and tracking[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Nashville:IEEE,2021:11779–11788.

[32] Wu Hai,Wen Chenglu,Li Wei,et al.Transformation-equivariant 3D object detection for autonomous driving[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington:AAAI Press,2023,37(3):2795–2802.

Small Object Detection Method Based on Multi-scale Fusion Using Vehicle-borne Laser-point Cloud

WANG Yintai[1,2], *CHAO Qun*[1,2*], *ZHANG Tengfei*[3], *WU Jiangdong*[1,2], *LIU Chengliang*[1,2]

(1.School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China;
2.State Key Laboratory of Mechanical System and Vibration, Shanghai Jiao Tong University, Shanghai 200240, China;
3.Intelligent Digitalization Academy, XCMG Research Institute, Xuzhou 221116, China)

Abstract:

Objective The development of unmanned fire trucks requires accurate 3D object detection in light detection and ranging (LiDAR) point clouds, particularly for small objects such as pedestrians and roadblocks, under stringent onboard computational constraints. Existing methods frequently exhibit low detection accuracy for small objects because of sparse point clouds and inadequate feature representation, whereas computationally complex models struggle to satisfy real-time processing requirements. In this study, we propose a lightweight detection method that enhances small-object perception through multiscale feature fusion, providing high detection accuracy and real-time performance suitable for unmanned fire truck applications.

Methods Raw LiDAR point clouds collected from multiple sensors mounted on an unmanned fire truck were voxelized to improve computational efficiency. The point clouds were partitioned into uniform voxels in the *x*–*y* plane, with each voxel containing up to *N* points. Each point was represented by a 10-dimensional feature vector comprising coordinates, intensity, and offsets relative to voxel centers and mean point positions. A lightweight fully connected layer projected these features into a *C*-dimensional vector, producing a tensor with the shape (***C***, *N*, *P*), where *P* represented the number of nonempty voxels.

An innovative point-to-voxel (PTV) attention mechanism was developed to aggregate point features within individual voxels. For each voxel feature tensor *L* of size $N \times C$, a global feature was extracted through max pooling and combined with a learnable query vector ***Q*** to generate a residual query (R_Q). Multihead attention was then applied using R_Q as the query and *L* as the keys and values, with the outputs scaled by a factor *d* to maintain gradient stability. This process aggregated the point features into a (***C***, ***P***) tensor, which was reshaped into a two-dimensional pseudo-image of size $C \times H \times W$.

The pseudo-image was processed using a 2D CNN backbone to extract multiscale features. Downsampling reduced the spatial resolution while increasing the channel depth, after which an efficient multiscale attention (EMA) module facilitated cross-channel and spatial feature fusion. The EMA module incorporated parallel 1 × 1 and 3 × 3 convolutional branches: the 1 × 1 branch captured channel relationships through grouping, whereas the 3 × 3 branch extracted local spatial features. Global average pooling generated channel attention, while a 1 × 1 convolution followed by sigmoid activation produced a spatial attention map. The weighted sum of the branch outputs was then added residually to the input to enhance feature representation. Deconvolution-based upsampling restored the spatial resolution, and the fused multiscale features were passed through a fully connected layer.

A Single Shot MultiBox Detector (SSD) head was employed for object classification and bounding box regression. Each 3D bounding box was parameterized using a seven-dimensional vector (x, y, z, w, h, l, θ), which represented the center coordinates, dimensions, and yaw angle. The loss function combined Smooth L1 loss for localization, softmax-based directional loss, and focal loss for classification to balance positive and negative samples. The term weights β_{loc}, β_{cls}, and β_{dir} were optimized, and positive anchors were defined using an intersection over union (IoU) threshold greater than 0.5.

A custom dataset was constructed from 2 558 frames of LiDAR data collected in fire truck operating scenarios and annotated into five classes: fire trucks, firefighting robots, cars, pedestrians, and roadblocks. The dataset was divided into training, validation, and test sets at a ratio of 7:2:1. Model training was performed using hardware equipped with an Intel Core i9–14900K CPU and an NVIDIA RTX A6000 GPU, with PyTorch 1.10, a batch size of 16, a learning rate of 0.003, and 160 training epochs.

Results and Discussions Evaluation metrics included the inference time per frame to assess real-time performance, as well as precision (P'), recall (R), average precision (A_P), and mean A_P (m_{AP}) to evaluate detection accuracy, with IoU thresholds of 0.50 and 0.35 used for true-positive and false-negative classification, respectively. On the custom dataset, the proposed method achieved a m_{AP} of 78.1%, with A_P values of 63.6% for fire trucks, 89.7% for firefighting robots, 81.2% for cars, 81.7% for pedestrians, and 74.3% for roadblocks. Small-object categories, particularly pedestrians and roadblocks, exhibited substantial improvements over the baseline methods: pedestrian A_P exceeded that of the next-best method by 14.3%, while roadblock A_P improved by 9.5%. The inference time was 14.39 ms per frame, with a model size of 60.6 MB and 5.05×10^6 parameters.

Comparative analysis against Second, CenterPoint, Point–Voxel Region-based Convolutional Neural Network (PV–RCNN), PointPillar, and the LiDAR-only variant of the transformation-equivariant 3D detector (TED–S) demonstrated superior performance in small-object detection. For pedestrians, the proposed method achieved an A_P of 81.7%, compared to 67.4% for PointPillar, while roadblock A_P reached 74.3%, compared to 64.8% for PointPillar. For larger objects, such as fire trucks, A_P was slightly lower (63.6% compared to 78.9% for Second), which was attributed

to the emphasis on intravoxel features rather than the global cross-voxel contextual information captured by more complex networks. In terms of real-time performance, the proposed method ranked second to PointPillar (11.50 ms); however, achieved higher overall detection accuracy.

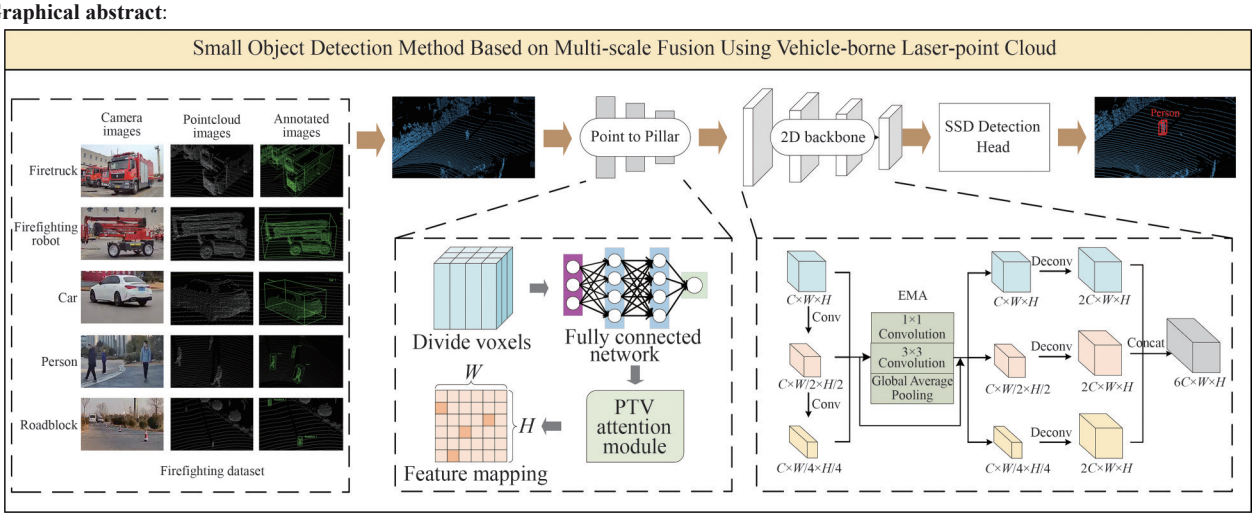
On the Karlsruhe Institute of Technology and Toyota Institute of Technology 3D Object Detection Dataset (KITTI) dataset, pedestrian A_P reached 67.4%, 61.2%, and 56.5% under easy, moderate, and hard difficulty levels, respectively. These results exceeded those of the baseline methods in small-object detection and further confirmed the generalizability of the proposed approach.

Ablation studies revealed the contributions of individual modules: the baseline PointPillar achieved m_{AP} of 72.0% at 11.50 ms. The addition of PTV attention increased m_{AP} to 76.8% (13.87 ms) by enhancing sparse point aggregation. The incorporation of EMA further improved m_{AP} to 78.1% (14.39 ms) through enhanced multiscale fusion. Visualizations confirmed accurate bounding boxes for multiscale objects, with fewer missed detections and false positives than PointPillar, particularly for distant, sparsely distributed pedestrians. The parallel branches and spatial attention of the EMA module effectively captured cross-scale dependencies and mitigated sparsity in distant objects. However, the sensitivity of 3D IoU to minor pedestrian offsets indicated the potential value of distance-based evaluation metrics for future research.

Conclusions In this study, we developed a lightweight LiDAR-based small-object detection method integrating voxelization, PTV attention for intravoxel feature aggregation, EMA for multiscale fusion, and SSD for regression. The proposed approach balanced computational efficiency and accuracy, achieving m_{AP} of 78.1% and an inference time of 14.39 ms on a custom unmanned fire truck dataset, with exceptional performance in detecting small objects. Real-world deployment confirmed its robustness in operational scenarios, demonstrating its suitability for resource-constrained autonomous systems.

Key words: autonomous driving; LiDAR point cloud; small object detection; attention mechanism; multi-scale fusion

Graphical abstract:



（编辑　陈　雪）

引用格式: Wang Yintai, Chao Qun, Zhang Tengfei, et al. Small object detection method based on multi-scale fusion using vehicle-borne laser-point cloud[J]. Advanced Engineering Sciences, 2026, 58(6): 328–339. [王寅泰, 潮群, 张腾飞, 等. 基于多尺度融合的车载激光点云小目标检测方法[J]. 工程科学与技术, 2026, 58(6): 328–339.]