计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(2):92−101 [doi: 10.15888/j.cnki.csa.009762] [CSTR: 32024.14.csa.009762]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

多模态融合的野外扬尘环境三维目标检测①

杨文浩[1,4], 况立群[1,4], 王　松[1,4], 张　珏[2,3,4]

[1](中北大学 计算机科学与技术学院, 太原 030051)
[2](智能采矿装备技术全国重点实验室, 太原 030032)
[3](山西太重智能采矿装备技术有限公司, 太原 030032)
[4](机器视觉与虚拟现实山西省重点实验室, 太原 030051)
通信作者: 况立群, E-mail: kuang@nuc.edu.cn

摘　要: 对于配备多种传感器的自动驾驶车辆, 在野外扬尘环境中进行高精度三维目标检测是一项重大挑战, 野外地形的多变性导致采集目标的区域特征差异性加剧, 同时扬尘颗粒物还会模糊目标特征. 为了克服这些困难, 本文提出了多模态特征动态融合的三维目标检测方法, 构建了多级特征自适应融合模块和特征对齐增强模块, 其中, 多级特征自适应融合模块动态调整模型对全局级特征和区域级特征的关注程度, 充分利用多级感受野, 减少区域差异对识别效果的影响; 而特征对齐增强模块则在多模态特征对齐之前增强感兴趣区域的特征表达, 有效抑制扬尘等干扰因素. 实验结果表明, 提出方法在自建野外数据集中比基线的平均精度提高了 2.79%, 在 KITTI 数据集的困难级别检测中提高了 1.7%, 表现出较好的鲁棒性和准确性.

关键词: 三维目标检测; 野外; 扬尘; 多模态融合; 点云

引用格式: 杨文浩,况立群,王松,张珏.多模态融合的野外扬尘环境三维目标检测.计算机系统应用,2025,34(2):92–101. http://www.c-s-a.org.cn/1003-3254/9762.html

Multi-modal Fusion for 3D Object Detection in Dusty Wilderness

YANG Wen-Hao[1,4], KUANG Li-Qun[1,4], WANG Song[1,4], ZHANG Jue[2,3,4]

[1](School of Computer Science and Technology, North University of China, Taiyuan 030051, China)
[2](National Key Laboratory of Intelligent Mining Equipment and Technology, Taiyuan 030032, China)
[3](Shanxi Taizhong Intelligent Mining Equipment and Technology Co. Ltd., Taiyuan 030032, China)
[4](Shanxi Key Laboratory of Machine Vision and Virtual Reality, Taiyuan 030051, China)

Abstract: It is a significant challenge for high-precision 3D object detection for autonomous vehicles equipped with multiple sensors in the dusty wilderness. The variable wilderness terrain aggravates the regional feature differences of detected objects. Additionally, dust particles can blur the object features. To address these issues, this study proposes a 3D object detection method based on multi-modal feature dynamic fusion and constructs a multi-level feature self-adaptive fusion module and a feature alignment augmentation module. The former module dynamically adjusts the model's attention to global-level features and regional-level features, leveraging multi-level receptive fields to reduce the impact of regional variances on recognition performance. The latter module bolsters the feature representation of regions of interest before multi-modal feature alignment, effectively suppressing interference factors such as dust. Experimental results show that compared with the average precision of the baseline, that of this approach is improved by 2.79% in the self-built wilderness dataset and by 1.7% in the hard-level test of the KITTI dataset. This shows our method has good robustness and precision.

Key words: 3D object detection; wilderness; dust; multi-modal fusion; point cloud

① 基金项目: 山西省科技重大专项计划"揭榜挂帅"项目 (202201150401021); 山西省基础研究计划 (TZLH20230818005, 202303021211153, 20220302122027); 山西省科技成果转化引导专项 (202104021301055); 山西省研究生科研创新项目 (2023KY613)
收稿时间: 2024-07-04; 修改时间: 2024-08-01; 采用时间: 2024-08-27; csa 在线出版时间: 2024-12-13
CNKI 网络首发时间: 2024-12-13

随着新能源车的飞速进步, 自动驾驶技术被广泛视为未来汽车产业的主要发展方向. 三维目标检测作为自动驾驶技术中的关键任务, 在学术界和工业界均备受关注. 然而, 当前的三维目标检测应用主要集中于城市道路[1–4], 对于复杂环境, 比如野外扬尘环境下的应用, 尚未得到充分的研究和考虑, 阻碍了自动驾驶朝着全面驾驶自动化发展.

复杂环境数据通常很难收集, 一些研究为了进行相关的三维目标检测研究, 常在现有城市数据集下进行复杂环境模拟或仿真[5,6]. 这种研究方法虽然具有一定的价值, 但它未能充分考虑到在实际环境中, 不同传感器在数据采集过程中可能遇到的诸多问题. 比如在扬尘环境中, 激光雷达传感器发射的光脉冲, 一旦遇到空气中的尘埃颗粒, 会发生反向散射和衰减, 使得激光雷达获得的点云数据会包含一些虚假返回[7]. 这种失真的数据会对模型性能造成不可预估的损失. 并且在模拟或仿真数据集上训练的算法往往也不能直接应用在真实环境中, 二者数据集分布存在差异.

为了对真实的扬尘环境下的三维目标检测进行研究, 本文团队在野外实地条件下, 采集了丰富的图像和点云数据, 其中扬尘环境是指在野外环境下车辆行进过程中, 由于地面的尘土被风吹动而升起飞扬的尘埃颗粒形成的模糊环境. 与在城市数据集上进行的模拟或仿真环境不同, 真实的野外环境中, 行进中的目标车辆往往面临颠簸的地形变化, 二者在激光雷达中目标成像的位置在高度方向有不同分布. 图 1 为数据集中目标车辆三维包围框底部坐标在垂直于地面方向 (即 Z 轴坐标) 的分布直方图, 数据分别来自 (a) KITTI 城市数据集和 (b) 自建野外扬尘数据集. 从图 1 中可以看到, 城市环境中目标包围框的高度方向分布较集中, 而野外环境中的分布则比较随机且分散. 这种区域之间的分布差异会加大模型提取有效特征的难度. 此外, 扬尘颗粒可能会影响相机的成像质量, 使得从拍摄的图像中获取到的目标特征变得模糊. 这些问题都对野外扬尘环境下的高精度的三维目标检测带来了挑战.

现有的三维目标检测方法分为单模态方法[8,9]和多模态方法[10,11]. 单模态方法从单个传感器获取数据并进行特征提取, 而多模态方法则从多个异构传感器获取数据并对不同模态的数据进行融合学习[12,13]. 野外车辆行驶过程中, 扬尘环境对各种传感器有着不同的影响, 单个传感器无法满足安全驾驶的需要, 利用从不同传感器获取的多模态数据进行融合有利于提升三维目标检测算法的精度. Yeong 等人[14]指出, 在复杂环境中, 多模态融合技术在整体感知性能上显著优于单独使用激光雷达或相机等单一传感器模态.

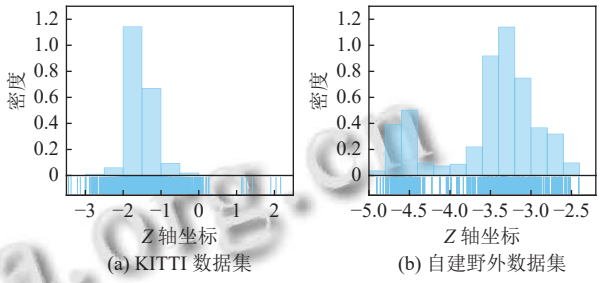


图 1 数据集目标高度分布对比

当前高精度多模态融合大致遵循两种方法: (1) 早期阶段融合, 使用相机的像素特征装饰激光雷达点云中的点, 以增加点的语义表达[15,16]. (2) 在模型管线中期进行融合, 即在早期进行特征提取后对各模态特征进行组合[17,18]. 然而, 早期融合由于引入了语义分割网络, 一般都会给模型带来不可忽略的计算开销, 此外, 对于扬尘环境, 尘埃形成的无效点也会直接加入到模型训练当中, 这也将带来额外的计算成本. 文献[19]对大雾天气条件下中期深度融合技术的效果进行了研究, 结果表明该融合方法具有较好的鲁棒性. 中期融合可以产生高质量的三维框, 使得模型能够对目标进行更好地识别.

为了平衡效率和精度, 当前高精度的多模态融合方法大都采用中期融合技术. 一些工作[20,21]基于点到像素的对应关系, 通过不同的融合算子将来自激光雷达主干的特征与来自图像主干的特征进行融合. 除了在主干网络中融合激光雷达特征和图像特征, 还有一类算法在候选框生成和 ROI (region of interest) 细化阶段进行融合[22,23]. 这类方法一般先从激光雷达检测器生成三维目标候选框, 然后将三维候选框投影到图像视图或鸟瞰图, 从视图中裁剪特征, 再将裁剪后的特征与三维候选框进行融合和进一步细化学习. 此外, 还有研究者[24]将前述两个阶段的处理流程结合在一起, 形成多级特征 (全局特征和区域特征), 以期汲取两者的优势.

鉴于野外扬尘环境对三维目标检测所带来的挑战, 模型应当不只是聚焦于全局信息, 更应细致地捕捉各个区域间的细微差异. 简单地将多级特征相加合并, 无

法充分利用多级特征的感受野, 从而很难满足野外扬尘环境下三维目标检测的准确性和鲁棒性要求. 为此, 本文提出了一个面向野外扬尘环境的三维目标检测方法, 该方法能够有效捕捉互补的多模态特征, 自适应地融合多级特征. 本文方法主要包括多级特征自适应融合和特征对齐增强两个核心内容. 其中, 多级特征自适应融合可以自适应地融合模型管线中不同阶段的多模态特征, 充分利用多级感受野, 更好地学习并适应区域之间的差异, 减缓在野外中目标高度方向分布多样性造成的识别影响; 特征对齐增强是一个即插即用的模块, 帮助模型在点云特征与图像特征对齐时增强感兴趣区域目标的特征表达, 从而有效降低扬尘等干扰因素对成像质量造成的不利影响.

1 模型架构

本文提出的模型架构如图 2 所示, 分为 3 个部分, 激光雷达管线、相机管线和结果融合与预测单元. (1) 激光雷达管线接收激光雷达点云数据的输入, 经过体素化后进入激光雷达主干网络, 激光雷达主干网络对体素点云进行特征提取, 得到点云全局特征. 全局特征经过 RPN (region proposal network) 网络生成 ROI 特征. (2) 相机管线获取相机图像, 通过相机主干网络对图像信息进行特征提取, 同时利用特征对齐增强模块动态聚合关键特征. 接着, 将图像特征分别与点云全局特征和 ROI 特征进行融合, 得到全局级特征 F_o 和区域级特征 F_r. (3) 使用多级特征融合模块自适应地将全局级特征 F_o 和区域级特征 F_r 融合起来, 得到鲁棒且有效的表征, 最后经过检测头进行分类与回归, 得到最终的预测结果. 本文网络结构的创新点体现在以下方面.

(1) 为了应对野外道路环境出现的区域特征差异的问题, 提出多级特征自适应融合模块, 与直接相加等简单的融合方式不同, 该模块通过预测全局级特征和区域级特征在融合过程中的权重, 使模型可以自适应调整不同级别特征在融合过程中的重要性, 有效适应区域之间的差异, 减缓目标高度方向分布多样性造成的识别影响.

(2) 针对野外扬尘模糊图像目标特征的问题, 进一步细化了相机主干网络, 在常规图像处理流程中嵌入了特征对齐增强模块, 该模块得益于卷积网络注意力的特性, 帮助模型在异构模态特征对齐时增强感兴趣区域目标的特征表达, 降低扬尘等干扰因素对成像质量造成的不利影响.

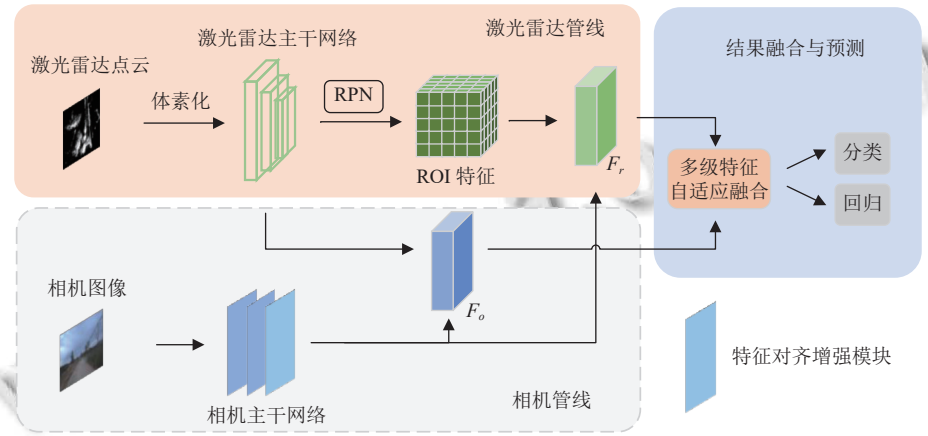


图 2 模型总体网络架构图

1.1 多级特征自适应融合

野外环境复杂, 道路状况千差万别, 坑洼不平, 这给激光雷达成像带来了不小的挑战. 激光雷达的信号可能会因为地面不平而产生跳跃. 如图 1 所示, 目标车辆的位置在高度方向的分布情况, 数据分别来自在城市中收集的 KITTI 数据集和自建野外扬尘数据集. 显然, KITTI 数据集中的分布比较集中, 而自建数据集中的分布比较分散且随机. 这是因为在城市中, 由于道路相对平坦, 车辆在激光雷达中的成像基本一致, 而在野外环境中, 车辆可能会经历更多的障碍和地形变化, 目标车辆位置在高度方向的分布更加多样化, 这种差异会导致模型在处理野外数据时出现偏差. 因此, 为了提高野外环境的目标识别能力, 模型不仅需要关注全局信息, 还要注意区域之间的差异.

针对此问题, 本文设计了多级特征自适应融合模块, 如图 3 所示. 该模块可以自适应地融合全局级和区域级的多模态特征, 充分利用多级特征的感受野, 使模型能够动态地调整其对不同级别特征的关注程度, 从而减缓目标的高度分布多样化造成的识别困难.

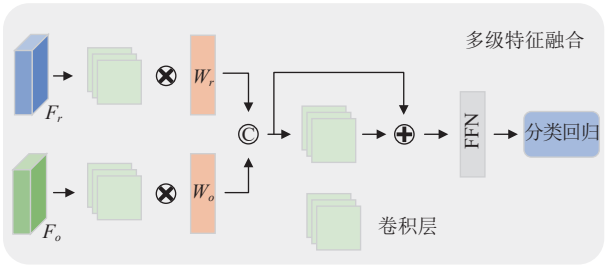


图 3 多级特征自适应融合

具体而言, F_o 和 F_r 分别表示已经融合点云和图像信息的全局级特征和区域级特征, 将 F_o 和 F_r 分别经过卷积层进行特征编码获得特征图 F_o^c 和 F_r^c, 计算公式如下:

$$F_o^c = S(BN(C_o(F_o))) \tag{1}$$

$$F_r^c = S(BN(C_r(F_r))) \tag{2}$$

其中, C_o 和 C_r 为卷积算子; BN (batch normalization) 为批归一化, 使特征图更加稳定; S 为非线性激活层, 使用的 $Sigmoid$ 函数.

为了自适应调整模型对不同级别特征的重视程度, 引入加权方法对多级特征进行驱动[25,26]. 即为全局级特征 F_o 和区域级特征 F_r 分别预测一个权重, 并用预测的权重对全局级特征 F_o 和区域级特征 F_r 进行加权. 通过这种加权方式, 模型可以在训练过程中自主学习到不同级别特征的重要性. 权重预测方法如图 4 所示.

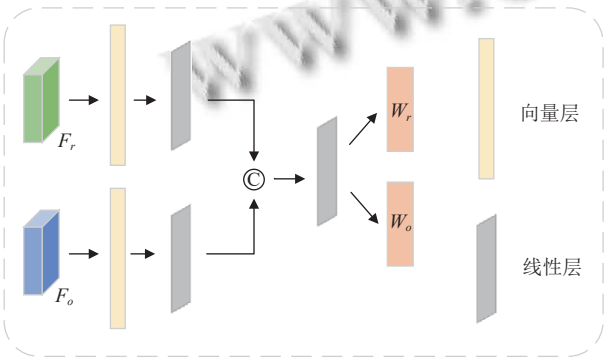


图 4 权重预测方法

进行权重预测之前先对输入特征进行预处理. 由于权重计算无需考虑结构信息, 对于特征 F_o 和 F_r, 直接通过展平操作 $Flatten(\cdot)$ 将多维特征转换为一维向量, 接着分别通过线性层 L_o 和 L_r 进行特征编码, 得到特征 F_o^w 和 F_r^w, 这样可以确保数据在展平后仍然保持在合适的分布范围, 操作方法如下:

$$F_o^w = L_o(Flatten(F_o)) \tag{3}$$

$$F_r^w = L_r(Flatten(F_r)) \tag{4}$$

然后使用 $CAT(\cdot)$ 操作将 F_o^w 和 F_r^w 连接起来, 并馈送入线性层, 得到最终的权重 W_o 和 W_r, 具体公式如下:

$$(W_o, W_r) = S(Linear(CAT(F_o^w, F_r^w))) \tag{5}$$

接着使用预测的权重 W_o 和 W_r 对编码后的全局级特征 F_o^c 和区域级特征 F_r^c 分别进行加权, 见图 3, 并将加权后的特征进行连接, 使用残差连接避免后期可能发生的梯度消失. 经过前馈网络 FFN (feed-forward network) 后作最后的预测.

1.2 特征对齐增强

野外环境中, 受道路与天气影响, 车辆行驶过程中常会产生扬尘现象, 这些空气中的颗粒物会降低环境的能见度, 模糊目标的轮廓和特征, 最终影响模型对目标的识别效果. 利用多模态融合的方法, 将激光雷达的数据与摄像头等其他传感器的数据进行结合, 可以提高对目标车辆的识别准确性和稳定性, 而多模态融合的前提是有效地对齐异构模态的特征. 因此, 扬尘环境中要求模型在特征对齐阶段将注意力更加聚焦于目标本身, 以便更好地捕捉目标的特征.

本文借鉴一个简单有效的卷积神经网络注意力[27], 设计了一个即插即用的特征对齐增强模块, 该模块可以在异构模态特征对齐阶段, 增强对感兴趣区域的注意程度, 使模型能够更准确理解和区分关键特征和无效背景信息, 从而提高对目标的识别能力.

特征对齐增强模块嵌入位置见图 2, 模块嵌入在相机管线主干网络中. 一般而言, 对于输入的图像信息, 通常都由成熟的图像处理算法进行特征提取, 然而在多模态特征融合架构中, 需要有针对性的网络设计来满足异构模态特征有效对齐的要求.

具体来说, 目前的多模态融合架构中基本分为激光雷达管线和相机管线, 在各管线中分别得到点云和图像特征后, 再通过一些设计的算子将异构特征进行关联、对齐和融合. 但这类架构设计基本聚焦在激光雷达管线, 为了更好地提取点云特有的三维结构特征,

激光雷达管线通常包括点云体素化、三维特征提取、锚框生成和 ROI 特征细化等多个阶段. 而相机管线设计则相对简易, 一般只配置一些经典的图像算法, 在处理扬尘颗粒污染过的图像时, 模型对感兴趣区域关注不足, 导致在关联和对齐阶段没有充分利用目标图像丰富的颜色和纹理信息.

而在人类学习和感知过程中, 即使周围环境存在许多干扰信息, 也往往能够专注于他们感兴趣的目标. 这是因为在视觉神经系统中, 那些与感兴趣目标相关的区域内的神经元表现出更高的活跃性, 这些神经元所携带的有效信息更加丰富, 而信息最丰富的神经元通常又会被更多地关注, 或者说, 这些神经元又会在后续视觉处理中被赋予更高的优先级[28].

对于图像信息, 感兴趣区域的特征往往有着目标详尽的语义和纹理信息. 因此, 感兴趣区域附近的“神经元”, 即神经网络节点, 应该更加活跃, 而那些由于扬尘现象在成像中造成的大量无效信息, 它们的表达应该被尽量抑制. 本文提出的特征对齐增强模块引入了一个以能量函数为内核的注意力机制[27], 该注意力机制可以计算出每个“神经元”的影响力, 从而加强关键特征在整个处理过程中的重要性. 能量函数如下:

$$ef(n,x_i)=\frac{1}{Su-1}\sum_{i=1}^{Su-1}(p-x_i)+(a-n)^2 \tag{6}$$

其中, n 和 x_i 分别代表目标神经元和其他神经元, i 是空间维度索引值, Su 则是由某一空间维度中的所有神经元数量. a 和 p 是两个不同的标量值, 设为 1 和−1, 代表活跃最值和抑制最值. 通过将能量函数 ef 最小化, 目标神经元将尽可能接近活跃最值, 其他神经元会趋向抑制最值.

受益于该注意力机制的强大特性, 本文通过将特征对齐增强模块嵌入到相机主干网络中, 模型能够在多模态特征对齐之前的图像特征提取阶段, 实现对感兴趣区域的特征的强化, 同时抑制无效背景的过度表达. 在消融实验中进一步验证了这一机制的有效性. 模块具体算法实现见算法 1.

算法 1. 特征对齐增强算法

//输入 x 和输出 y 特征维度一致, 皆为 4D 特征[B, C, H, W], 分别为批量大小、通道数、高和宽.

(1) FeatureAlignEnhance(x):
(2) x = Conv(x)
(3) B, C, H, W=x.*size*()
(4) S=H×W − 1
(5) d=(x–x.*mean*(*dim*=(2, 3))).*pow*(2)
(6) e=d/(4×(*d.sum*(*dim*=(2, 3))/S+0.1))+0.5
(7) y = *ReLU*(*BN*(x×*Sigmoid*(e)))
(8) return y

1.3 损失函数

相对于单阶段目标检测方法, 双阶段目标检测方法能够更准确地定位和分类目标, 在面对复杂环境、目标特征模糊和尺度变化等问题时表现更为鲁棒. 针对前述野外扬尘环境下出现的问题, 本文采用双阶段方法[29]检测目标. 模型整体的损失函数由第 1 阶段损失函数和第 2 阶段损失函数组成, 计算公式如下:

$$\mathcal{L}=\mathcal{L}_1+\mathcal{L}_2 \tag{7}$$

其中, $\mathcal{L}_1$ 和 $\mathcal{L}_2$ 分别为第 1 阶段损失函数和第 2 阶段损失函数. 第 1 阶段生成目标建议, 并在第 2 阶段进行细化, 所以 $\mathcal{L}_1$ 采用 RPN 损失函数. 第 2 阶段损失 $\mathcal{L}_2$ 由 3 部分组成, 分别是分类损失 $\mathcal{L}_{\text{bCross}}$, 回归损失 $\mathcal{L}_{\text{wSmL1}}$, 即:

$$\mathcal{L}_2=\alpha\mathcal{L}_{\text{bCross}}+\beta\mathcal{L}_{\text{wSmL1}} \tag{8}$$

其中, α 和 β 为平衡不同损失的超参数. 本文自建数据集只有汽车一个类别, 分类损失 $\mathcal{L}_{\text{bCross}}$ 采用二元交叉熵函数, 回归损失 $\mathcal{L}_{\text{wSmL1}}$ 使用 SmoothL1 函数.

2 实验结果及分析

本文在自建的野外扬尘数据集上对提出的模型性能进行了评估, 此外, 为了验证模型的泛化能力, 也在流行的三维目标检测数据集 KITTI 上进行了实验, 与当前性能良好的一些方法做了对比, 最后对提出的方法进行了消融实验和可视化分析, 进一步验证提出方法的有效性.

2.1 数据集与评价指标

在真实的野外自然环境中, 由于路面车辆运动或风力作用会使土壤颗粒悬浮在空气中, 从而产生扬尘现象. 本文的实验数据在这种野外扬尘的环境下, 并且有着良好光照条件的情况下, 通过车载平台采集.

采集平台上安装的传感器包括海康威视 MV-CA020-20GC 可见光相机一台, 禾赛 64 线激光雷达一台, 分别用来采集 RGB 图像数据和点云数据. 可见光相机分辨率为 1920×1200, 采集正前方 RGB 图像, 激光雷达采集 360°点云数据, 点云数据进行切割预处理后保留采集平台正前方 0–180°数据, 得到自然光照充

足的 7350 组 (点云/RGB 图像) 野外扬尘数据.

之后分别对图像和点云数据进行标注, 为了便于对原始数据进行处理, 遵循流行的三维目标检测实践, 将标注数据按照 KITTI 数据集的标签格式进行转换和整理, 得到可用于训练、验证和测试的自建真实野外扬尘数据集. 其中训练集、验证集、测试集按 7:2:1 进行划分. 由于在实际的野外扬尘环境中, 行人极少, 因此构建的扬尘数据集目前仅包含“汽车”这一类别. 在后续的工作中, 考虑增加其他类别, 以丰富数据集的多样性. 自建数据集不同模态分布情况如表 1 所示.

表 1 自建野外数据集分布

类别	点云图数量	RGB图像数量	目标数量
汽车	7350	7350	8190

自建数据集的评价指标参考了 KITTI 数据集的官方指标. 为了更严格地评估模型性能, 选择计算三维检测框的准确率, 而非鸟瞰视图 BEV (bird's eye view) 下的检测框准确率. 采用平均精度 (average precision, *AP*) 作为评价指标, 用来衡量模型识别汽车类别的准确性, 计算 *AP* 时, 将汽车类别的交互比 (intersection over union, IoU) 设为 0.7. 与最新的 KITTI 评价指标一致, 抛弃旧的通过 11 个召回位置计算 *AP*, 本文采用 40 个召回位置来修正指标, 最终得到更加科学、可以更好反映模型能力的 *AP* 值. *AP* 计算公式为:

$$AP_{RN} = \frac{1}{N}\sum_{r\in RN}\rho_{\text{interp}}(r) \tag{9}$$

其中, *RN* 即召回率, 40 个召回位置即为 *R*40, $R40 = \{1/40, 2/40, 3/40, \cdots, 1\}$. 内插函数 ρ_{interp} 定义为 $\rho_{\text{interp}}(r) = \max_{r':r'\geqslant r}\rho(r')$, $\rho(r)$ 为召回 r 的精度.

2.2 数据集与评价指标

本实验基于一台配备两块 NVIDIA GeForce RTX 3090 GPU 以及 Intel(R) Xeon(R) Gold 5218R CPU 的服务器进行训练与测试, 服务器操作系统为 Ubuntu 18.04.6 LTS. 深度学习实验平台基于 Cuda 11.1.1、Python 3.7.12 搭建, 使用 Cudnn 8 进行运算加速, 采用的深度学习框架为 PyTorch 1.10.0. 使用 OpenCV-Python 和 Open3d 0.17.0 分别对图像和点云进行可视化.

对于自建野外数据集, x、y、z 轴上的检测范围分别为[0, 56] m, [−6.4, 0.8] m, [−4 , 0] m. 为了节省计算成本, 实验中输入的 RGB 图像数据是经过将原相机数据处理后得到的更便于模型学习的分辨率为 960×600 的 RGB 图片. 考虑到点云数据处理中的效率和速率需要平衡, 模型中的点云分支使用的主干网络是 Voxel R-CNN[23], 本实验将输入的点云数据进行体素化, 单个体素的尺寸设置为 (0.05, 0.05, 0.1) m, 并且每个体素中包含点的数量最多为 5 个, 训练时体素数量最多为 16000 个, 测试时不超过 40000 个.

网络使用 Adam 优化器进行端到端训练, 训练的 epoch 为 60 个, 单个 GPU 的 batch_size 设置为 2, 优化器的初始学习率设置为 0.01, 衰减策略采用 onecycle, 权重衰减系数为 0.01, 动量因子为 0.9. 后处理阶段采用 NMS (非极大值抑制), 阈值设置为 0.55.

2.3 实验结果与性能对比

自建野外数据集. 表 2 显示了模型在自建野外数据集上的结果 (模态中 L 指点云, C 指图像). 自建数据集处理过的类别目前只有汽车一类, 为了更好地评估模型性能, 采用主流的可以更有效反应算法检测性能的评价指标, 见第 2.1 节评价指标介绍. 提出的方法在自建数据集上的 *AP* 值达到 76.32%, 与之前的方法对比 (由于本数据集缺乏路面信息, 对比方法中基于路面的增强方法都没有配置), 不仅优于只使用激光雷达单模态的方法, 也超越了多模态方法. 其中, 相较于基准模型 LoGoNet, 本文方法提升 2.79%. 鉴于野外环境的识别特性, 提出的方法对于野外扬尘环境下的三维目标检测任务具有一定的适用性.

表 2 自建野外数据集上不同方法检测结果对比

方法	年份	模态	汽车*AP* (%)
SECOND[30]	2018	L	74.12
PV-RCNN[31]	2020	L	74.74
Voxel-R-CNN[29]	2021	L	72.85
CasA[32]	2022	L	72.31
FocalsConv[10]	2022	L+C	75.20
LoGoNet[24]	2023	L+C	74.25
Ours	2024	L+C	**76.32**

KITTI 数据集. 本文还在 KITTI 公共数据集上进行了对比实验, 见表 3 所示 (模态中 L 指点云, C 指图像). 为了评价一致与公平, 在与其他主流模型作对比时, 只计算汽车一类的指标. 与对比模型的评价指标一致, 汽车交互比 IoU 设为 0.7, 采用 40 个召回位置计算 *AP* 值. 从表 3 中可以看出, 提出的模型在汽车类别上所有难度级别的 *AP* 值相比于基准模型都有所提升, 特别是困难这一级别, *AP* 提高 1.7%. 表明本方法较好的鲁棒性.

表 3 KITTI 数据集上不同方法检测结果对比

方法	年份	模态	汽车*AP* (%)		
			容易	中等	困难
SECOND[30]	2018	L	88.61	78.62	77.22
PV-RCNN[31]	2020	L	92.10	84.36	82.48
Voxel-R-CNN[29]	2021	L	92.38	85.29	82.86
CAT-Det[21]	2022	L+C	90.12	81.46	79.15
FocalsConv[10]	2022	L+C	92.26	85.32	82.95
LoGoNet[24]	2023	L+C	92.04	85.04	84.31
Ours	2024	L+C	**92.40**	**85.37**	**85.74**

自动驾驶矿山场景数据集[33]. 该数据集采集自矿山实际作业场景, 其环境特点是道路崎岖、灰尘多. 见图 5 所示, 分别是不同浓度下的矿山场景. 采用数据集中的 3D 平均精度指标来评估模型在矿山数据集的性能表现. 实验设置中, 点云在 x、y、z 轴上的检测范围分别是[−2.8, 78.8] m, [−45.8, 55] m, [−4, 11] m, 体素尺寸为 (0.05, 0.06, 0.3) m, RGB 图像分辨率为 1536×2048. 实验结果见表 4 所示, 提出方法在各类别的检测精度均优于其他方法.

(a) 扬尘较弱 (b) 扬尘较强

图 5 自动驾驶矿山场景

表 4 矿山场景数据集上不同方法检测结果对比

方法	年份	模态	Truck@3D-*AP* (%)
SECOND[30]	2018	L	40.27
PV-RCNN[31]	2020	L	54.59
Voxel-R-CNN[29]	2021	L	55.37
CasA[32]	2022	L	57.42
FocalsConv[10]	2022	L+C	57.93
LoGoNet[24]	2023	L+C	58.22
Ours	2024	L+C	**59.31**

2.4 消融实验

本文消融实验主要在自建野外数据集上进行, 以验证提出的方法中每个模块对最终性能的影响. 本文方法主要有两个模块, 即多级特征自适应融合模块和特征对齐增强模块. 各组件的作用如表 5 所示.

由表 5 可知, 对于自建数据集上的评价指标, 多级特征自适应融合模块带来 2.3% 的性能增益. 野外环境中目标车辆受地面不平影响, 在三维点云成像中位置分布差异明显. 多级特征自适应融合模块自适应地改变各管线关键特征的权重, 融合区域级特征和全局级特征后, 模型形成鲁棒的表征, 这将进一步加强对上下文的理解, 从而减缓区域特征差异对目标识别造成的不利影响. 添加特征对齐增强模块后, 汽车 *AP* 值进一步提升 0.48%. 野外汽车行进过程中与地面作用形成的扬尘一般聚集在尾部区域, 因此扬尘点往往会模糊目标特征. 特征对齐增强模块可以动态聚合关键的感兴趣区域特征, 在异构模态对齐之前为多模态融合管线中的各级特征提供关键、丰富的语义和颜色信息, 为后续流程生成更好的多模态特征, 缓解扬尘的干扰.

表 5 自建野外数据集上消融实验结果对比

多级特征自适应融合	特征对齐增强	汽车*AP* (%)
×	×	74.25
√	×	75.96
√	√	76.32

此外, 基于特征对齐增强是一个即插即用的模块, 考虑到多模态特征对齐时, 加强点云数据中感兴趣区域特征同样至关重要, 本文将该模块嵌入到了点云 BEV 网络中, 如表 6 所示. 与嵌入在相机主干网络相比, 效果反而变差. 分析可能是点云 BEV 特征虽然和图像特征的维度一致, 但点云 BEV 特征是由原始的三维点云压缩而来, 隐含着点云特有的三维结构信息, 使得特征对齐增强模块可能无法捕捉到这种隐式的三维特征, 造成检测精度下降.

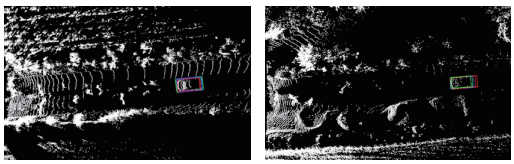
表 6 特征对齐增强模块不同嵌入位置结果对比

点云BEV网络	相机主干网络	汽车*AP* (%)
√	×	72.13
×	√	74.79

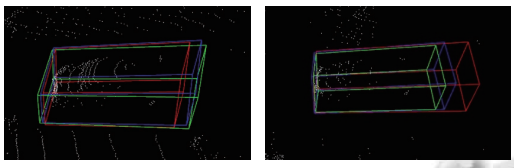
2.5 可视化分析

可视化分析见图 6, 展示了基线方法和本文提出的方法在自建数据集上的三维检测结果对比, 从左到右分别是两组野外不同道路环境和不同扬尘密度下的场景, 从上到下依次是 (a) 图像场景 (RGB 图片)、(b) 对应的点云场景 (点云俯瞰图) 和 (c) 对应的点云场景中的车辆目 (目标三维框近视图). RGB 图片直观地呈现了野外扬尘环境的真实情况, 点云图中红色为目标三维框真值, 绿色和蓝色分别为基线方法和本文方法的检测结果. 可以观察到提出的方法在野外不同场景中, 三维检测框都更接近于真值, 验证了本文提出方法的有效性.

(a) 图像场景



(b) 对应的点云场景



(c) 对应的点云场景中的车辆目标

图 6 自建数据集可视化检测结果对比

2.6 目标高度分布变化与实验结果

基于锚框的三维目标检测方法在模型训练时, 首先在激光雷达点云中生成一批固定位置和大小的锚框, 然后再进行进一步的训练和预测.

野外路面由于地形变化会造成采集车辆和目标车辆不在同一水平面上, 这种高度差异在激光雷达成像中会导致目标位置的显著变化, 从而增加模型学习的难度. 本文提出的多级特征自适应融合模块能够动态调整对区域差异的注意程度, 提高模型对高度变化的适应能力. 见图 7 所示, 其中, LoGoNet 为基准模型, Ours 为本文方法. 在不同高度分布上提出方法表现得比基准模型更好.

2.7 扬尘现象影响分析与实验结果

野外环境中, 车辆行驶时与不平整路面作用常产生不同浓度的扬尘, 本文以在相机成像中, 扬尘在空气中不同浓度大小来验证扬尘现象对三维目标检测的影响程度. 具体见图 8 所示, 分别为较弱、中等和较强, 图片依照扬尘浓度由小到大排列. 图中绿色线条为扬尘与可见目标之间的边界, 以表明目标在扬尘环境中的模糊程度. 依据上述定义, 将自建数据集分别划分为 3 个子集, 随后, 将模型分别在各个子集上进行实验验证, 并记录下实验结果, 见图 9 所示, 其中, LoGoNet 为基准模型, Ours 为本文所提出的方法. 可以看到, 本文提出方法在各个级别上的检测精度都优于基准模型.

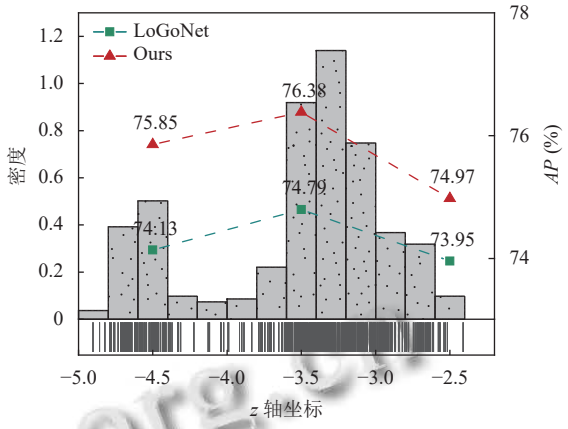


图 7 目标高度不同分布检测结果对比

较弱 中等 较强

图 8 野外扬尘不同浓度展示

3 结束语

通过引入自建数据集, 本文分析和总结了自动驾驶车辆在野外扬尘环境中进行目标识别和定位时存在的区域差异和特征模糊等问题. 针对这些问题, 本文提出了多模态特征动态融合的三维目标检测方法. 该方法可以充分利用互补的激光雷达和图像信息, 动态融合目标关键特征. 在自建数据集上的实验表明, 提出的方法能有效提高了三维目标检测精度, 并在 KITTI 数据集上进行对比实验, 显示出方法对高难度目标的识别能力. 由于野外扬尘环境中目标稀少, 且数据集收集困难, 本文自建数据集只有汽车类别. 未来将进一步解决目标类别单一的问题, 使研究更加深入和全面.

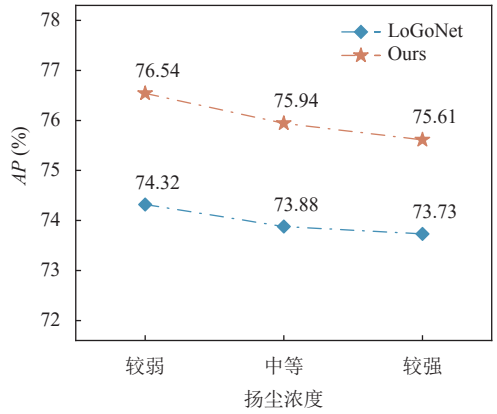


图 9 扬尘不同浓度实验结果对比

参考文献

1 Mao JG, Shi SS, Wang XG, *et al*. 3D object detection for autonomous driving: A comprehensive survey. International Journal of Computer Vision, 2023, 131(8): 1909–1963. [doi: 10.1007/s11263-023-01790-1]

2 葛同澳, 李辉, 郭颖, 等. 基于双融合框架的多模态 3D 目标检测算法. 电子学报, 2023, 51(11): 3100–3110. [doi: 10.12263/DZXB.20230414]

3 霍威乐, 荆涛, 任爽. 面向自动驾驶的三维目标检测综述. 计算机科学, 2023, 50(7): 107–118. [doi: 10.11896/jsjkx.220700090]

4 Guo JY, Kurup U, Shah M. Is it safe to drive? An overview of factors, metrics, and datasets for driveability assessment in autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(8): 3135–3151. [doi: 10.1109/TITS.2019.2926042]

5 Hahner M, Sakaridis C, Bijelic M, *et al*. LiDAR snowfall simulation for robust 3D object detection. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 16343–16353.

6 Hahner M, Sakaridis C, Dai DX, *et al*. Fog simulation on real LiDAR point clouds for 3D object detection in adverse weather. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 15263–15272.

7 Vattem T, Sebastian G, Lukic L. Rethinking LiDAR object detection in adverse weather conditions. Proceedings of the 2022 International Conference on Robotics and Automation. Philadelphia: IEEE, 2022. 5093–5099.

8 陈易男. 自动驾驶场景中基于单目图像的三维目标检测研究 [硕士学位论文]. 杭州: 浙江大学, 2023.

9 张冬冬, 郭杰, 陈阳. 基于原始点云的三维目标检测算法. 计算机工程与应用, 2023, 59(3): 209–217. [doi: 10.3778/j.issn.1002-8331.2109-0239]

10 Chen YK, Li YW, Zhang XY, *et al*. Focal sparse convolutional networks for 3D object detection. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 5418–5427.

11 刘越, 刘芳, 武奥运, 等. 基于自注意力机制与图卷积的 3D 目标检测网络. 计算机应用, 2024, 44(6): 1972–1977.

12 Wang YJ, Mao QY, Zhu HQ, *et al*. Multi-modal 3D object detection in autonomous driving: A survey. International Journal of Computer Vision, 2023, 131(8): 2122–2152. [doi: 10.1007/s11263-023-01784-z]

13 彭湃, 耿可可, 王子威, 等. 智能汽车环境感知方法综述. 机械工程学报, 2023, 59(20): 281–303.

14 Yeong DJ, Velasco-Hernandez G, Barry J, *et al*. Sensor and sensor fusion technology in autonomous vehicles: A review. Sensors, 2021, 21(6): 2140. [doi: 10.3390/s21062140]

15 Vora S, Lang AH, Helou B, *et al*. PointPainting: Sequential fusion for 3D object detection. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 4603–4611.

16 Xu SQ, Zhou DF, Fang J, *et al*. FusionPainting: Multimodal fusion with adaptive attention for 3D object detection. Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference. Indianapolis: IEEE, 2021. 3047–3054.

17 Chen XZ, Ma HM, Wan J, *et al*. Multi-view 3D object detection network for autonomous driving. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6526–6534.

18 Chen XY, Zhang TY, Wang Y, *et al*. FUTR3D: A unified sensor fusion framework for 3D detection. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver: IEEE, 2023. 172–181.

19 Mai NAM, Duthon P, Khoudour L, *et al*. 3D object detection with SLS-fusion network in foggy weather conditions. Sensors, 2021, 21(20): 6711. [doi: 10.3390/s21206711]

20 Sindagi VA, Zhou Y, Tuzel O. MVX-Net: Multimodal VoxelNet for 3D object detection. Proceedings of the 2019 International Conference on Robotics and Automation. Montreal: IEEE, 2019. 7276–7282.

21 Zhang YN, Chen JX, Huang D. CAT-Det: Contrastively augmented Transformer for multimodal 3D object detection. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 898–907.

22 Ku J, Mozifian M, Lee J, *et al*. Joint 3D proposal generation and object detection from view aggregation. Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid: IEEE, 2018. 1–8.

23 Bai XY, Hu ZY, Zhu XG, *et al*. TransFusion: Robust LiDAR-camera fusion for 3D object detection with Transformers. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 1080–1089.

24 Li X, Ma T, Hou YN, *et al*. LoGoNet: Towards accurate 3D object detection with local-to-global cross-modal fusion. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 17524–17534.

25 Huang TT, Liu Z, Chen XW, *et al*. EPNet: Enhancing point features with image semantics for 3D object detection. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 35–52.

26 Wu XP, Peng L, Yang HH, *et al*. Sparse fuse dense: Towards high quality 3D detection with depth completion. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 5408–5417.

27 Yang LX, Zhang RY, Li LD, *et al*. SimAM: A simple, parameter-free attention module for convolutional neural networks. Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021. 11863–11874.

28 Reynolds JH, Chelazzi L. Attentional modulation of visual processing. Annual Review of Neuroscience, 2004, 27: 611–647. [doi: 10.1146/annurev.neuro.26.041002.131039]

29 Deng JJ, Shi SS, Li PW, *et al*. Voxel R-CNN: Towards high performance voxel-based 3D object detection. Proceedings of the 35th AAAI Conference on Artificial Intelligence. AAAI, 2021. 1201–1209.

30 Yan Y, Mao YX, Li B. SECOND: Sparsely embedded convolutional detection. Sensors, 2018, 18(10): 3337. [doi: 10.3390/s18103337]

31 Shi SS, Guo CX, Jiang L, *et al*. PV-RCNN: Point-voxel feature set abstraction for 3D object detection. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 10526–10535.

32 Wu H, Deng JH, Wen CL, *et al*. CasA: A cascade attention network for 3-D object detection from LiDAR point clouds. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 5704511.

33 Li YC, Li ZX, Teng SY, *et al*. AutoMine: An unmanned mine dataset. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 21276–21285.

(校对责编: 张重毅)