DOI: 10.19650/j.cnki.cjsi.J2513943

一种改进的鸟瞰图视角下相机/激光雷达融合感知算法

夏若炎[1,2],徐晓苏[1,2]
(1. 微惯性仪表与先进导航技术教育部重点实验室 南京 210096; 2. 东南大学仪器科学与工程学院 南京 210096)

摘 要:在自动驾驶感知任务中,通过将不同模态的信息投影到统一的空间表示,实现基于鸟瞰图的相机和激光雷达特征多模态融合已成为主流研究范式。虽然 BEVFusion 等代表性框架能够实现较高的三维目标检测精度,但其在二维图像特征向 BEV 空间的视角转换过程中依赖深度预测,该模块不仅模型复杂、参数冗余,还存在推理效率低、内存消耗高等问题,对硬件资源提出了较高的要求,限制了模型在边缘设备或资源受限场景中的部署与应用。针对上述问题,在 BEVFusion 框架基础上,围绕视角转换过程的精度与效率瓶颈展开研究,提出了一种融合相机与激光雷达信息的 BEV 视觉特征优化算法。该算法利用激光雷达提供的深度信息替代图像深度预测,通过将其嵌入图像特征表达过程,实现对原有视角转换路径的结构性简化,并对 BEV 空间构建与池化模块进行了精简重构,有效降低了计算复杂度。实验结果表明,在保持三维物体检测精度不变的前提下,优化后方案将关键模块推理时间缩短至原方案的 16%,端到端推理速度提升 83%,峰值显存占用降低 27%,同时显著减轻了对输入图像分辨率的限制,增强了模型对算力资源的适应能力,提升了其在实际部署中的可行性。

关键词: 激光雷达;相机;融合感知;鸟瞰图;模型优化
中图分类号: TH701 **文献标识码**: A **国家标准学科分类代码**: 460.4099

An improved camera-LiDAR fusion perception algorithm in the bird′s-eye view perspective

Xia Ruoyan[1,2], Xu Xiaosu[1,2]
(*1. Key Laboratory of Micro-inertial Instrument and Advanced Navigation Technology, Ministry of Education, Nanjing 210096, China; 2. School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China*)

Abstract: In autonomous driving perception tasks, a multi-modal fusion of camera and LiDAR features based on a bird′s-eye view has become a mainstream research paradigm to combine information from different modalities into a unified spatial representation. Although representative frameworks such as BEVFusion achieve high 3D object detection accuracy, they rely heavily on depth prediction during the perspective transformation from 2D image features to the BEV space. This depth module is often complex, parameter-intensive, and results in low inference efficiency and high memory consumption, posing challenges for deployment on edge devices or resource-constrained platforms. To address these issues, we build upon the BEVFusion framework and focus on improving the accuracy and efficiency of the perspective transformation process. A BEV visual feature optimization algorithm is proposed, which integrates camera and LiDAR information by embedding LiDAR-provided depth data into the image feature representation, replacing the original depth prediction module. Additionally, the BEV space construction and pooling modules are restructured for computational efficiency. Experimental results show that, without compromising 3D detection accuracy, the proposed method reduces the inference time of key modules to 16% of the original, improves end-to-end inference speed by 83%, and lowers peak memory usage by 27%. It also significantly reduces sensitivity to input image resolution, enhancing adaptability to varying compute resources and improving deployment feasibility in real-world applications.

Keywords: LiDAR; camera; fusion perception; bird′s-eye view; model optimization

收稿日期:2025-04-18 Received Date: 2025-04-18

0 引　　言

环境感知任务作为自动驾驶算法领域的上游任务,对精度和实时性有很高的要求。早期的激光雷达(light detection and ranging, LiDAR)点云与相机图像融合算法往往分为两个阶段进行,即首先通过候选框将二维图像初步映射到三维空间,随后将其与激光点云相结合来生成最终的物体检测结果,如文献[1-3]所述。随着神经网络近年来不断地更新发展,目前的研究重点已经转向了更统一的端到端融合模型。

现有的端到端融合算法大致可分为数据级融合与特征级融合两类。数据级端到端融合算法有两种思路,第一种是从图像中提取颜色特征映射到对应的点云,典型算法如 Vora 等[4]提出的 PointPainting、Wang 等[5]提出的 PointAugmenting、Xu 等[6]提出的 FusionPainting 等。该种方式在处理稀疏雷达点云时语义上会有损失,另一种思路是反过来将点云投影给图像,为对应的像素附加深度值,对应的算法有 Qi 等[7]提出的 FrustumPointNet 和 Wang 等[8]提出的 FrustumConvNet 等。针对点云数据的稀疏性,Wang 等[9]提出 Sparse2Dense 算法,在融合前对点云进行上采样来实现深度补全。

特征级融合算法需要先将激光雷达和相机采集的数据分别进行特征提取,然后通过神经网络将提取的特征进行融合,最后根据融合后的特征进行目标检测。特征级融合算法弥补了点云与图像在结构特征上的巨大差异,因此成为当前研究的主流方向。2023 年 Chen 等[10]提出的 3D 融合变换器模型(fusion transformer for 3D, FUTR3D)作为一种典型的端到端特征级统一传感器融合框架可用于几乎任何传感器配置,其中自定义的特征融合器模块兼容相机、毫米波雷达和不同分辨率激光雷达之间任意的特征值融合。类似的特征级融合算法还有 Wu 等[11]提出的 MVFusion 和 Bai 等[12]提出的 TransFusion。二者都引入了上下文机制强化融合模块,区别在于前者关注多个相机的内模态融合,致力于解决多视角之间的边界断裂问题;后者将上下文机制用于雷达和相机的融合,使用位置感知编码引导二者的特征统一。

此外,周志伟等[13]则在数据和特征两级进行融合,来弥补单级融合的缺陷。也有研究工作在后端进行语义级融合,对点云和图像数据分别进行目标检测后再对两个检测结果进行匹配与跟踪,择优输出最终感知结果,如文献[14-16]所述。

近期,基于三维鸟瞰图(bird′s-eye-view,BEV)视角下的端到端特征级融合感知技术得到快速发展,2020 年 Philion 等[17]提出“提升-投影-输出”(lift-splat-shoot, LSS)算法,将图像信息投影到鸟瞰图视野生成视觉 BEV 特征进行感知的理论获得了广泛的关注。BEV 视角下的图像特征与传统的视锥体特征相比更加便于与激光雷达特征融合,因此视觉 BEV 特征很快的被广泛应用于激光雷达相机融合算法。其中 Liu 等[18]提出 BEVFusion 模型,将 BEV 图像特征与激光雷达特征融合,结合两种传感器的优势,将多模态融合技术引入了鸟瞰图领域,获得了显著的精度优势。在此基础上,2022 年 Chen 等[19]和 Borse 等[20]不约而同地指出了当前框架在融合模块的拼接粗糙性,前者由此提出了一种特征对齐损失函数来衡量投影相机特征与 LiDAR 特征之间的相关性,弥补了坐标对齐误差给特征融合带来的潜在不准确性。Jiao 等[21]提出多尺度多深度融合模型(multi-scale multi-depth fusion, MSMDFusion),用激光雷达点来估计摄像头特征的 3D 位置;Wang 等[22]则提出一种统一的多模态转换器(unified transformer, UniTR)机制,通过预先分配深度给每个摄像头特征,对图像投影模块提出了改进。

在国内,张炳力等[23]引入注意力机制优化融合模块,在实车搭载平台上验证了算法的可迁移性。针对露天矿山的特殊场景。崔文等[24]额外引入光场数据与雷达点云和图像在 BEV 视角下进行融合,提高了无人驾驶卡车对天气恶劣、多碎石环境的感知精度。金宇锋等[25]针对不同模态数据融合时的错位现象,采取了变形注意力机制进行优化,实现了精度提升。孙备等[26]采用交叉注意力机制,改进无人机对地伪装目标的检测定位。于睿等[27]提出一种基于自监督深度学习的热成像与激光雷达融合深度补全方法,用于在低光照或无光照的条件下生成像素级稠密的深度图。

上述研究推动了融合感知算法在不同领域、不同场景的应用,但特征级算法通常面临数据量庞大、运行效率低以及内存占用高等问题。对图形处理器(graphics processing unit,GPU)硬件配置的依赖性提升了 BEV 感知技术在实际应用中的部署成本,制约了该技术的实际应用。此外,现有方法在传感器特征提取阶段的边际效益也日益减弱,制约了整体性能的进一步提升。

针对上述问题,一些研究工作从算法或工程角度进行了优化。例如,宋建辉等[28]提出了一种高性能混合网络(improvement performance hybridnets, IPHNet)以更准确地完成实时感知任务;Long 等[29-30]则利用英伟达实验室推出的底层推理加速技术,对 BEVFusion 模型进行训练后优化,提升推理速度。然而,这些优化方法往往高度依赖特定骨干网络结构或硬件平台,限制了模型的泛化能力与部署灵活性。

因此,本研究提出了一种新的多传感器鸟瞰图特征投影方法,将激光雷达深度信息嵌入图像特征表达中,代替原有的离散深度预测模块。实验结果显示,在同样

GPU 配置下,优化后的模块在保持感知精度基本不变的前提下,将关键模块推理时间缩短至原方案的 16%,端到端推理速度提升 83%,峰值显存占用降低 27%。同时,优化后模型的全类平均正确率(mean average precision, mAP)与原模型持平,验证了该方法在提升效率的同时不牺牲检测精度。

本研究的主要贡献为:

1) 通过实验验证了 BEVFusion 中基于 LSS 方式的深度特征估计模块精度较低,且图像特征投影模块并未充分利用深度信息;同时,去除深度信息不会显著降低感知模型的预测效果;

2) 对现有的视角变换模块进行优化,通过将激光雷达深度信息嵌入图像特征表达中,代替了原有的深度预测模块,并对 BEV 空间构建与池化模块进行了重构;

3) 采用不同尺度图像结构的对比实验结果表明,优化后的算法能够更快处理更大空间尺度的图像特征,同时可以在资源有限的硬件平台上运行,增强了硬件适配性。

1 BEVFusion 的可优化性分析

本文的研究主要基于开源框架 BEVFusion 展开,BEVFusion 是一种典型的端到端训练大模型,尽管在三维感知任务中效果出色,但对于训练设备和时间成本都有很高的要求,限制了其在资源受限场景下的广泛应用。如图 1 所示,BEVFusion 的主要框架的组成模块和流程为:图像特征提取模块和点云特征提取模块将图像与点云信息转化为 BEV 特征,来自两种模态的 BEV 特征沿着通道维度进行拼接后,再输入目标检测模块,实现三维目标检测、地图语义分割等下游任务。

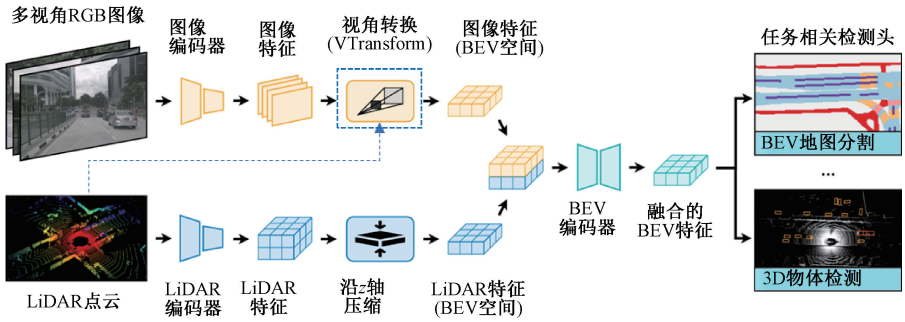

图 1 BEVFusion 流程

Fig. 1 Pipeline of BEVFusion

1.1 BEVFusion 耗时分析

为探究模型的计算瓶颈,本研究对推理时延进行模块级分解,在 RTX 4080 GPU 单卡硬件环境下,使用官方预训练模型对 nuScenes 验证集进行推理测试,各模块时延分布如表 1 所示。

表 1 模型各个模块用时表

Table 1 Time consumption of each module

模块名称	模块用时/ms
雷达特征编码	20.4
图像特征编码	8.2
图像特征视角变换(VTransform)	42.2
融合特征编码器	1.6
多任务检测头	5.9
推理总用时	78.3

测试结果表明,单次推理平均耗时 78.3 ms, 其中图像特征视角转换(VTransform)占用了整个模型端到端推理大约 54%的运算时间,该模块中高分辨率信息的加载和复杂矩阵运算已成为制约推理实时性的关键因素。因此如果能针对这一部分进行优化,那么对提升整体的推理效率是显著的。

1.2 视角变换模块有效性分析

BEVFusion 的视角转换 VTransform 模块包含 3 部分,分别为基于概率分布的深度预测模块、BEV 空间构建模块、和特征聚合池化模块,如图 2 所示。3 个模块均可优化,具体分析见后文。

1)基于概率分布的深度预测

LSS 算法首次提出了将环视图像特征投影到 BEV 空间中的多区间深度预测方法,基于给定大小的 BEV 网格分别统计每个网格区域的语义信息和深度信息。BEVFusion 在 LSS 基础上进行改进,引入激光雷达信息辅助深度预测,深度预测模块的输入为图像特征编码器基于多视角图像生成的图像特征 $\boldsymbol{F}_{cam}$ 和原始点云 $\boldsymbol{P}_{liDAR}$,算法具体流程为:

(1)将 $\boldsymbol{P}_{liDAR}$ 通过外参转换到图像坐标系下,并且根据图像特征 $\boldsymbol{F}_{cam}$ 的分辨率进行体素化,生成结构化特征 $\boldsymbol{F}_{liDAR}$,过程称为 $F_{DTransform}$;

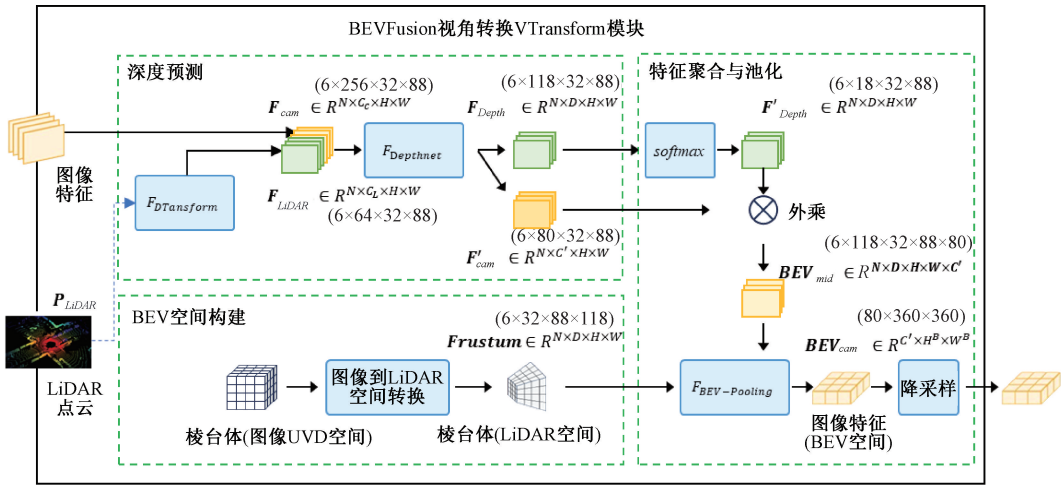

图 2　BEVFusion 视角转换模块流程

Fig. 2　Pipeline of VTransform module in BEVFusion

(2)将 $\boldsymbol{F}_{liDAR}$ 与 $\boldsymbol{F}_{cam}$ 沿着通道维度拼接,输入二维卷积深度估计算法 $F_{Depthnet}$,输出离散深度预测结果和降采样后的图像特征。

基于上述流程,深度预测模块的计算过程如式(1)所示。

$$\begin{aligned}&\boldsymbol{F}_{Lidar}=F_{DTransform}(\boldsymbol{P}_{liDAR})\\&\boldsymbol{F}'_{cam},\boldsymbol{F}_{Depth}=F_{Depthnet}(Concat(\boldsymbol{F}_{cam},\boldsymbol{F}_{LiDAR}))\end{aligned}\tag{1}$$

式中:$\boldsymbol{F}_{Depth}$ 为离散深度预测结果;$\boldsymbol{F}'_{cam}$ 为降采样的图像特征;$Concat$ 为拼接操作。$F_{Depthnet}$ 将图像特征中的每个像素在其对应的成像射线方向上划分为多个离散的深度区间后,预测其落在各个深度区间的概率分布。$\boldsymbol{F}_{Depth}$ 的输出格式为(N,D,H,W),其中 D 表示深度离散区间个数,N 为相机数,$H\times W$ 为图像特征的分辨率。

为直观验证多区间深度估计算法的效果,随机选取 3 张不同场景下的深度预测结果,对深度通道按概率峰值压缩后可视化,结果如图 3 所示。

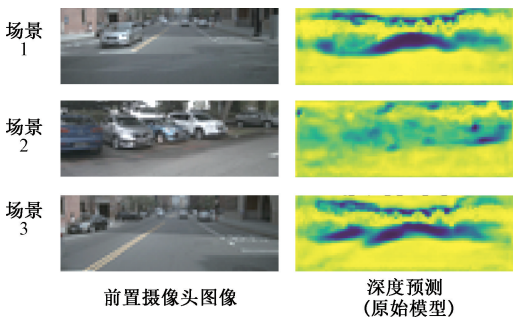

图 3　深度预测结果的可视化

Fig. 3　Visualization of the depth prediction

从右侧的深度图可以看出,$F_{Depthnet}$ 输出的深度预测结果整体呈现模糊、缺乏清晰边界的形态,表明深度估计算法难以对目标轮廓进行精确分离。这揭示了多区间的深度概率估计存在一定的不确定性,无法有效还原真实场景中的深度分布。

2)BEV 空间构建与特征聚合池化

为了将二维图像特征提升到三维空间,算法将 $\boldsymbol{F}_{Depth}$ 进行归一化操作后,与 $\boldsymbol{F}'_{cam}$ 外乘升维为 3D 场景特征中间值 $\boldsymbol{BEV}_{mid}$,并且构造 BEV 网格对中间值进行聚合和池化操作,得到视觉 BEV 特征 $\boldsymbol{BEV}_{cam}$,这一过程表达如公式(2)所示。

$$\begin{cases}\boldsymbol{BEV}_{mid}=\boldsymbol{F}'_{cam}\times\mathrm{softmax}(\boldsymbol{F}_{Depth})\\\boldsymbol{BEV}_{cam}=F_{BEV-Pooling}(\boldsymbol{Frustum},\boldsymbol{BEV}_{mid})\end{cases}\tag{2}$$

其中,softmax 为归一化操作,确保每个像素在所有深度区间上的概率分布之和为 1,从而将图像特征的深度估计由回归问题转化为多分类任务。通过外积运算将图像特征与深度概率分布相结合,实现从 2D 到 3D 的特征升维。具体而言,该过程将每个像素的特征沿其成像射线方向扩展至多个深度区间,最终生成具有空间语义信息的 3D 特征表示。

$F_{BEV-Pooling}$ 算法负责将体素特征投影至 BEV 空间,$\boldsymbol{Frustum}$ 是一个基于相机分辨率和成像范围设计的固定矩阵,通过预定义的相机内外参矩阵,算法在其中生成了每个相机视角的像素与鸟瞰图像素之间的的映射索引。基于此索引,$F_{BEV-Pooling}$ 将 3D 场景特征中间值 $\boldsymbol{BEV}_{mid}$ 投影至 BEV 空间的对应网格单元,对落入每个 BEV 网格的所有体素特征向量进行池化,池化结果即为视觉 BEV 特征 $\boldsymbol{BEV}_{cam}$。BEV 空间的构建以及特征的聚合池化过程涉及大量浮点运算,显著增加了推理过程的计算负担,影响整体推理效率。

3)深度预测有效性分析

为系统评估不同多区间深度估计算法对生成视觉

BEV 特征精度的影响,随机选取一个典型场景,设计并对比了 2 种不同深度预测方法的特征表示。具体方法为:对生成的 BEV 特征矩阵沿通道维度进行平均处理,获得可视化图像,结果如图 4 所示。

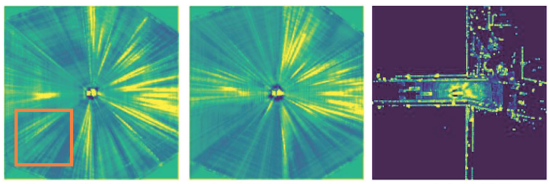

图 4 视觉 BEV 特征与激光雷达 BEV 特征可视化对比

Fig. 4 Comparison between the BEV features from camera and liDAR

实验设置具体为:

(1)去除结构化雷达特征,采用无雷达辅助的纯视觉深度预测方法,对生成的图像 BEV 特征进行可视化(图 4 左),相应的深度预测模块更新如式(3)所示。

$$\boldsymbol{F}'_{cam},\boldsymbol{F}_{Depth}=\boldsymbol{F}_{Depthnet}(\boldsymbol{F}_{cam}) \tag{3}$$

(2)使用原始模型配置,采用有激光雷达点云辅助的深度预测方法,对生成的图像 BEV 特征可视化(图 4 中);

(3)直接使用点云生成的激光雷达 BEV 特征图,作为对比参考(图 4 右)。

从图 4 左侧的两张图像 BEV 可视化图来看,无论是否引入点云辅助,深度预测结果在鸟瞰图中普遍呈现沿视线方向的放射状纹理。其中,亮度平均的放射线条表示该方向上预测不确定性较高,系统倾向于采用均匀分布将图像特征平均赋值至该区域对应的 BEV 网格。在存在遮挡或非道路区域的场景下,放射线明显缩短,说明深度概率分布集中于物体的实际空间位置。引入点云辅助信息后,在建筑物等复杂结构区域(如图 4 中方框所示),原先存在的异常放射线被有效抑制,说明深度预测准确性有所提升。然而,与图 4 右侧直接基于点云生成的激光雷达 BEV 特征相比,融合后的视觉深度预测结果仍存在一定模糊性,表明尽管融合点云能够优化预测效果,但由于视觉传感器的深度测量存在先天不足和不确定性,其改进存在上限。

1.3 视角变换模块冗余性分析

在前述结论的基础上,进一步探讨深度估计模块的精度对 BEVFusion 最终感知任务性能的影响。BEVFusion 架构不仅包含视觉特征编码器,还引入了点云特征编码器,其通过动态体素化+稀疏编码方式从点云数据中提取激光雷达特征。随后,BEV 特征融合模块将视觉 BEV 与激光雷达 BEV 特征进行拼接后再进一步在语义级别进行融合提炼,并将融合结果输入至多任务检测头。

然而,由于深度预测模块在某些场景下精度偏低,故提出一个假设:在后端的 BEV 特征融合过程中,特征融合和检测模块可能通过自适应权重机制,更多地依赖于激光雷达提供的准确深度信息,从而削弱了对视觉深度预测的依赖。若该假设成立,则生成阶段中引入点云的深度信息对最终检测结果的贡献将显得冗余,不仅增加了显存占用和计算开销,也可能并未显著提升检测精度。

1)消融实验

为了验证这一猜想进行了消融实验,通过人为干预的方式修改 $\boldsymbol{F}_{Depth}$ 为不具备语义意义的虚拟数值,运行 nuScenes 数据测试集和评估集,观察其对三维目标检测 mAP 的影响,从而分析深度估计模块的重要性。实验设计了 4 种修改模式:所有深度区间的概率值均为 1、-1、0.008 5(在深度通道 $D=118$ 时满足平均概率归一化)以及随机数。实验设计分为两个阶段:

(1)基准测试:加载 BEVFusion 官方预训练模型定量评估其深度预测模块在目标域的表现;

(2)自适应修正验证:基于预训练模型进行了迭代次数为 6 的微调训练,验证模型能否通过有限迭代自动校正深度概率偏移。

实验结果如表 2 所示。

表 2 修改 $\boldsymbol{F}_{Depth}$ 后 BEVFusion 在验证集上取得的推理结果

Table 2 Testing results of BEVFuion on nuScnees(val) after modifying $\boldsymbol{F}_{Depth}$

$\boldsymbol{F}_{Depth}$ 修改模式	全类平均正确率(*mAP*)	
	预训练模型(test)	重新训练模型(eval)
原始模型	71.5	68.1
全为 1	41.8	68.0
全为 0.008 5	66.0	68.1
全随机数	42.5	68.0
全为-1	65.3	67.9

由表 2 可以看到,在未进行微调的情况下,$\boldsymbol{F}_{Depth}$ 替换为全 1 或随机数显著降低了检测精度,表明这类修改会扰乱模型原有的分布假设;而当 $\boldsymbol{F}_{Depth}$ 采用归一化后的均值 0.008 5 或-1 时,模型的推理精度与原始输入情况下相近,说明在预训练模型中,视觉转换模块对深度维度的建模效果更接近于一种平均汇聚机制,未对特定深度位置赋予明显权重。

表 2 右侧展示了在深度输入被扰乱后,重新训练模型所获得的性能。经过微调训练后,所有改动版本的精度几乎恢复到与原始模型一致的水平,尤其是原本精度下降明显的“全为 1”和“全随机数”两种配置,其 mAP 从 41.8 和 42.5 恢复到 68.0。这种结果表明,即便输入使

用的是与真实场景完全不符的深度估计值,模型仍能通过训练学习到忽略无效深度信息、聚焦于其余有效特征源,推理性能几乎无损。该现象进一步验证了:当前深度估计模块对 BEV 感知性能的贡献有限,其所提供的信息在推理过程中未起到决定性作用,甚至可被模型在训练中自动规避。

2)可视化对比

为进一步验证深度预测模块的冗余性,在量化检测精度之外,还对消融实验中生成的视觉 BEV 特征进行了直观可视化分析,结果如图 5 所示。

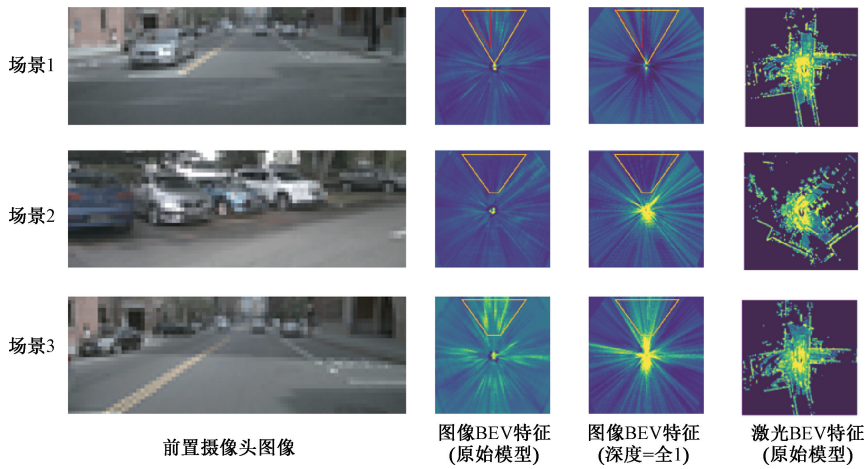

图 5　消融实验中的视觉 BEV 特征可视化效果

Fig. 5　Visualization of BEV feature map in ablation study

图 5 展示了 3 个不同的测试场景中的 BEV 特征可视化结果,每个场景分别包括:基于真实深度估计结果生成的视觉 BEV 特征、消融实验中深度概率值均为 1 时生成的视觉 BEV 特征和激光雷达 BEV 特征,针对每个场景的具体说明为:

(1)场景 1:存在局部近距离目标

在原始视觉 BEV 图像中,图像特征投影集中于目标存在区域,对应区域之外则呈现均匀分布。这一现象同样出现在消融模型中,进一步证明即使强制设定统一深度,模型仍可借助图像特征通道学习感知空间结构。

(2)场景 2:近距离静态目标聚集

该场景包含大量近距离静止车辆,在视觉 BEV 图像中,图像特征高度集中在摄像头位置附近,远距离区域基本无图像投影。而消融模型也呈现相同特征,这表明:模型在训练中已经学会将图像中的深度信息内嵌进特征通道中,而非仅依赖显式深度预测。

(3)场景 3:空旷区域,目标缺失

在该场景中,摄像头视野中几乎无 3D 物体。无论是原始模型还是消融模型,其特征可视化结果都呈现出均匀、明亮的放射状分布,说明在无明确目标的情况下,图像特征在整个深度方向均匀投影。

3)实验结论

根据以上实验结果,可得出判断:

深度估计模块为图像特征提供了从 2D 到 3D 的投影几何约束,但最终 3D 检测性能更依赖于 BEV 空间中跨模态特征的聚合质量。对于 BEVFusion 这类多模态融合模型而言,网络具有容错性,能够通过端到端的训练学习出自动补偿深度误差的机制,从而减弱深度分布不准确所带来的负面影响。另一方面,模型在训练过程中可以学习到从图像特征中获取场景结构信息,而不需要深度估计模块的显式预测结果。因此在计算资源受限的条件下,针对深度预测模块进行优化可以在保证精度不变的情况下提升运算效率。

1.4　现有模型耗时原因分析

视觉 BEV 特征提取环节中,各个部分的实际运算值和计算复杂度具体为:

1)视觉 BEV 特征的输入为 $\boldsymbol{F}_{cam} \in R^{N\times C_c\times H\times W}$,实际实验中 $N=6$ 代表相机数,$C_c=256$ 代表图像特征的通道数,$H\times W=32\times 88$ 代表图像特征的分辨率。根据式(2)可知,$\boldsymbol{F}_{LiDAR}$ 与 $\boldsymbol{F}_{cam}$ 具有相同的特征分辨率和不同的特征通道数,在实际运算中 $\boldsymbol{F}_{LiDAR} \in R^{N\times C_L\times H\times W}$ 的通道数 $C_L=64$,$\boldsymbol{F}_{LiDAR}$ 与 $\boldsymbol{F}_{cam}$ 拼接后的特征通道数为 320。

2)根据式(3),$F_{Depthnet}$ 模块基于拼接特征生成深度预测结果,$\boldsymbol{F}_{Depth}$ 的结构为(6,118,H,W)。同时,$\boldsymbol{F}_{cam}$ 的通道数经过降采样缩小为80,因此经过外乘运算后 $\boldsymbol{BEV}_{mid}$ 的结构为(6,118,H,W,80),外乘算法同时涉及乘法和加法计算,因此其计算复杂度为 $O(2\cdot 6\cdot 118\cdot 32\cdot 88\cdot 80)$,在典型设置下约为 3.2 亿次乘加运算。

3)在 $F_{BEV-Pooling}$ 阶段,算法首先定义了与原始图像分辨率大小相同的空间索引 $\boldsymbol{Frustum} \in R^{N\times D\times H\times W}$,随后通过像素平面与 BEV 空间的坐标系变换矩阵建立起 $\boldsymbol{Frustum}$ 与 $\boldsymbol{BEV}_{Cam}$ 之间的映射关系。基于映射后的索引,池化算法通过累加求和的计算方式将 $\boldsymbol{BEV}_{mid}$ 聚合为视觉 BEV 特征 $\boldsymbol{BEV}_{Cam}$。池化过程的计算复杂度为 $O(6\cdot 118\cdot 32\cdot 88\cdot 80)$,即每个体素特征进行一次加法操作,总计约为 1.6 亿次加法操作。

综上,整个视觉 BEV 特征提取模块涉及了读取百万个双精度浮点数接近 4.8 亿次运算操作,因此对推理的效率带来了压力。

2 视角转换模型的优化

基于上述的对视觉转换模块的可优化性分析,提出一种 BEVFusion 视觉特征提取过程的结构精简和功能重构方法,设计了一种无需深度分类估计的新型图像投影算法与模型构型。改进后的方法在不损失检测精度的前提下,显著提升了推理效率并降低了内存资源消耗。基于新算法的 VTransform 模块如图 6 所示,主要包括 3 方面:内嵌点云深度的图像特征降采样、BEV 空间构造,以及特征聚合与 BEV 池化重构。

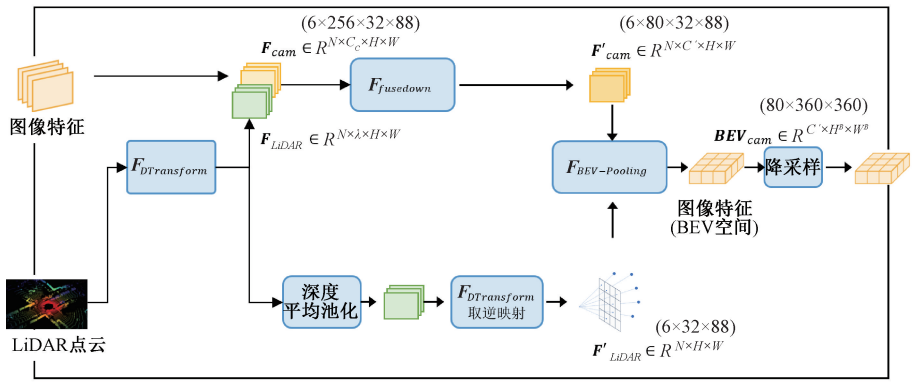

图 6 改进后的图像到 BEV 视角转换流程

Fig. 6 Improved image-to-BEV view transformation

2.1 内嵌点云深度的图像特征降采样

本文完全移除了原始架构中的深度估计模块,不再采用 $F_{Depthnet}$ 对每个像素进行深度概率分布的预测。取而代之,提出了一个基于轻量级卷积神经网络(convolutional neural network, CNN)的降采样模块 $F_{fusedown}$,旨在实现两项功能:对输入图像特征进行通道压缩,使其与后端处理模块的输入维度保持一致;将激光雷达提供的深度信息直接融合进图像特征通道中,实现特征级别的融合。

除此之外,方案保留了 $F_{DTransform}$ 模块用于提取点云深度信息,但对其结构进行了优化。在原始设计中,体素化后的点云被表示为一个具有 64 个通道的深度矩阵。然而,考虑到点云数据本身存在稀疏性与大量无效区域,较高的通道数可能引入冗余计算和不必要的内存开销。为此,对不同通道数进行了对比实验。实验结果表明,在通道数设为 8 的情况下,模型在保持感知性能的同时显著降低了计算成本与显存占用。因此,最终方案将 $F_{DTransform}$ 模块的深度通道数压缩为 8。

$F_{fusedown}$ 的输入包含来自图像编码器的图像特征以及 $F_{DTransform}$ 提取的点云深度信息。在算法内,二者拼接后通过卷积神经网络融合,输出融合后的图像特征。该设计使得模型能够在训练过程中自动学习深度误差的补偿机制,并将其内嵌于图像特征表达中,从而摆脱对深度预测模块 $F_{Depthnet}$ 的依赖。

2.2 BEV 空间构造

本方法摒弃了原方案在 BEV 空间构造中的棱台体建模与映射变换流程,改为直接利用点云的位置为图像特征提供坐标,具体流程为:

1)对 $F_{DTransform}$ 输出的多通道深度值进行平均池化,得到图像特征尺度下每个像素对应的深度值;

2)将深度值从像素空间逆映射回 LiDAR 坐标系,进而确定其在 BEV 空间中的位置;

3)最终生成的位置矩阵为 $\boldsymbol{F}'_{LiDAR} \in R^{N\times H\times W}$,每个像素特征对应一个精确空间坐标点。此方法将每张图像的投影从 3D 棱台体简化为一个 2D 的空间坐标网格,显著降低了空间复杂度与计算开销。

2.3 特征聚合与 BEV 池化重构

除了不再采用多区间深度概率表示,特征聚合与 BEV 池化结构中的外积操作也随之被移除。图像特征与对应的位置坐标可直接作为 $F_{BEV-Pooling}$ 的输入,完成投影与聚合,无需额外的体素构建步骤。

为了与新的输入结构适配, $F_{BEV-Pooling}$ 重新定义了池化结构的输入维度。图像特征输入数据结构从原先的五维度张量 $\boldsymbol{BEV}_{mid} \in R^{N\times D\times H\times W\times C'}$ 变为了四维张量 $\boldsymbol{F}'_{cam} \in R^{N\times C'\times H\times W}$,在 $D=118$ 的典型配置下,不仅将池化层的空间占用率减少到之前的 0.008 5 倍,更进一步降低了视觉转换模块的运算时长。

综上所述,优化后的模型从各方面对原设计进行了重要的结构精简,去除了冗余的深度维度,并避免了棱台体空间映射,从而推理效率显著提升,显存消耗显著降低。

3　实验验证

针对优化方法的有效性和可迁移性,分别在公开数据集和实验室开发的样机平台,对优化算法进行量化评估和定性的可视化结果展示与分析。

3.1　优化模型在公开数据集上的量化评估

针对优化方法的有效性,在完全一致的环境下将优化算法与原始 BEVFusion 算法进行对比,采用 Motional 团队发布的公开数据集 nuScenes 进行量化的评估。nuScenes 是一个面向自动驾驶场景的大规模多模态感知数据集,包含激光雷达与 6 个环视相机(前、前左、前右、后左、后右、后)采集的数据,覆盖 360° 全景视野。数据集包含超过 1 000 个动态场景,每个场景时长约 20 s,并为每一帧提供了完整的相机内参、外参以及高质量的三维真值框标注。

1)评价指标说明

在实验设置方面,分别对比了原始算法与改进后算法在 3 个维度的性能表现,即:

(1)检测精度指标:使用 nuScenes 官方评估标准,包括平均精度指标 mAP 与归一化检测分数(nuScenes detection score,NDS)对比模型在三维目标检测任务中的感知能力;

(2)推理效率指标:统计两种算法在推理阶段的平均用时(ms),以评估模型的实时处理能力;

(3)内存资源占用:测量模型运行时显存消耗峰值,验证改进结构在资源使用上的优化效果。

2)优化模型精度验证

本实验基于 BEVFusion 的官方预训练模型开展,采用相同的参数配置对原始模型和优化后的模型分别进行了微调训练。在训练时冻结了点云与图像的特征编码器网络,仅开放视角变换模块及其后续结构的参数。训练过程采用了 4 张 NVIDIA RTX4090 显卡,单批次(batch size)设置为 8,训练最大迭代次数固定为 6。

如 2.1 节所述,为验证不同深度通道系数对图像特征投影的引导能力,在 $F_{DTransform}$ 环节设置不同深度通道数($\lambda=0,1,2,8,64$)进行了对比实验,并将优化模型在各组配置下的三维物体检测结果与原始模型进行了对比,结果如表 3 所示。

表 3　原始模型与优化视角转换模块的推理结果对比

Table 3　Results from original model and the optimized VTransform module

模型与深度通道系数	验证集		测试集	
	mAP	*NDS*	*mAP*	*NDS*
优化模型($\lambda=0$)	67.6	70.7	–	–
优化模型($\lambda=1$)	67.9	70.6	–	–
优化模型($\lambda=2$)	68.1	71.0	–	–
优化模型($\lambda=8$)	**68.2**	**71.2**	**68.2**	**71.2**
优化模型($\lambda=64$)	68.1	71.1	–	–
原始模型	**68.5**	**71.4**	**70.2**	**72.9**

从实验结果可以看出,优化后的模型在精度及误差指标与原始模型基本一致,未出现明显的精度下降,验证了在剔除深度预测模块后,模型性能依然得以保持。同时 $\lambda=8$ 时取得最优精度,表明该通道配置在精度与效率之间达到了良好平衡。其中测试集上的精度略低于原始模型,主要原因是原始模型在提交测试集结果时,采用了训练集与评估集的联合训练策略,而本文受限于时间未进行联合训练,仅在训练集上进行训练。

3)推理效率与内存资源占用结果和对比

为进一步验证优化后模型在不同硬件配置下的适应能力,特别是在更大的输入图像分辨率时对硬件内存资源和执行效率的需求,实验选择了不同图像特征输入尺寸,在 NVIDIA RTX4080 和 RTX4090D 两种显卡平台上进行了推理测试,二者的参数差异如表 4 所示。

表 4　RTX4080 移动版与 RTX4090D 桌面版配置对比

Table 4　Configuration comparison between RTX 4080 mobile and RTX 4090D desktop

参数	RTX 4090D (桌面版)	RTX 4080 (移动版)
GPU 芯片	AD102	AD102
CUDA 核心数	14 592	7 424
显存容量	24 GB GDDR6X	12 GB GDDR6
显存位宽/bit	384	192
显存带宽/(GB·s^{-1})	1 008	432
最大功耗/W	450	150
FP32 理论性能/TFLOPS	73.5	24.7

所有的模型推理实验采用单卡执行，batch-size 设置为 1。在两种硬件平台上，模型分别采用了相同的学习参数，因此检测精度一致，仅在资源占用和推理效率上表现出差别。NVIDIA RTX 4090D 为 RTX 4090 的简化版本，性能约为正规版本的 90%。另一方面，实时性实验采用的 NVIDIA RTX 4080 为移动端版本，虽然移动端 RTX4080 在命名上与标准版一致，但其实际性能约为标准版本的 60%～70%，因此模型在二者的表现更能反映本算法在低成本实验室条件下的部署能力。

模型推理过程中的 GPU 内存占用量及 VTransform 模块的执行时间如表 5 所示，其中空单元表示该条件下，因内存溢出无法运算。

表 5 原模型与优化模型内存占用量和耗时对比

Table 4 Comparison of memory usage and time consumption between the original model and the optimized mode

硬件平台	输入图像特征尺寸	GPU 内存占用/GB		VTransform 部分用时/ms	
		原模型	优化后模型	原模型	优化后模型
RTX 4080 移动版	32×88	11.1	8.3	42.2	6.8
	64×176	–	8.7	–	12.6
	128×352	–	9.3	–	31.3
RTX 4090D	32×88	16.5	14.1	18.7	5.7
	64×176	21.8	14.3	63.8	11.1
	128×352	–	14.7	–	22.0

从表 5 可见，当使用默认图像特征尺寸 32×88 时，原始模型和优化模型在两个平台上均可运行。在相对配置更低的 RTX-4080 平台上，优化模型在不牺牲检测精度的前提下，将 VTransform 模块的执行时间从 42.2 ms 减少到 6.8 ms，仅为原模型的 16%；相应的，一次完整推理用时从原始模型的 78.3 ms 下降到 42.9 ms，整体推理效率提升约 83%，显著提升了推理实时性。显存占用方面，模型总的峰值内存消耗从 11.1 GB 下降到 8.7 GB，减少了约 27%。

当输入图像特征尺寸放大 2 倍至 64×176 后，原始模型在 RTX-4080 上出现内存溢出；在 RTX-4090 上虽可运行，但执行时间则从 18.7 ms 增加到 63.8 ms，实时性显著降低。优化算法则只增加到 12.6 ms，整体推理的实时性下降不明显。

图像特征放大 4 倍至 128×352 后，原始模型在两个平台上都无法执行。与之相比，优化后的模型在两种平台上均保持稳定运行，VTransform 模块的最大耗时不超过 31.3 ms，展现出良好的扩展性与实时性保障。

除此之外，表 4 中 VTransform 模块的执行时间为每次推理都重新计算图像与激光雷达之间几何变换的情况下测得，即假设相机内参、外参及与激光雷达的相对位姿在运行过程中动态变化。相机与雷达之间位姿的动态变换使得原模型无法依赖预计算缓存，严重影响其实时性与适用范围。

相较之下，优化方案不依赖于预计算，即便在每次推理过程中动态生成转换关系，也能够在多尺度图像输入场景下保持高效推理性能，显著提升了模型的通用性与工业落地能力。

综上所述，优化算法在 RTX 4090/4080 移动端平台上都能够实现稳定的实时检测，充分体现了模型结构在效率与计算资源受限条件下的优越性。相比于依赖桌面级高算力的算法，优化算法更贴近实际应用场景，在边缘计算、车载平台及移动设备部署中具有更强的实用价值。

3.2 样机实验

除了公开数据集，分别在校园内部道路和校外公共道路使用样机平台进行了实验，进一步测试优化算法在实际应用中的稳定性和实时性。数据采集平台搭载了 1 台 Ouster128 线激光雷达和 6 台海康相机，另外配备了霍尼韦尔的 IMU 和司南导航 GNSS 系统等传感器，采集环境包括停车场、密集人群和夜间主干道等多种特殊场景，采用优化算法进行目标检测的可视化结果见后文。

停车场环境的目标检测结果如图 7 所示，作为结构简单、检测目标类别最少的环境，车辆具有平行分布、互不完全遮挡的特点，优化算法取得了最优效果，即使在较远距离处，系统仍展现出较强的远距离检测能力，能识别出远处的交通锥和车辆。同时，系统在部分目标存在遮挡的条件下依然具备良好的检测能力，能够识别出被栏杆或树木遮挡的车辆。

图 7 停车场环境目标检测结果可视化

Fig. 7 Visualization of object detection in parking lot

图 8 展示了行人聚集的拥挤单向双车道环境下的目标检测结果。该场景中,前视角由于视觉重叠以及目标密度极高,导致部分密集人群未能被准确检测。然而,在后视相机视角下,优化算法成功检测出了后方密集区域中的行人群体,说明模型能够充分利用多视角信息来弥补局部感知盲区。此外,尽管远处车辆和电动车受到部分遮挡,模型仍准确完成了目标识别与三维框预测,体现出其对复杂遮挡情况下目标轮廓与特征的提取能力。

图 8　行人聚集的单向双车道目标检测结果可视化

Fig. 8　Visualization of object detection results on a one-way dual-lane road with crowded pedestrians

图 9 展示了夜间主干道环境的目标检测结果。尽管该场景下整体光线昏暗、可见度较低,尤其是部分区域几乎无路灯照明,但优化算法依然成功检测到了处于阴影区域的黑色汽车以及停靠在路边的摩托车等低反射目标。这表明模型在处理低光照条件下仍然具有较强的鲁棒性,确保了即使在极端照明环境中也能维持可靠的感知性能。

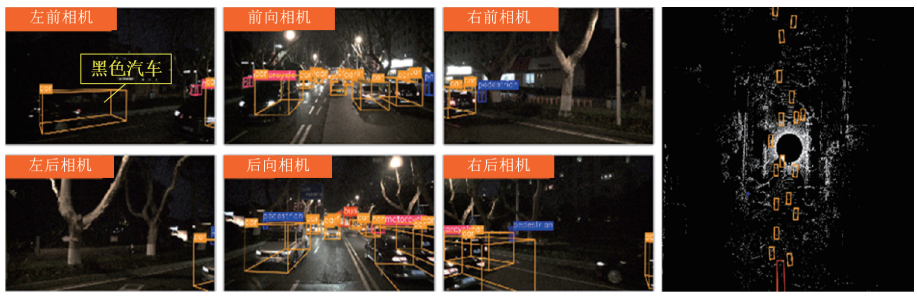

图 9　夜间主干道环境目标检测结果可视化

Fig. 9　Nighttime visualization of object detection results on a main road

图 10 展示了高速公路环境中算法对远距离高速移动目标的检测能力。该场景下车辆刚刚驶出高架桥阴影区域,光照条件发生剧烈变化,因此对感知算法的鲁棒性提出了较高要求。优化算法在此条件下仍然从前视相机视野中成功检测出远处同向行驶车道中的目标,确保系统能够提前获取足够的空间和时间冗余进行高效决策,证明了算法的可靠性。

图 11 展示了城市主干道环境的目标检测结果。从图中可以看出,尽管场景中存在大量密集排布的护栏结构、动态行人和多类别静态障碍物,优化算法仍能稳定输出准确的三维检测结果。前方行人、路边的交通锥、主车道中的各类汽车均被准确识别与定位,说明优化算法在城市高密度场景下具备较强的泛化能力与细节感知能力,适用于实际城市自动驾驶应用需求。

图 10　高速公路环境远距离目标检测结果可视化

Fig. 10　Visualization of far-distance object detection result on highway

图 11 城市主干道环境目标检测结果可视化

Fig. 11 Visualization of object detection results on a main road

此外,针对优化算法的实时性,分别统计基准算法 BEVFusion 和优化后的模型在 RTX 4080 平台上的平均前向推理耗时,并取其倒数计算得到各自的理论最大每秒可处理帧数(frames per second,FPS)。在相同条件下,BEVFusion 的 FPS 为 12.7,而优化模型的 FPS 为 23.3,这一对比结果证明了优化算法不仅可以在自搭建平台上有效运行,而且保持了较高的实时性。

综上所述,本研究分别在公开数据集上和样机平台上对优化算法进行了实验,结果展示了算法在不同硬件设备和环境条件下均可以保持稳定的检测精度和高于原模型的推理效率,证明了优化算法的鲁棒性。

4 结 论

本研究提出了相机/激光雷达融合的 BEV 视觉特征优化算法。与传统依赖深度估计与棱台体建模的视角转换方式不同,优化算法通过构建图像特征与体素化点云之间的直接映射关系,显著简化了视图变换流程。

该方法有效消除了转换过程中的冗余步骤,在保持检测精度的同时,大幅降低了内存占用,并显著提升了推理效率。在不同性能的计算平台上均表现出良好的执行性能和可扩展性。

在 nuScenes 数据集上的大量实验表明,优化后的模型在精度上与当前主流高精度方法相当,同时具备更强的实时性与资源适应能力,拓展了其应用领域,为多传感器融合感知任务提供了一种更高效、更灵活的解决方案。

参考文献

[1] CHEN X ZH, MA H M, 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.

[2] 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, 2018:1-8.

[3] DENG J, CZARNECKI K. MLOD: A multi-view 3D object detection based on robust feature fusion method[C]. 2019 IEEE Intelligent Transportation Systems Conference, 2019: 279-284.

[4] VORA S, LANG A H, HELOU B, et al. Pointpainting: Sequential fusion for 3D object detection[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020:4603-4611.

[5] WANG CH W, MA CH, ZHU M, et al. Pointaugmenting: Cross-modal augmentation for 3D object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition, 2021:11794-11803.

[6] XU SH Q, ZHOU D F, FANG J, et al. Fusionpainting: Multimodal fusion with adaptive attention for 3D object detection[C]. 2021 IEEE International Intelligent Transportation Systems Conference, 2021:3047-3054.

[7] QI C R, LIU W, WU CH X, 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.

[8] WANG ZH X, JIA K. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3D object detection[C]. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2019: 1742-1749.

[9] WANG T Y, HU X W, LIU ZH ZH, et al. Sparse2Dense: Learning to densify 3D features for 3D object detection[J]. Advances in Neural Information Processing Systems, 2022, 35: 38533-38545.

[10] CHEN X Y, ZHANG T Y, WANG Y, et al. FUTR3D: A unified sensor fusion framework for 3D detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023:172-181.

[11] WU Z ZH, CHEN G L, GAN Y ZH, et al. Mvfusion: Multi-view 3D object detection with semantic-aligned radar and camera fusion[C]. 2023 IEEE International

Conference on Robotics and Automation, 2023: 2766-2773.

[12] BAI X Y, HU Z Y, ZHU X G, et al. Transfusion: Robust liDAR-camera fusion for 3D object detection with transformers[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 1090-1099.

[13] 周志伟, 周建江, 王佳宾, 等. 基于雷达和视觉多级信息融合的目标检测网络[J]. 电子测量技术, 2024, 47(24):110-117.
ZHOU ZH W, ZHOU J J, WANG J B, et al. Target detection network on multi-level information fusion of radar and vision[J]. Electronic Measurement Technology, 2024, 47(24):110-117.

[14] 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.
ZHENG SH W, LI W H, HU J Y. Vehicle detection in the traffic environment based on the fusion of laser point cloud and image information[J]. Chinese Journal of Scientific Instrument, 2019, 40(12):143-151.

[15] 李研芳, 黄影平. 基于激光雷达和相机融合的目标检测[J]. 电子测量技术, 2021, 44(5):112-117.
LI Y F, HUANG Y P. Target detection based on the fusion of lidar and camera[J]. Electronic Measurement Technology, 2021, 44(5):112-117.

[16] 冯明驰,高小倩,汪静姝,等. 基于立体视觉与激光雷达的车辆目标外形位置融合算法研究[J]. 仪器仪表学报, 2021, 42(10):210-220.
FENG M CH, GAO X Q, WANG J SH, et al. Research on the fusion algorithm of vehicle object shape-position based on stereo vision and lidar[J]. Chinese Journal of Scientific Instrument, 2021, 42(10):210-220.

[17] PHILION J, FIDLER S. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D[C]. Computer Vision-ECCV 2020, 2020:194-210.

[18] LIU ZH J, TANG H T, AMINI A, et al. Bevfusion: Multi-task multi-sensor fusion with unified bird′s-eye view representation[C]. 2023 IEEE International Conference on Robotics and Automation, 2023: 2774-2781.

[19] CHEN Z H, LI ZH Y, ZHANG SH Q, et al. Autoalign: Pixel-instance feature aggregation for multi-modal 3D object detection[C]. Thirty-First International Joint Conference on Artificial Intelligence, 2022: 827-833.

[20] BORSE S, KLINGNER M, KUMAR V R, et al. X-align: Cross-modal cross-view alignment for bird′s-eye-view segmentation[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023:3287-3297.

[21] JIAO Y, JIE Z Q, CHEN SH X, et al. Msmdfusion: Fusing LiDAR and camera at multiple scales with multi-depth seeds for 3D object detection[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023:21643-21652.

[22] WANG H Y, TANG H, SHI SH SH, et al. Unitr: A unified and efficient multi-modal transformer for bird′s-eye-view representation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023:6792-6802.

[23] 张炳力, 潘泽昊, 姜俊昭,等. 基于交叉注意力机制的多模态感知融合方法[J]. 中国公路学报, 2024, 37(3):181-193.
ZHANG B L,PAN Z H,JIANG J ZH,et al. Multi-modal perception fusion method based on cross attention[J]. China Journal of Highway and Transport, 2024, 37(3): 181-193.

[24] 崔文, 朱文奇, 王康. 基于多模态融合 BEV 露天矿无人驾驶矿卡感知算法[J]. 工矿自动化, 2023, 49(S2):125-129.
CUI W, ZHU W Q, WANG K. Perception algorithm for unmanned mining trucks in open-pit mines based on multi-modal fusion BEV[J]. Industry and Mining Automation, 2023, 49(S2):125-129.

[25] 金宇锋, 陶重犇. 基于 Transformer 的融合信息增强 3D 目标检测算法[J]. 仪器仪表学报, 2023, 44(12): 297-306.
JIN Y F, TAO CH B. Fusion information enhanced method based on Transformer for 3D object detection[J]. Chinese Journal of Scientific Instrument, 2023, 44(12): 297-306.

[26] 孙备, 党昭洋, 吴鹏, 等. 多尺度互交叉注意力改进的单无人机对地伪装目标检测定位方法[J]. 仪器仪表学报, 2023, 44(6):54-65.
SUN B, DANG ZH Y, WU P, et al. Multi scale cross attention improved method of single unmanned aerial vehicle for ground camouflage target detection and localization[J]. Chinese Journal of Scientific Instrument, 2023, 44(6): 54-65.

[27] 于睿, 马国梁, 郭健, 等. 基于自监督学习的热成像与激光雷达融合深度补全方法[J]. 仪器仪表学报, 2025, 46(1): 170-181.
YU R, MA G L, GUO J, et al. Self-supervised learning-based depth completion method using thermal imaging and LiDAR fusion [J]. Chinese Journal of Scientific Instrument, 2025, 46(1): 170-181.

[28] 宋建辉, 刘鑫, 庄爽, 等. 面向无人驾驶的多任务环境感知算法研究[J]. 电子测量与仪器学报, 2025, 39(1): 122-132.
SONG J H, LIU X, ZHUANG SH, et al. Research on multi-task environment perception algorithm for unmanned driving[J]. Journal of Electronic Measurement and Instrumentation, 2025, 39(1): 122-132.

[29] LONG X, ZAN X, XUE J, et al. An acceleration inference implementation of bevfusion with mqbench on xavier[C]. 2023 China Automation Congress, 2023: 8665-8670.

[30] HUANG F, LIU S, ZHANG G, et al. Deployfusion: A deployable monocular 3D object detection with multi-sensor information fusion in bev for edge devices[J]. Sensors, 24(21):7007.

作者简介

夏若炎,2022 年于东南大学获得学士学位,现为东南大学硕士研究生,主要研究方向为自动驾驶,深度学习和多模态感知。
E-mail:amanda1381@ 163. com

Xia Ruoyan received her B. Sc. degree from Southeast University in 2022. Now she is a master's student in Southeast University. Her main research interests include multimodal perception, deep learning and autonomous driving.

徐晓苏,1982 年于东南大学获得学士学位,1985 年于东南大学获得硕士学位,1991 年于东南大学获得博士学位,现为东南大学二级教授,主要研究方向为控制理论与工程及导航定位技术。
E-mail:xxs@ seu. edu. cn

Xu Xiaosu received his B. Sc. degree from Southeast University in 1982, received his M. Sc. degree from Southeast University in 1985, received his Ph. D. degree from Southeast University in 1991. Now he is a Rank-2 Professor at Southeast University. His main research interests include control theory and engineering, navigation and positioning technologies.